Karlsruhe Institute of Technology

Statistical Guarantees for Generative Models as Distribution Estimators

Zur Erlangung des akademischen Grades eines

DOKTORS DER NATURWISSENSCHAFTEN

von der KIT-Fakultät für Mathematik des
Karlsruher Instituts für Technologie (KIT)
genehmigte

DISSERTATION

von

Lea Maria Kunkel, M.Sc.

Tag der mündlichen Prüfung:	3. Dezember 2025
Referent:	Prof. Dr. Mathias Trabs
Korreferenten:	Prof. Dr. Tilmann Gneiting
	Prof. Dr. Arnak S. Dalalyan

Abstract

Generative models have emerged as a symbol of artificial intelligence, enabling computers to mimic human behavior based on large datasets. While the empirical results are impressive, the theoretical understanding lags behind. Naturally, using more training data should result in a better model. In a rigorous mathematical setting, we aim to bound a model's error by a declining function of the number of samples. In this thesis, we study such upper bounds for two models: Generative Adversarial Networks (GANs) and Flow Matching. Furthermore, we will extend Flow Matching to the setting of conditional distribution estimation. Along the way, we will also investigate classical kernel-based methods for distribution estimation.

Since their introduction in 2014, GANs have evolved from the initial Vanilla setup to several adaptations. The statistical literature mainly focuses on Wasserstein GANs and their generalizations, which can build on the theory of optimal transport. In contrast, statistical results for Vanilla GANs are limited to very specific settings. To bridge this gap, we establish a connection between Vanilla GANs and the Wasserstein-1 distance by leveraging the neural network architecture commonly used in practice. This enables us to transfer various results, such as dimension reduction properties, to Vanilla GANs. Our findings finally explain the empirical success of these early generative models.

Flow Matching is a very recent generative model introduced in 2023 that just emerged as an alternative to diffusions, the current state-of-the-art. Consequently, theoretical results are limited. First, we demonstrate the natural connection between Flow Matching and the kernel density estimator. Then, we prove that, in an over-parameterized setting, Flow Matching achieves minimax optimal rates in the Wasserstein distance. Additionally, we study the regularity of the underlying dynamics, which are essential to statistical bounds. This enables us to derive rates using smaller networks, improving and extending the few preceding results.

Subsequently, we study Flow Matching in a conditional setting. This allows us to employ Flow Matching in a forecasting context. After extending the model in a mathematically reasonable way and showing the connection to a Nadaraya-Watson-type estimator, we connect proper scoring rules, which are a common way to measure model error in forecasting settings, to the concept of risk in statistical learning. Then we study rates of convergence in the risk associated to the Fourier-Score. After deriving lower bounds, we prove that the Nadaraya-Watson-type estimator is minimax optimal. For certain dimensions, we extend our results to the conditional Flow Matching estimator. Finally, we demonstrate Flow Matching's practical capability to estimate a conditional distribution.

Acknowledgements

First of all, I would like to thank Mathias Trabs for his excellent supervision over the past three years. Besides your inspiring and joyful approach to mathematics and research, I am deeply grateful for your ability to encourage and support people around you. No matter how bad or hopeless a situation seemed, I could always count on you to assess it exactly as I needed. I feel very privileged to be your student.

Furthermore, I would like to express my gratitude to Tilmann Gneiting and Arnak Dalalyan, who kindly agreed to serve as a referee for this thesis.

Additionally, I thank Oliver Stein for his didactically ingenious lecture *Nichtlineare Optimierung* which sparked my interest in mathematics after a rather unpleasant start of my Bachelor studies.

I would also like to thank Franz Nestmann for sharing his passion and talent for teaching with me. Thank you, Franz, for encouraging me from my time as a student-tutor until now and for being a great person for productive discussions - both on a professional and personal level.

My time at the Institute of Stochastics at KIT was truly wonderful, largely thanks to my amazing colleagues, who made every break the highlight of the day.

Furthermore, I would like to mention two students who, of course, must remain anonymous. Their determination and drive, regardless of how bad things get, are truly admirable. Seeing you succeed was one of the best moments of the last three years, and motivated me greatly.

Most importantly, I would like to thank my parents, my sister Anna, and my boyfriend Steffen for their unwavering support. Papa, thank you for being the best role model of integrity and hard work. Mama, thank you for teaching me how to care deeply about something. Anna, thank you for making me the proudest sister. Lastly, Steffen, thank you for making it impossible not to smile when I'm with you. Without you having my back every single day, neither of my 2025 theses would have been possible.

Prior publications and further declarations

A significant portion of this thesis is drawn from a previous publication and two preprints, which are available on arXiv.
Chapter 3 and Section 2.4 are based on the publication

- Kunkel, L. & Trabs, M. (2025b). A Wasserstein perspective of Vanilla GANs. Neural Networks, 181:106770.

Section 4.1, Section 5.2 and Section 5.4 are based on the preprint

- Kunkel, L. & Trabs, M. (2025a). On the minimax optimality of flow matching through the connection to kernel density estimation. arXiv preprint arXiv:2504.13336.

An earlier version of this work was subject to

- Kunkel, L. (2025). Flow Matching from a KDE perspective. In: Oberwolfach Rep. 22 (2025), No. 1, p. 398 - 390. Based on joint work with Mathias Trabs.

Section 5.5 is based on the preprint

- Kunkel, L. (2025) Distribution estimation via flow matching with Lipschitz guarantees. arXiv preprint arXiv:2509.02337.

Parts of the reasoning of the proofs of Section 5.5 are moved to Section 2.5. Furthermore, Section 5.3 is a mingling of both preprints. Mathias Trabs has agreed to the presentation of the results of Kunkel & Trabs (2025b) and Kunkel & Trabs (2025a) in this thesis. Direct quotes of the above mentioned publication and preprints appear throughout the thesis. Concerning the preprints Kunkel & Trabs (2025a) and Kunkel (2025), the direct quotes refer to the version available on the date of the submission of this thesis, which is the 8th of October 2025. Furthermore, it is intended to submit the results of Chapter 6.

Declaration about the use of AI tools This thesis was written in accordance with the KIT guidelines on the use of generative AI at KIT. By 1.d) of this guideline, the use of generative AI should be documented transparently whenever appropriate and necessary. The author considers all immediate influences of generative AI on this thesis to belong to this category. In this context, two AI models have been used: ChatGPT by OpenAI and DeepL by DeepL SE. Parts of the text were checked for typos and grammatical errors, and occasionally suggestions for synonyms and modification of word order were used. No paragraphs were rewritten by artificial intelligence. Furthermore, there no information was added to the text by AI tools.
In the context of the implementations, generative AI was used in data processing, data presentation and related tasks including troubleshooting. Every output was carefully checked and several sanity-checks were performed to assure the correctness. All plots originate from the author's ideas. The author considers herself the legal owner of all graphics in this thesis besides the KIT logo, the samples from the MNIST dataset presented in Figure 1.1 and the underlying map in Figure 6.4.

Notations

Relations, set of numbers and other basics

$\mathbb{R}$	Real numbers
$\mathbb{R}_{>0}$ or $(0,\infty)$	Positive real numbers
$\mathbb{R}_{\geq 0}$ or $[0,\infty)$	Nonnegative real numbers
$\mathbb{N}$	positive integers
$\mathbb{N}_0$	Nonnegative integers
$a \lesssim b$	$a \leq c \cdot b$ for a constant $c \in (0,\infty)$
$a \gtrsim b$	$a \geq c \cdot b$ for a constant $c \in (0,\infty)$
$a \asymp b$	$a = c \cdot b$ for a constant $c \in (0,\infty)$
$\mathcal{O}$	Landau notation
$\lceil x \rceil$	Largest integer not smaller than $x \in \mathbb{R}$
$\lfloor x \rfloor$	Largest integer not larger than $x \in \mathbb{R}$
$a \vee b$	Maximum of $a, b \in \mathbb{R}$
$a \wedge b$	Minimum of $a, b \in \mathbb{R}$
Ω°	Interior of the set Ω
$\mathbb{1}_\Omega$	Indicator function on the set Ω

Linear algebra

$\dim(\mathcal{X})$	Dimension of the space $\mathcal{X}$
$x^\top$	Transpose of the vector x
$A^\top$	Transpose of the matrix A
$V^\perp$	Orthogonal complement of the set V
$\langle \cdot, \cdot \rangle$	Euclidean inner product on $\mathcal{X}$
$\lvert \cdot \rvert$	Euclidean norm on $\mathcal{X}$
$\lvert \cdot \rvert_p$	p-norm on $\mathcal{X}$
$\lVert \cdot \rVert$	Spectral norm
$\mathrm{id}_\mathcal{X}$	Identity mapping on $\mathcal{X}$
$I_\mathcal{X}$ or $I_{\dim(\mathcal{X})}$	Matrix corresponding to $\mathrm{id}_\mathcal{X}$
A^{-1}	Inverse of the matrix A (if existent)
$A \succeq 0$	Matrix A is positive semi-definite
$A \succ 0$	Matrix A is positive definite
$A \preceq 0$	Matrix A is negative semi-definite
$A \prec 0$	Matrix A is negative definite
$A \succeq B$	A greater than or equal B in the Loewner order
e_i	i-th basis vector of the canonical basis of $\mathcal{X}$
$\lvert A \rvert_{\ell^0}$	Number of nonzero entries of the matrix A
$\mathrm{tr}(A)$	Trace of the matrix A

Analysis

$\operatorname{supp}(f)$	Support of the function f
$\frac{\partial}{\partial x_i} f$	Partial derivative of f with respect to x_i
∇f	Gradient of f
H_f	Hessian of f
$D_x f$	Jacobian of a vector valued function f with respect to $x \in \mathcal{X}$
D^k	Mixed partial derivative w.r.t. a multiindex k
$\operatorname{div}(f)$	Divergence of f
$\mathcal{F}f$	Fourier transform of f, see (2.10)
$\circ$	Concatenation of functions or sets of functions

Probability and measure theory

$\mathcal{B}_{\mathcal{X}}$	Borel σ-algebra on $\mathcal{X}$
δ_x	Dirac measure in $x \in \mathcal{X}$
$\frac{\mathrm{d}\mu}{\mathrm{d}\nu}$	Radon-Nikodym density of μ with respect to ν
$\ll$	Absolute continuity
$\mathbb{P}^X$	Distribution of the random variable X
$X \sim \mathbb{P}$	X is a random variable whose distribution is $\mathbb{P}$
$X \sim p$	X is a random variable whose density is p
$\mathbb{E}_{X\sim\mathbb{P}}[X]$ or $\mathbb{E}[X]$	Expected value of X
$\operatorname{Var}_{X\sim\mathbb{P}}(X)$ or $\operatorname{Var}(X)$	Variance of X
$\operatorname{Cov}_{X\sim\mathbb{P}}(X)$ or $\operatorname{Cov}(X)$	Covariance of X
$\mathcal{N}(a, \Sigma)$	Normal distribution with mean a and covariance Σ
$\mathcal{U}[a, b]$	Uniform distribution on $[a, b]$
$p \propto f$	Density p is proportional to f
$p^{\otimes n}$	n-fold product measure of the probability measure corresponding to the density p
φ_μ	Characteristic function of μ, see (2.11)
M_μ	Moment generating function of μ, see (2.12)
K_μ	Cumulant generating function of μ, see (2.13)

Function spaces

C^k	k-times continuously differentiable functions, $k \in [0, \infty]$
$\operatorname{Lip}(L)$	Lipschitz-L functions on Ω bounded by B, see (2.2)
$\mathcal{H}^\alpha$	α-Hölder functions, see (2.4)
$B^\alpha_{1,\infty}$	α-Besov functions, see (2.5)
$W^{k,\infty}$	Sobolev space, see (2.9)
H^s	Fractional Sobolev space, see (6.9)
$\|\cdot\|_\infty$	Supremum norm, see (2.1)
$\|\cdot\|_1$	L_1-norm, see (2.1)

$\|\cdot\|_{C^\beta}$	Supremum norm of the derivatives up to order β, see (2.7)
$\|\cdot\|_{\mathcal{H}^\alpha}$	α-Hölder norm, see (2.3)
$\|\cdot\|_{B^\alpha_{1,\infty}}$	α-Besov norm, see (2.6)
$\|\cdot\|_{W^{k,\infty}}$	Sobolev norm, see (2.8)
$\|\cdot\|_{H^s}$	Fractional Sobolev norm, see (6.10)
$\mathcal{N}(\tau, \mathcal{A}, \|\cdot\|_\infty)$	Covering number of a set $\mathcal{A}$ with respect to the supremum norm, see Definition 2.10

Distances between distributions

W_1	Wasserstein-1, see Definition 2.1
W_2	Wasserstein-2, see Definition 2.1
TV	Total variation, see Definition 2.3
KL	Kullback-Leibler divergence, see Definition 2.4
JS	Jensen-Shannon, see Definition 2.5

Abbreviations

ReLU	Rectified linear unit
ReQU	Rectified quadratic unit
KDE	Kernel density estimator
GAN	Generative adversarial network
ODE	Ordinary differential equation
SDE	Stochastic differential equation
CNF	Continuous normalizing flow
i.i.d.	Independent, identically distributed
PDF	Probability density function
CDF	Cumulative distribution function
NW	Nadaraya–Watson
CRPS	Continuous ranked probability score
s.t.	subject to

Contents

Chapter 1

Introduction

Generative models have emerged as one of the most prominent symbols of modern artificial intelligence tools. The idea that computers could mimic human activities used to be unimaginable, or at least futuristic. While technological advancements in this field are particularly rapid, the theoretical understanding of these models lags behind. This alone is enough to arouse interest in the mathematical foundations of generative models. Moreover, a profound understanding is necessary to adjust societal and legislative regulations in areas substantially impacted by generative models such as intellectual property. For instance image generation raises fundamental questions concerning artistic work and creativity.

Image generation also serves as the prototypical task of the models studied in this thesis. Assume we have a sample of images encoded in numerical values, which could for example be pictures of handwritten versions of the digit 3 with a fixed amount of pixels. For every pixel, a numerical value represents a certain gray scale. Now we want to build a model based on these observations, that creates a new picture looking like a handwritten 3. New means that the picture should not copy one of the 3s from the observations, but learn from the samples to create a seemingly handwritten 3 by itself just like another human would do. On a pixel level this means outputting a value for each pixel such that the corresponding picture is recognized as a 3 by humans. Instead of human creativity or distinct fine motor skills, the mathematical model uses a draw from a probability distribution and a function that transforms the draw to values whose colorization ideally looks like a 3. Another draw of the same probability distribution leads to a different 3. The process of finding this function is called *training*, the process of generating a new picture is called *data generation*. This general setting is illustrated in Figure 1.1. The precise model of Figure 1.1 and all subsequent images in this introduction can be found in Appendix A.
Figure 1.1 also illustrates what is known as dimension reduction. On the left hand side, the function G_1 is constructed such that the number of pixels of the input of G_1 is as large as the number of pixels in the samples. On the right hand side, the number of pixels of the input of G_2 is much smaller, instead of 784, only 25 pixels are used. As we can see, this leads to generated images that are not visibly worse than generated images of G_1.

While here the assessment of the author whether one 3 is not visibly worse than another 3 hopefully coincides with the judgment of the reader, it is far from a precise quantification of

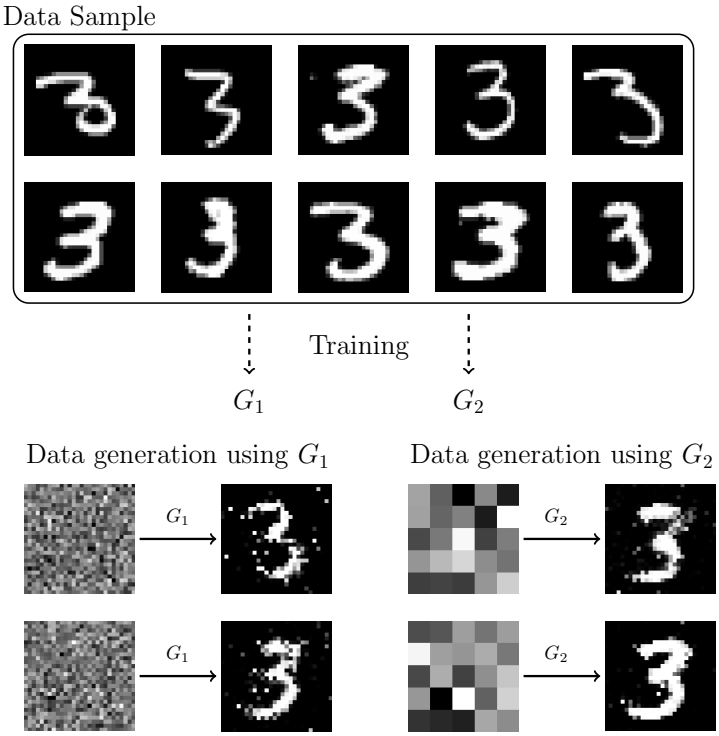


Figure 1.1.: Two different generative models trained on pictures of handwritten 3s. The first model, represented by the function G_1 transforms draws from the 784-dimensional Gaussian into images that look like a handwritten 3. The second model, represented by the function G_2, uses draws from the 25-dimensional Gaussian for the same task.

difference. The natural question, how well the model is in imitating humans writing down 3s, cannot be answered quantitatively by human judgment. Looking at the underlying mathematical construct, one can quickly find a better approach to quantifying these differences. The key objective of this thesis is to analyze how well a given model is depending on the number of samples that were used for training.

Generative models and their evaluation

To introduce the underlying mathematical construct, suppose we observe n independent, identically distributed (i.i.d.) observations $X_1, ..., X_n$ from an unknown probability distribution $\mathbb{P}^*$ on some measurable space $(\mathcal{X}, \mathcal{B}_\mathcal{X})$. Further, choose another probability distribution $\mathbb{U}$ on another measurable space $(\mathcal{Z}, \mathcal{B}_\mathcal{Z})$, the *latent* distribution. We want to learn a measurable, deterministic function $G\colon \mathcal{Z} \to \mathcal{X}$ such that the distribution of a transformed sample Z from $\mathbb{U}$, denoted by $\mathbb{P}^{G(Z)}$, is a good imitation of $\mathbb{P}^*$. In statistical language, this would be referred to as $\mathbb{P}^{G(Z)}$ being a good estimator of $\mathbb{P}^*$. As we are not approximating a certain property of a distribution, but the entire distribution as such, this is called distribution estimation. The goal is that this imitation of the unknown distribution is as good as possible. Naturally, this can either be the case by choosing a good latent distribution $\mathbb{U}$ or by choosing a good function G. The generative approach is to focus on the latter, but aim for a preferably small dimension of $\mathcal{Z}$. Thus, suppose we have a set $\mathcal{G}$ of potential functions for G and want to choose the best possible G. This conceptual setting of a generative model is summarized in Figure 1.2.

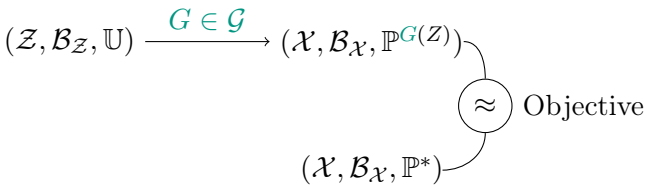


Figure 1.2.: Conceptual setting of generative models. $\mathbb{U}$ represents the latent distribution. A sample Z from $\mathbb{U}$ is transformed using a function G from the generator class to mimic the unknown distribution $\mathbb{P}^*$.

From Figure 1.2 two questions arise immediately. The first is in which way to measure how well $\mathbb{P}^{G(Z)}$ approximates $\mathbb{P}^*$. Thereupon, the second concerns the choice of the set $\mathcal{G}$ and the function G.

Quantifying the difference between two probability distributions is a question that arises in many settings, particularly in statistics. Thus, it has been of theoretical interest for a long time. There are several metrics and divergences that have been studied in various settings. Since, in general, metrics on spaces of probability distributions are not equivalent, the concrete choice can heavily influence the generative model. For distances that are not metrics, this is at least equally important. Note that in this thesis, a distance refers to any dissimilarity measure, not necessarily a metric. In this introduction, we are going to focus on one key aspect that a distance should have in this setting: The distance used for evaluation should again be chosen such that there is a meaningful interpretation of one choice of G being better than another in terms of this distance. Additionally, a lot of distances are related to each other, some being stronger and immediately implying bounds in other distances. Thus, choosing a rather strong distance is beneficial. As we shall see, stronger distances and meaningful interpretations act contrarily in some cases.

A straightforward choice of G is then such that it minimizes a distance between $\mathbb{P}^{G(Z)}$ and $\mathbb{P}^*$ over all $G \in \mathcal{G}$. This can either be the same distance that was chosen to evaluate the model, but is of course not restricted to this choice. Subsequently, an easy naive approach is to set $\mathcal{G}$ as large as possible, including all $\mathcal{B}_\mathcal{Z}$-$\mathcal{B}_\mathcal{X}$-measurable functions. However, for an efficient sampling from the generated distribution, we need to be able to evaluate the function quickly. This implies that the function needs to be implementable on a computer and the mathematical components of the functions must be sufficiently simple. At the same time, the function class should ideally be flexible and large. Therefore, $\mathcal{G}$ is typically chosen as some set of neural networks.

In the first chapters, we are going to use the Wasserstein-1 distance between two probability distributions $\mathbb{P}$ and $\mathbb{Q}$, on $\mathbb{R}^d$ equipped with the Euclidean norm $|\cdot|$ defined as

$$\mathsf{W}_1(\mathbb{P}, \mathbb{Q}) = \inf_{\pi \in \Pi(\mathbb{P}, \mathbb{Q})} \int_{\mathbb{R}^d} |x - y| \, \mathrm{d}\pi(x, y),$$

where $\Pi(\mathbb{P}, \mathbb{Q})$ is the set of all joint distributions whose marginals are $\mathbb{P}$ and $\mathbb{Q}$. The Wasserstein-1 distance can be interpreted as the minimal effort in terms of the Euclidean norm required to shift the mass from $\mathbb{P}$ to $\mathbb{Q}$. This distance is very well studied in the context of optimal transport, see Villani (2008), and profits from a nice dual form. It is weak enough to allow for meaningful interpretations and strong enough to metrize weak convergence. This makes the Wasserstein-1 distance a frequent choice when evaluating distribution estimators, see for example Liang (2017); Huang et al. (2022); Chen et al. (2020); Lee et al. (2025); Stéphanovitch et al. (2024); Berry & Sauer (2017); Berenfeld & Hoffmann (2021); Divol (2022); Wu & Wu (2022); Gao et al. (2024b); Schreuder et al. (2021); Vardanyan et al. (2024). The precise application of the Wasserstein-1 distance in these references will be discussed in the course of this thesis. In this introduction, the references will mostly be presented as short, enumerative examples. In the corresponding chapters, we will discuss the literature in more detail and emphasize the contributions of this thesis based on it.
In other settings, the choice of the evaluation distance will be motivated by the area of application. In Chapter 6 we are going to study generative models in forecasting settings. Thus, we use distances based on *proper scoring rules*, which is the classical tools to evaluate forecasts.

Generative adversarial networks

The first method we are going to study are **Generative Adversarial Networks (GANs)**, which were introduced by Goodfellow et al. (2014). In the aforementioned theoretical setting, this model uses the shifted Jensen-Shannon distance as an optimization criterion. This leads to the following objective function, called the *Vanilla GAN*:

$$\inf_{G\in\mathcal{G}} \mathrm{JS}(\mathbb{P}^{G(Z)}, \mathbb{P}^*) - \log(4) = \inf_{G\in\mathcal{G}} \sup_{\substack{D \text{ measurable} \\ D(\mathcal{X})\subset(0,1)}} \mathbb{E}_{\substack{X\sim\mathbb{P}^* \\ Z\sim\mathbb{U}}} \big[\log D(X) + \log\big(1 - D(G(Z))\big)\big]. \quad (1.1)$$

Without going into detail about the existence of the minimum and maximum here, this shows a classical minimax-game: the function D, which is called the *discriminator* is chosen so that its values are as close as possible to 1 wherever the mass of $\mathbb{P}^*$ lies and as close as possible to 0 wherever the mass of $\mathbb{P}^{G(Z)}$ lies. When evaluated empirically, using observation of $\mathbb{P}^*$ and $\mathbb{P}^{G(Z)}$, the discriminator can be interpreted as a classifier. In adversarial position, the *generator* G wants to transform Z such that the discriminator assigns values as close as possible to 1 on the mass of $\mathbb{P}^{G(Z)}$.
In practice, G and D are parameterized using neural networks whose parameters are optimized successively. This implies that the Jensen-Shannon distance is not calculated directly, but is rather estimated via the function D. The GAN approach of determining the function G directly is classical for early generative models in the 2010s. Another prominent example is the Wasserstein autoencoder (Tolstikhin et al., 2018). The use of the Jensen-Shannon distance leads to the problem that in case $\mathbb{P}^{G(Z)}$ and $\mathbb{P}^*$ are singular, this distance is by definition maximal. This is regularly the case if $\dim(\mathcal{Z}) < \dim(\mathcal{X})$, which refers to the right hand side of Figure 1.1, where the number of pixels in the latent space is much lower than the number of pixels in the generated image. As this is an important setting in practice, the underlying Jensen-Shannon distance has

been replaced by several other distances. The most famous replacement leads to the Wasserstein GAN (Arjovsky et al., 2017) using the dual formulation of the 1-Wasserstein distance

$$\inf_{G\in\mathcal{G}} \mathsf{W}_1(\mathbb{P}^{G(Z)},\mathbb{P}^*) = \inf_{G\in\mathcal{G}} \sup_{D\in\mathrm{Lip}(1)} \mathbb{E}_{\substack{X\sim\mathbb{P}^*\\ Z\sim\mathbb{U}}}\big[D(X)-D(G(Z))\big]. \tag{1.2}$$

Just like in Vanilla GANs, the Wasserstein GAN has the classical adversarial structure: the discriminator function W is chosen such that it discriminates as well as possible between $\mathbb{P}^*$ and $\mathbb{P}^{G(Z)}$, the generator function G tries to make this discrimination as hard as possible. The difference lies in the objective function and the restriction on the function D. Again, the set of Lipschitz 1 functions is in practice replaced by neural networks. The Wasserstein distance allows for a more meaningful interpretation of $\mathsf{W}_1(\mathbb{P}^{G_1(Z)},\mathbb{P}^*) = \mathsf{W}_1(\mathbb{P}^{G_2(Z)},\mathbb{P}^*)$ for $G_1, G_2 \in \mathcal{G}$ also in case of $\dim(\mathcal{Z}) < \dim(\mathcal{X})$. As already mentioned, the Wasserstein distance is very well studied. Hence, most theoretical results focus on the statistical properties of Wasserstein GANs or close relatives: next to optimization and asymptotic properties (Biau et al., 2021), error decompositions in Kullback-Leibler divergence, the Hellinger distance and the Wasserstein distance (Liang, 2021), dimension reduction settings (Schreuder et al., 2021; Tang & Yang, 2023) and general rates of convergence in the Wasserstein distance in several settings (Liang, 2017; Huang et al., 2022; Chen et al., 2020; Lee et al., 2025; Stéphanovitch et al., 2024) have been studied. These results will be reviewed more detailed in Chapter 3.

Although Vanilla GANs were introduced much earlier, the statistical understanding remained largely very limited. Biau et al. (2020) derived a central limit theorem and Puchkin et al. (2024) used smooth neural networks to evaluate the model in the Jensen-Shannon distance itself assuming that $\mathcal{Z} = \mathcal{X}$. Despite the theoretical limitations, Vanilla GANs did work in practice even in cases where $\dim(\mathcal{Z}) < \dim(\mathcal{X})$. In fact, the model used for Figure 1.1 is a Vanilla GAN. In Chapter 3 we are going to bridge this gap. This allows us to obtain the first rate of convergence for Vanilla GANs allowing for singular measures.

The prototypical analysis of GANs separates the error that occurs due to the use of neural networks instead of measurable or Lipschitz-1 functions and the error that the model (1.1) (or (1.2) respectively) itself causes. Using this approach, it is unclear how to obtain convergence results for the Vanilla GAN when $\dim(\mathcal{Z}) < \dim(\mathcal{X})$. In order to analyze Vanilla GANs, we are going to take another perspective: instead of paying for the use of neural networks in the proof, we exploit this restriction of the underlying model. We show that using a set of neural networks instead of all measurable functions actually enables the model to cope with the case $\dim(\mathcal{Z}) < \dim(\mathcal{X})$. As a network class, we use Hölder continuous feedforward ReLU networks and derive a novel approximation result suited for our conditions. This result extends Gühring et al. (2020), who approximate a function and its derivative. Due to the restriction of the discriminator class, a uniform bound on the approximation error such as Yarotsky (2017); Kohler & Langer (2021); Schmidt-Hieber (2020) is not enough for our purpose. Notably, we do not need to use smooth networks, which aligns more closely to practice. We derive a rate of convergence with respect to the Wasserstein-1 distance for a broad class of unknown distributions $\mathbb{P}^*$. If the intrinsic dimension of $\mathbb{P}^*$ is smaller than the dimension of $\mathcal{X}$, our rate depends on the dimension of

the latent space $\mathcal{Z}$ and thus circumvents the curse of dimensionality. In the end, we demonstrate our theoretical findings on synthetic data.

Generative Flow Matching

The second model we are going to analyze is typical of generative models in the 2020s: instead of obtaining G as the minimizer of some optimization problem directly, models are designed such that underlying dynamics are approximated. This seemingly more complicated approach often leads to objectives that are much easier. Distances between probability measures are usually hard to calculate in high dimensions, e.g. the Wasserstein distance on $\mathbb{R}^d$ for $d \geq 2$. As we saw, one key feature of GANs is that this distance is also approximated using a neural network. The second model we are going to study is **Flow Matching**, which is one example of this more recent approach of generative models. In Flow Matching the mapping G is replaced by a function with an additional time input $\psi\colon \mathcal{X} \times [0,1] \to \mathcal{X}$. This function is obtained as the solution of an ordinary differential equation (ODE),

$$\frac{\partial \psi_t}{\partial t}(x) = v_t(\psi_t(x)), \quad \psi_0(x) = x, \quad \forall x \in \mathcal{X}, \tag{1.3}$$

where $v\colon \mathcal{X} \times [0,1] \to \mathcal{X}$ is a vector field. Note that the input dimension in space is the same as the output dimension. The vector field v is chosen so that for a fixed latent distribution $\mathbb{U}$ on $\mathcal{X}$, $Z \sim \mathbb{U}$ and $t \in [0,1]$, the distribution of $\psi_t(Z)$ has nice properties. Additionally, v should be constructed so that (1.3) has a unique solution. By the boundary condition (1.3) we know that $\mathbb{P}^{\psi_0(Z)} \sim \mathbb{U}$. We want to construct the model so that $\mathbb{P}^{\psi_1(Z)} \approx \mathbb{P}^*$. Note that for the generative model in the GAN setting, these two distributions would be enough. However the Flow Matching model provides estimates for all $t \in [0,1]$. In the setting of Lipman et al. (2023), both $\mathbb{U}$ and $\mathbb{P}^*$ are assumed to admit densities, p and p^* respectively. The vector field v_t is then constructed such that the density p_t of $\psi_t(Z)$ is given by

$$p_t(x) = \int p\Big(\frac{x - \mu_t(y)}{\sigma_t}\Big) p^*(y)\, \mathrm{d}y, \tag{1.4}$$

where $\mu\colon \mathcal{X} \times [0,1] \to \mathcal{X}$ is a *mean shift* function and $\sigma\colon [0,1] \to \mathbb{R}_{>0}$ is a *variance function*. These two functions are chosen such that

$$p_0(x) = p(x), \quad p_1(x) = \int p(\frac{x-y}{\sigma_{\min}}) p^*(y)\, \mathrm{d}y,$$

where $\sigma_{\min} \in (0,1)$. In case of $\mathcal{X} = \mathbb{R}$, Figure 1.3 illustrates a flow where p_0 is the density of the standard Gaussian.

The corresponding vector field v can be obtained through an equivalent formulation of (1.3) that connects a vector field to the corresponding density path via a partial differential equation. This vector field should be approximated by a function class $\mathcal{M}$, i.e. a function $\hat{v} \in \mathcal{M}$ is chosen as

$$\inf_{\tilde{v} \in \mathcal{M}} \mathbb{E}_{\substack{t \sim \mathcal{U}(0,1) \\ X_t \sim p_t}} [|\tilde{v}_t(X_t) - v_t(X_t)|^2]. \tag{1.5}$$

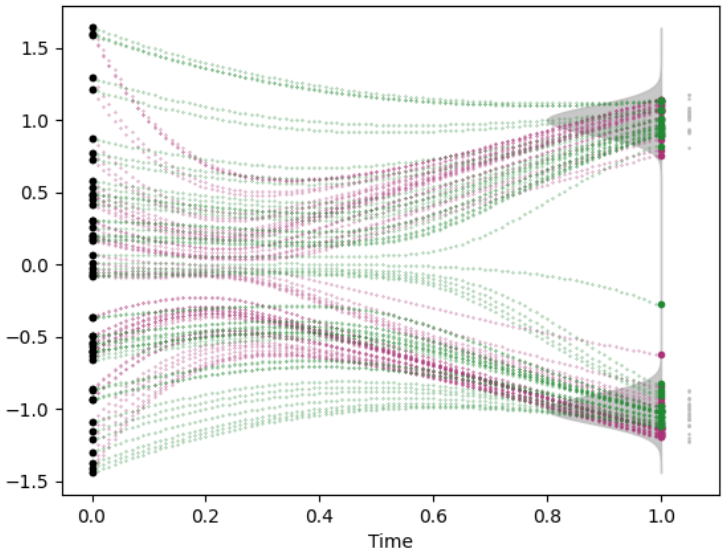


Figure 1.3.: Flow from $\mathbb{U} = \mathcal{N}(0,1)$ over time $t \in [0,1]$ to a bimodal Camelback distribution for two different variance functions, $\sigma_t^{(1)} = 1 - (1 - \sigma_{\min})t$ (green dots) and $\sigma_t^{(2)} = \sigma_{\min}^t$ (red dots). Each model is trained on $n = 50$ samples (grey dots) and uses the linear mean shift.

In practice, $\mathcal{M}$ is a class of neural networks. The function for the generative model ψ_1 is then obtained by solving the ODE (1.3). Compared to GANs, we thus do not obtain a generator function directly, but rather the underlying vector field of this generator function. To see why this is beneficial, we need to take one more step: Neither p_t nor v_t are accessible in practice, as they depend on the unknown distribution p^*. Thus, the important observation of Lipman et al. (2023) is that in case of a parameterized class $\mathcal{M}$, the gradients of (1.5) with respect to the parameters of $\tilde{v}$ are the same as the gradients of

$$\mathbb{E}_{\substack{t\sim\mathcal{U}(0,1) \\ X\sim\mathbb{P}^* \\ X_t\sim p(\frac{\cdot-\mu_t(X)}{\sigma_t})}} \left[|\tilde{v}_t(X_t) - v_t(X_t|X)|^2\right], \tag{1.6}$$

where $v_t(\cdot|x), x \in \mathcal{X}$ is a vector field that can be derived from the setting in closed form. Looking at the empirical counterpart of (1.6)

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\substack{t\sim\mathcal{U}(0,1) \\ X_t\sim p(\frac{\cdot-\mu_t(X_i)}{\sigma_t})}} \left[|\tilde{v}_t(X_t) - v_t(X_t|X_i)|^2\right], \tag{1.7}$$

we can finally see the advantage of Flow Matching. We are in full control of every term in (1.7) and the objective to minimize is a simple least squares problem. This easy objective function is one of the key characteristics of generative models in the 2020s. Another very famous approach are diffusion models (Song et al., 2021), which are based on stochastic differential equations (SDEs). Instead of a vector field, the gradient of the log-density, also called the score, of a constructed time dependent density is approximated. This approach leads to a least squares

problem similar to (1.7). Diffusion models leverage the theory of score matching. This also indirectly explains the name Flow Matching for the ODE based model.

For our theoretical analysis of Flow Matching, we choose the Wasserstein-1 metric as an evaluation distance. Additionally to the previous mentioned advantages of this metric, this allows for comparability to the results obtained for Vanilla GANs. As we are interested in the performance of the distribution estimation model, we aim to analyze $\mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{\hat{\psi}_1(Z)})$, where $\hat{\psi}$ is the solution to the ODE (1.3) using the minimizer of (1.7) for the vector field. Given the structure of the Flow Matching model, it is natural to choose a reference ψ^* that is the solution of the ODE (1.3) using a reference vector field v^* and decompose the error in the following way

$$\mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{\hat{\psi}_1(Z)}) \leq \underbrace{\mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{\psi_1^*(Z)})}_{\text{reference error}} + \underbrace{\mathsf{W}_1(\mathbb{P}^{\psi_1^*(Z)}, \mathbb{P}^{\hat{\psi}_1(Z)})}_{\text{approximation error}}.$$

For the reference v^*, there are two natural candidates: the first one is v from (1.5), which is the vector field leading to the density path (1.4). This corresponds to the choice of the scarce previous theoretical works in this area of research (Fukumizu et al., 2025; Gao et al., 2024b). Hereafter, we are going to call this choice the *population reference*. There is another candidate for v^*. Going backwards from the empirical least squares problem (1.7), we show that there are empirical counterparts of (1.5), of the vector field and of (1.4). The empirical counterpart of (1.4) is given by

$$p_t^n(x) = \frac{1}{n}\sum_{i=1}^{n} p\Big(\frac{x - \mu_t(X_i)}{\sigma_t}\Big). \tag{1.8}$$

This is related to the kernel density estimator using the density p as a kernel, which has been studied in the statistical literature for a long time. We are going to call this the *empirical reference*. In this thesis, we are going to study both settings, the empirical and the population reference.

The reference error $\mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{\psi_1^*(Z)})$ is the Wasserstein distance between distributions with densities (1.4) or (1.8) and the unknown distribution $\mathbb{P}^*$. In the mentioned literature, this error is typically forced to be negligible by choosing $\sigma_{\min}$ extremely small. In kernel density estimation in contrast, the choice of the bandwidth, which corresponds to the choice of $\sigma_{\min}$ in our setting, is of utmost importance for the capability to profit from smoothness in the unknown distribution. In the analysis via the empirical reference, studying $\mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{\psi^*(Z)})$ corresponds to studying the kernel density estimator.

The classical literature on kernel density estimation focuses on error bounds in the mean squared error or the L^1 error (Tsybakov, 2009; Devroye & Lugosi, 2012; Scott, 1992). Albeit these results sometimes imply bounds on Wasserstein distance, the corresponding bounds are suboptimal. Recently, the use of the Wasserstein-1 distance became more popular in the analysis of kernel density estimation, particularly on unknown manifolds (Berry & Sauer, 2017; Berenfeld & Hoffmann, 2021; Divol, 2022; Wu & Wu, 2022), but the only optimal bounds so far are for very specific kernels that do not include the Gaussian kernel. The use of the Gaussian kernel corresponds to using $\mathbb{U} = \mathcal{N}(0, I_d)$ as latent distribution. We are going to show that the kernel

density estimator can estimate certain unknown densities with a rate of convergence that is optimal up to logarithmic factors. Our results allow for standard choices of kernels such as the Gaussian kernel. Furthermore, in case the unknown distribution is supported on a linear subspace, we show that the curse of dimensionality can be circumvented. This result and the use of the empirical reference enables us to profit from a careful choice of $\sigma_{\min}$.

In the analysis of the population reference, the reference error $\mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{\psi_1^*(Z)})$ itself is minimized by choosing $\sigma_{\min}$ as small as possible. In the limit case $\sigma_{\min} \to 0$, (1.4) implies for the convolution $p_1(x) \to p^*(x)$. However, we are going to see that even in this case, the theoretical analysis of the entire model can profit from a carefully chosen $\sigma_{\min}$.

The second step is to bound the error caused by approximation, $\mathsf{W}_1(\mathbb{P}^{\psi_1^*(Z)}, \mathbb{P}^{\hat{\psi}(Z)})$. While the approximation of underlying dynamics leads to the easier objectives (1.6) and (1.7), in the theoretical analysis we need to track down the effects of the vector field approximation on the corresponding generated distributions. Albeit the stability of solutions of ODEs is a question that arises in various problems, Grönwall's inequality is the method of choice in such general settings, see for example Albergo & Vanden-Eijnden (2023); Benton et al. (2024); Gao et al. (2024b); Fukumizu et al. (2025); Stéphanovitch et al. (2025). This leads to bounds that depend exponentially on the Lipschitz constant of the vector field, which is one of the key difficulties in the theoretical analysis of Flow Matching. In contrast, diffusion models, whose similarities to Flow Matching will be motivated in Section 5.1.2, can profit from Girsanov's theorem which circumvents this issue (Chen et al., 2023a,b,c; Oko et al., 2023; Tang & Yang, 2024; Azangulov et al., 2024; Zhang et al., 2024; Yakovlev & Puchkin, 2025). As a side effect of a fast evolving area of research, some of the very recent results contain critical flaws. We defer a comment to Chapter 5.

In our analysis via the empirical reference, we are going to look at the over-parameterized setting. Using the empirical reference allows us to obtain bounds without making the trade-off in network size that is typical in settings of empirical risk minimization over a set of functions, by far non-exhaustive examples include in context of GANs Liang (2021), in context of diffusions Oko et al. (2023); Yakovlev & Puchkin (2025) and more general in context of score-based generative models Stéphanovitch et al. (2025). This is also the approach of the previous results for Flow Matching, see Gao et al. (2024b); Fukumizu et al. (2025). Therefore we can compensate the dependency on the Lipschitz constant and ultimately obtain rates of convergence for a large class of bounded unknown distributions that are minimax optimal up to a logarithmic factor. We can also extend our dimension reduction result for unknown distributions supported on a linear subspace to the Flow Matching estimator.

In the analysis via the population reference, we are going to employ a more classical approach. Using a Bernstein-type inequality results in a bidirectional effect when larger networks are used. This prohibits the compensation approach we exploited before. Thus, we start our analysis with a detailed study of the Lipschitz constant of the population reference v. This study includes lower bounds on the Lipschitz constant and a collection of assumptions that guarantee control over the exponential term caused by the use of Grönwall's lemma. Furthermore, our results hold for a broad class of variance functions σ_t. General variance functions have, to the best of the author's

knowledge, not been studied in a statistical context. However, the optimal-transport-based result of Tsimpos et al. (2025) shows that common choices are not optimal, which indicates the need for general results. Afterwards we show that these assumptions are met by a class of unbounded, non-log-concave distributions. Then, we use higher order smoothness of the population reference v, which can be controlled by an appropriately chosen $\sigma_{\min}$, to improve the subsequent rate of convergence. Although this rate is not optimal in a minimax sense, it improves existing results of Gao et al. (2024b) for the estimation of unbounded distributions using Flow Matching.

Conditional distribution estimation via Flow Matching

So far, we have focused on learning how to approximate an unknown distribution $\mathbb{P}^*$. Albeit the goal has been the generation of new samples, the ability to sample in a cheap way paves the way to estimating characteristic quantities, such as the mean or the variance of the unknown distribution, via classical estimation methods. While this is of interest on its own, many applications need specific characteristics based on some additional information that can vary. One example of such applications is temperature prediction for the next day based on today's temperature and air pressure. Depending on the purpose, different values are of interest: A hiker probably wants to know an interval in which the actual temperature lies with high probability in order to provide corresponding equipment. An airport controller is rather interested in worst case bounds in order to prevent accidents. In both cases, a simple point forecast corresponding to a mean regression problem is insufficient. Rather, knowledge of the distribution given the available information is necessary.

In a mathematical setting, this corresponds to estimating a conditional distribution and inferences thereof. Thus, we now assume we observe n samples $(X_1, W_1), ..., (X_n, W_n)$ from the joint distribution $\mathbb{P}^*_{\mathcal{X},\mathcal{W}}$. Our goal is to estimate the conditional distribution $\mathbb{P}^*_{\mathcal{X}|w}$ for $w \in \mathcal{W}$.

A straightforward approach to adapt Flow Matching to the setting of conditional distribution estimation is to equip the vector field with an additional input in (1.7), leading to

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\substack{t\sim\mathcal{U}(0,1) \\ X_t\sim p(\frac{\cdot-\mu_t(X_i)}{\sigma_t})}} [|\tilde{v}_t(X_t, W_i) - v_t(X_t|X_i)|^2]. \tag{1.9}$$

In this setting $\tilde{v}\colon [0,1] \times \mathcal{X} \times \mathcal{W} \mapsto \mathcal{X}$, where $\mathcal{W}$ is the space of the covariates. From a machine learning perspective, this is a special instance of a *guided* Flow Matching model (Zheng et al., 2023). These models inter- and extrapolate between the unconditional and the conditional model. So far there are, to the best of the author's knowledge, no theoretical results of this model. Conditional diffusions have been analyzed by Tang et al. (2025).
We start by showing that the Flow Matching model corresponds to vector fields that are well defined only on a finite set of points. While this hinders further statistical analysis, a simple

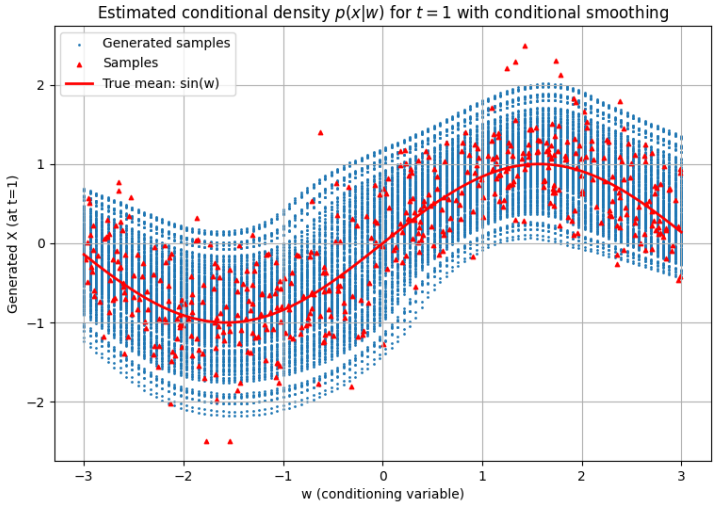


Figure 1.4.: Conditional density estimation with smoothing in the covariates, based on 500 samples, $W \sim U[-3,3]$, $X \sim \mathcal{N}(\sin(W), 0.5^2)$, 200 latent samples are chosen once and then put through the model for different values of w.

adaptation avoids this. Thus, we introduce

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\substack{t\sim\mathcal{U}[0,1] \\ X_t\sim p(\frac{\cdot-\mu_t(X_i)}{\sigma_t}) \\ W\sim K_{h_w}(\cdot-W_i)}} \left[\, |\tilde{v}_t(X_t, W) - v_t\left(X_t|X_i\right)|^2 \,\right], \tag{1.10}$$

where K_{h_w} is a kernel with bandwidth h_w independent of t. Given a value $w \in \mathcal{W}$ and a minimizer $\hat{v}$ of (1.10) over some function class, we can solve the adapted ODE

$$\frac{\partial \psi_{t,w}}{\partial t}(x) = \hat{v}_{t,w}(\psi_{t,w}(x)), \quad \psi_{0,w}(x) = x, \quad \forall x \in \mathcal{X}, \forall w \in \mathcal{W}. \tag{1.11}$$

The solution $\hat{\psi}_{t,w}$ leads to the estimated conditional distribution $\mathbb{P}^{\hat{\psi}_{t,w}(Z)}$. Figure 1.4 illustrates a toy example in this setting. As we can sample from $\mathbb{P}^{\hat{\psi}_{t,w}(Z)}$ at the small cost of sampling from $\mathbb{U}$ and applying the function $\hat{\psi}_{t,w}$, we can obtain estimates for nearly arbitrary properties of $\mathbb{P}^*_{\mathcal{X}|w}$.

In order to analyze the theoretical capability of (1.10) we again need to choose an evaluation metric. As motivated above, conditional distribution estimation is the theoretical framework of probabilistic forecasting. To achieve comparability, we adapt *proper scoring rules* to choose an evaluation metric. Scoring rules are designed to evaluate an estimated distribution given a true observation (Gneiting & Raftery, 2007). The characteristic *proper* refers to the preferable property that the true distribution minimizes the scoring rule within a class of alternatives. Common choices include the energy score S_E by Gneiting et al. (2007), defined for some $\beta \in (0,2)$ which maps a distribution $\mathbb{P}$ and a value $x \in \mathcal{X}$ to

$$S_E(\mathbb{P}, x) := \mathbb{E}_{X\sim\mathbb{P}}\left[|X - x|_2^\beta\right] - \frac{1}{2}\mathbb{E}_{X,X'\overset{\text{i.i.d}}{\sim}\mathbb{P}}\left[|X - X'|_2^\beta\right]. \tag{1.12}$$

The energy score is closely related to energy statistics by Székely (2003).

As visible from the example (1.12), a scoring rule can be used to compare a predictive distribution $\mathbb{P}$ to a realization x. It accounts for both, the location of x relative to $\mathbb{P}$, represented by the first term in (1.12), and the spread of the predictive distribution, represented by the second term in (1.12). In a first step, we connect the concept of proper scoring rules to the notions of risk and thus, empirical risk minimization, which is fundamental to statistical learning. In a nutshell, we are only interested in the difference between the model estimator and the best possible estimator that would be chosen in case of full information.

Afterwards, we adopt the use of the empirical reference as in the unconditional case. This reveals that the model (1.10) is closely connected to a Nadaraya-Watson-type estimator and thus the classical extension of the kernel density estimator to the setting of conditional density estimation (Hall et al., 1999). Proper scoring rules have hardly been investigated in the context of statistical learning, except for Pic et al. (2023). First, we study lower bounds with respect to the risk associated with the Fourier score, a generalization of the energy score (1.12). Then we show that the Nadaraya-Watson-type estimator achieves this rate and is thus minimax optimal. For the energy score, we obtain a rate that is as fast as the corresponding rate in a mean regression problem evaluated in the weighted L_2 distance, see Györfi (2002). Afterwards, we use our results to derive a rate of convergence for the Flow Matching model using the energy score.
Ultimately, we apply our model to classical forecasting datasets, weather prediction tasks, and toy examples that illustrate the behavior of our estimator. Our experiments show that the Flow Matching estimator is a promising approach to estimating conditional distributions.

Outline

This thesis is structured as follows.
Chapter 2 begins with preliminaries, including general definitions of function spaces and probability theory concepts. We discuss the distances between probability distributions necessary for this thesis. Then, we briefly introduce the concept of proper scoring rules, connect them to the previously presented distances, and define the scoring rules that will be used later. Afterwards, we provide a precise definition of feedforward ReLU networks, which serve as the prototype for the neural networks employed in subsequent chapters. In Section 2.4, we review the approximation result used in Chapter 5 and Chapter 6 and derive a novel approximation result suited for the setting in Chapter 3. Subsequently, we recall functional and concentration inequalities needed in Section 5.5. In the end, we present a conceptual proof of an oracle inequality that serves as a starting point for subsequent proofs.
Chapter 3 studies the Vanilla GAN. First, we introduce the Vanilla GAN distance, which characterizes the optimization problem (1.1). Section 3.2 investigates the relationship between the Vanilla GAN distance and the Wasserstein distance. We demonstrate that, while not equivalent, the two distances are compatible with each other. Using this relationship, we derive an oracle inequality for the Vanilla GAN, where $\mathcal{G}$ is a nonempty compact set and $\mathcal{D}$ is a

set of Lipschitz functions. We show that Vanilla GANs can avoid the curse of dimensionality. Afterwards, we consider the situation where $\mathcal{G}$ and $\mathcal{D}$ consist of neural networks. Here we relax the Lipschitz condition to a α-Hölder condition and prove a convergence rate for the Vanilla GAN with network generator and discriminator. Subsequently, we derive a convergence rate for Wasserstein-type GANs with a network generator and discriminator using our approximation result. This allows us to directly compare Vanilla GANs to Wasserstein GANs. In the end, we illustrate our theoretical results with a numerical example based on synthetic data.

Chapter 4, revisits the kernel density estimator. We derive a rate of convergence in the Wasserstein distance for the kernel density estimator, which is optimal up to logarithmic constants. Afterwards we show that, in case the unknown distribution is supported by a linear subspace, the kernel density estimator can overcome the curse of dimensionality.

Chapter 5 investigates Flow Matching. First, we provide a brief overview of related models. Then, in Section 5.2, we demonstrate the connection to kernel density estimation. Afterwards, we introduce the two reference models and derive a general error decomposition. Section 5.4 studies the over-parameterized setting, using the results of Chapter 4 to obtain minimax optimal rates up to logarithmic constants. Section 5.5 investigates smaller networks. After studying the Lipschitz constant of the vector field v from (1.5) for general variance functions, we derive a rate of convergence.

Chapter 6 begins by adapting Flow Matching to the conditional distribution estimation setting. Then we connect the concepts of proper scoring rules to risk. We derive a lower bound on the risk related to the Fourier score and show that the Nadaraya-Watson-type estimator achieves this rate. Subsequently, we demonstrate that the Flow Matching estimator can be minimax optimal in the energy score. Ultimately, we apply the Flow Matching estimator to toy examples, classical forecasting datasets, and weather prediction tasks.

Chapter 7 presents an overall conclusion and further avenues for research.

Throughout the thesis, the proofs are placed at the end of each chapter. An exception from this is made in Chapter 2, where most proofs consist of references and brief remarks. The proof of the novel approximation result is again moved to the end of Chapter 2.

Chapter 2

Foundations and network approximation

In this chapter, we recall definitions from the theory of function spaces and concepts from probability theory that are essential for the subsequent chapters. Most importantly, we define the distances between probability distributions and scoring rules used in this thesis and introduce feedforward rectified linear unit (ReLU) neural networks. We also prove a novel approximation result needed for Chapter 3. In the end, we present a short outline of a proof of a convergence rate in distribution estimation, that will serve as a starting point for the subsequent chapters.
We assume that basic concepts of probability theory as well as fundamentals from calculus and linear algebra, as those listed in the notation overview without a reference to a definition, are known. Further, we will always adapt the definitions to the Euclidean setting studied in this thesis. To prevent potential confusion concerning the dimension we will consider spaces $\mathcal{X}$ and $\mathcal{Y}$ which are finite-dimensional Cartesian products of real numbers in this chapter. Likewise, we are going to denote the probability measures in this chapter with Greek lower case letters, reserving blackboard bold letters for later.
Whenever referring to a probability measure μ on $\mathcal{X}$, we consider a probability space $(\mathcal{X}, \mathcal{B}_{\mathcal{X}}, \mu)$. As standard in statistics, when $X \sim \mu$, we assume the canonical setting $X\colon (\mathcal{X}, \mathcal{B}_{\mathcal{X}}, \mu) \to (\mathcal{X}, \mathcal{B}_{\mathcal{X}})$ and $X(x) = x$. Due to this, we are going to use the terms probability measure and distribution as synonyms.

2.1. Preliminaries

In this section, we will collect most of the definitions of function classes needed in the subsequent chapters. Furthermore, we are going to fix the notion of the Fourier transform, the moment generating function and the cumulant generating function used in this thesis. Since both concepts are ubiquitously used, we refrain from further comments on the literature.

2.1.1. Function classes

Given a function $f\colon \Omega \to \mathcal{Y}$, where $\Omega \subset \mathcal{X}$, the supremum norm and the L_1-norm are defined by

$$\|f\|_{\infty,\Omega} \coloneqq \operatorname*{ess\,sup}_{x\in\Omega} |f(x)| \quad \text{and} \quad \|f\|_{1,\Omega} \coloneqq \int_{\Omega} |f(x)| \, \mathrm{d}x. \tag{2.1}$$

In case the domain Ω is clear, we frequently omit it.

Lipschitz and Hölder functions In case $\mathcal{X}$ is equipped with $|\cdot|_q$, we denote the set of bounded Lipschitz functions by

$$\mathrm{Lip}(L, B, \Omega) \coloneqq \Big\{ f \colon \Omega \to \mathbb{R} \;\Big|\; \|f\|_{\infty,\Omega} \leq B, \;\; \frac{|f(x) - f(y)|}{|x - y|_q} \leq L, x, y \in \Omega \Big\}. \tag{2.2}$$

The set of unbounded Lipschitz functions is abbreviated by $\mathrm{Lip}(L, \Omega) \coloneqq \mathrm{Lip}(L, \infty, \Omega)$. By Rademacher's theorem (Evans, 2010, Theorem 6), a Lipschitz function is differentiable almost everywhere. For $\alpha \in (0, 1]$ we define the α-Hölder norm by

$$\|f\|_{\mathcal{H}^\alpha(\Omega)} \coloneqq \max \Big\{ \|f\|_\infty, \operatorname*{ess\,sup}_{x,y \in \Omega} \frac{|f(x) - f(y)|}{|x - y|_q^\alpha} \Big\} \tag{2.3}$$

and the α-Hölder ball of functions with Hölder constant $\Gamma > 0$ as

$$\mathcal{H}^\alpha(\Gamma, \Omega) \coloneqq \big\{ f \colon \Omega \to \mathbb{R} \;\big|\; \|f\|_{\mathcal{H}^\alpha(\Omega)} \leq \Gamma \big\}. \tag{2.4}$$

In particular, $\mathrm{Lip}(L, B, \Omega) \subseteq \mathcal{H}^\alpha(\max(L, 2B), \Omega)$ for any $\alpha \in (0, 1)$.

Besov spaces In case $\mathcal{Y} = \mathbb{R}$, $\alpha \in (0, 1]$ and $M \in \mathbb{R}_{>0}$, the Besov ball $B^\alpha_{1,\infty}(M, \Omega)$ is defined as

$$B^\alpha_{1,\infty}(M, \Omega) \coloneqq \big\{ f \in L_1(\Omega) \colon |f|_{B^\alpha_{1,\infty}(\Omega)} < M \big\}, \tag{2.5}$$

where

$$|f|_{B^\alpha_{1,\infty}(\Omega)} \coloneqq \sup_{t>0} t^{-\alpha} \omega_1(f, t)_1, \quad \omega_1(f, t)_1 \coloneqq \sup_{0<|h|\leq t} \int |f(x) - f(x + h)| \, \mathrm{d}x.$$

If $x + h \notin \Omega$, the integrand is taken to be zero. Replacing $|f|_{B^\alpha_{1,\infty}(\Omega)} < M$ with $|f|_{B^\alpha_{1,\infty}(\Omega)} < \infty$ in (2.5), we obtain the Besov space $B^\alpha_{1,\infty}(\Omega)$. The corresponding norm on $B^\alpha_{1,\infty}(\Omega)$ is

$$\|f\|_{B^\alpha_{1,\infty}(\Omega)} \coloneqq \|f\|_{L_1(\Omega)} + |f|_{B^\alpha_{1,\infty}(\Omega)}. \tag{2.6}$$

Smooth function classes $C^k(\Omega)$ denotes the set of functions whose component functions are k-times continuously differentiable with bounded derivatives in the supremum norm. For a multi-index $k \in \mathbb{N}_0^{\dim(\mathcal{X})}$ with $|k|_1 = \sum_{i=1}^{\dim(\mathcal{X})} k_i$, we write

$$D^k \coloneqq \frac{\partial^{|k|_1}}{\partial x_1^{k_1} \cdots \partial x_d^{k_{\dim(\mathcal{X})}}}$$

and for $\beta \in \mathbb{N}$ we define

$$\|f\|_{C^\beta} \coloneqq \max_{k \colon |k| \leq \beta} \|D^k f\|_\infty. \tag{2.7}$$

Sobolev spaces For Section 2.4, we are going to need Sobolev spaces. The set of locally integrable functions is given by

$$L^1_{\mathrm{loc}}(\Omega) \coloneqq \Big\{ f \colon \Omega \to \mathbb{R} \;\Big|\; \int_K |f(x)| \, \mathrm{d}x < \infty, \text{ for all compact } K \subset \Omega^\circ \Big\}.$$

A function $f \in L^1_{\text{loc}}(\Omega)$ has a weak α-th derivative, $D^\alpha_w f$ for a multi index $\alpha \in \mathbb{N}_0^{\dim(\mathcal{X})}$, provided there exists a function $g \in L^1_{\text{loc}}(\Omega)$ such that

$$\int_\Omega g(x)\phi(x)dx = (-1)^{|\alpha|} \int_\Omega f(x) D^\alpha \phi(x) dx \quad \text{for all } \phi \in C^\infty(\Omega) \text{ with compact support.}$$

If such a g exists, we define $D^\alpha_w f \coloneqq g$. For $f \in L^1_{\text{loc}}(\Omega)$ and $k \in \mathbb{N}_0$ the Sobolev norm is

$$\|f\|_{W^{k,\infty}(\Omega)} \coloneqq \max_{|\alpha| \leq k} \|D^\alpha_w f\|_{\infty,\Omega}\,, \tag{2.8}$$

with semi-norm

$$|f|_{W^{k,\infty}(\Omega)} \coloneqq \max_{|\alpha| = k} \|D^\alpha_w f\|_{\infty,\Omega}\,.$$

The Sobolev space

$$W^{k,\infty}(\Omega) \coloneqq \{f \in L^1_{\text{loc}}(\Omega) : \|f\|_{W^{k,\infty}(\Omega)} < \infty\} \tag{2.9}$$

is a Banach space (Brenner & Scott, 2008, Theorem 1.3.2).
Note that $\text{Lip}(L, B, \Omega) \subset W^{1,\infty}(\Omega)$, since $\|f\|_{W^{1,\infty}} \leq \max(L, B)$ for any $f \in \text{Lip}(L, B, \Omega)$.

2.1.2. Fourier transform and related objects

Let $f\colon \mathcal{X} \to \mathbb{R}$ be a such that $\|f\|_1 < \infty$. Then we define the Fourier transform of f as

$$\mathcal{F}f(u) \coloneqq \int e^{\mathrm{i}\langle x,u\rangle} f(x)\, \mathrm{d}x, \tag{2.10}$$

where i is the imaginary unit. Compared to the classical definition, we neglect the constant and invert the sign.
The characteristic function of a distribution μ on $\mathcal{X}$ is defined as

$$\varphi_\mu(u) \coloneqq \mathbb{E}_{X\sim\mu}\big[e^{\mathrm{i}\langle X,u\rangle}\big]. \tag{2.11}$$

In case μ admits a density p with respect to the Lebesgue measure, then

$$\varphi_\mu(u) = \mathcal{F}p(u),$$

which explains our deviation from the classical definition of the Fourier transform.
The moment generating function of a distribution μ on $\mathcal{X}$ is defined as

$$M_\mu(t) \coloneqq \mathbb{E}_{X\sim\mu}\big[e^{\langle X,t\rangle}\big], \tag{2.12}$$

if the term on the right hand side exists for $t \in (-h, h)$ for some $h > 0$. In this case, for a multiindex $\ell \in \mathbb{N}_0^{\dim(\mathcal{X})}$

$$D^\ell M_\mu(0) = \mathbb{E}_{X\sim\mu}[X_1^{\ell_1} \cdot \ldots \cdot X_{\dim(\mathcal{X})}^{\ell_{\dim(\mathcal{X})}}].$$

The cumulant generating function is defined as the logarithm of the moment generating function,

$$K_\mu(t) = \log\left(\mathbb{E}_{X\sim\mu}\left[e^{\langle X,t\rangle}\right]\right). \tag{2.13}$$

The ℓ-th cumulant is defined as

$$\kappa_\ell \coloneqq D^\ell K_\mu(0).$$

In case $\dim(\mathcal{X}) > 1$, the cumulant is also called the joint cumulant.

2.2. Distances between probability measures

To quantify how close a generated distribution is to the distribution that should be imitated, we need to choose a distance (not necessarily a metric) between the two probability measures. In the following, we introduce some examples of distances between probability measures and connect them if possible.

2.2.1. Wasserstein distance

Of high importance to this thesis is the Wasserstein distance, it will be used as both an evaluation metric and as an underlying concept to some of the models we are going to study.

Definition 2.1. *(Wasserstein distances) Let d be a metric on $\mathcal{X}$, and let $p \in [1,\infty)$. For two probability measures μ,ν on $\mathcal{X}$,*

$$\mathsf{W}_p(\mu,\nu) \coloneqq \Big(\inf_{\pi\in\Pi(\mu,\nu)} \int_{\mathcal{X}} d(x,y)^p \,\mathrm{d}\pi(x,y)\Big)^{1/p} \tag{2.14}$$

is called the Wasserstein distance of order p between μ and ν. $\Pi(\mu,\nu)$ denotes the set of all joint probability measures π on $\mathcal{X}$ with marginals μ and ν.

The Wasserstein distance is the optimal value of an optimal transport problem: if the metric is interpreted as a cost function, then the Wasserstein distance is the minimal cost necessary to transport the mass of measure μ to ν. The metric d can be replaced by other suitable cost functions and the Euclidean setting can be generalized to Polish metric spaces. An example will be presented in context of Definition 2.3. Since both generalizations are beyond the needs for this thesis, we refer to Villani (2008) for a comprehensive analysis of the fascinating field of optimal transport.
In the following, we present properties of the Wasserstein distance which are of significance to this thesis.

Metric As shown by Villani (2008, p. 94), the Wasserstein distance satisfies the axioms of a metric, i.e. for $p \in [1,\infty)$

$$\begin{aligned}
&\mathsf{W}_p(\mu,\nu) = \mathsf{W}_p(\nu,\mu), && \text{(symmetry)},\\
&\mathsf{W}_p(\mu,\nu) \le \mathsf{W}_p(\mu,\rho) + \mathsf{W}_p(\rho,\nu), && \text{(triangle inequality)},
\end{aligned}$$

$$\mathsf{W}_p(\mu,\nu) \geq 0, \mathsf{W}_p(\mu,\nu) = 0 \iff \mu = \nu, \qquad \text{(definiteness)}.$$

On the space of probability measures with finite p-th moments, W_p is finite. Combined with the above, W_p is a metric (Villani, 2008, p. 95).

Strength Additionally, W_p metrizes weak convergence in the space of probability measures with finite p-th moment (Villani, 2008, Theorem 6.9). This means that if $(\mu_k)_{k\in\mathbb{N}}$ is a sequence of probability measures with finite p-th moment and μ is another measure with finite p-th moment, then

$$(\mu_k)_{k\in\mathbb{N}} \text{ converges weakly to } \mu \quad \iff \quad \mathsf{W}_p\left(\mu_k, \mu\right) \overset{k\to\infty}{\longrightarrow} 0.$$

Duality In case of $p = 1$ the Kantorovich duality (Villani, 2008, Theorem 5.10) leads to the following useful dual representation of W_1: For any μ, ν with finite first moment we have

$$\mathsf{W}_1(\mu,\nu) = \sup_{W\in\mathrm{Lip}(1)} \mathbb{E}_{\substack{X\sim\mu \\ Y\sim\nu}}\big[W(X) - W(Y)\big]. \tag{2.15}$$

In case of discrete measures on $\mathcal{X}$, this is exactly the same duality as the duality in linear programs. The following illustration is inspired by Solomon (2018).
Consider discrete probability measures $\mu = \sum_{i=1}^{k_1} \delta_{x_i} v_i$ on $\mathcal{Z} = \{x_1, ..., x_{k_1}\} \subset \mathcal{X}$ and $\nu = \sum_{i=1}^{k_2} \delta_{y_i} w_i$ on $\mathcal{W} = \{y_1, ..., y_{k_2}\} \subset \mathcal{X}$ with $v_i, w_i \geq 0$, $\sum_{i=1}^{k_1} v_i = \sum_{i=1}^{k_2} w_i = 1$. For the cost of transporting mass from x_i to y_j, we equip $\mathcal{X}$ with the q-norm $|\cdot|_q$ and define for $i \in \{1, ..., k_1\}$ and $j \in \{1, ..., k_2\}$

$$c_{ij} \coloneqq |x_i - y_j|_q.$$

Let T_{ij} be the total amount of mass that should be transported from x_i to y_j and denote

$$T \coloneqq \begin{pmatrix} T_{11} & \cdots & T_{1k_2} \\ \vdots & \ddots & \vdots \\ T_{k_1 1} & \cdots & T_{k_1 k_2} \end{pmatrix}.$$

This is a linear optimization problem:

$$\begin{aligned} (LP): \min_{T\in\mathbb{R}^{k_1\times k_2}} & \sum_{i=1}^{k_1}\sum_{j=1}^{k_2} T_{ij} c_{ij} \\ \text{s.t. } & T_{ij} \geq 0, && \forall\, i, j, \\ & \sum_{j=1}^{k_2} T_{ij} = v_i, && \forall\, i \in \{1, ..., k_1\}, \\ & \sum_{i=1}^{k_1} T_{ij} = w_j, && \forall\, j \in \{1, ..., k_2\}. \end{aligned}$$

Define the $(k_1 + k_2) \times k_1 k_2$-dimensional matrix

$$A \coloneqq \begin{pmatrix} 1 & 0 & \dots & 0 & 1 & 0 & \dots & 0 & \dots & 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 & 0 & 1 & \dots & 0 & \dots & 0 & 1 & \dots & 0 \\ \vdots & & \ddots & & & & \ddots & & & & & \ddots & \vdots \\ 0 & 0 & \dots & 1 & 0 & 0 & \dots & 1 & \dots & 0 & 0 & \dots & 1 \\ -1 & -1 & \dots & -1 & 0 & 0 & \dots & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & -1 & -1 & \dots & -1 & \dots & 0 & 0 & \dots & 0 \\ \vdots & & & & & & & & & & & & \vdots \\ 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & \dots & -1 & -1 & \dots & -1 \end{pmatrix}$$

and

$$t \coloneqq \begin{pmatrix} T_{11} \\ \vdots \\ T_{k_1 1} \\ T_{21} \\ \vdots \\ T_{k_1 k_2} \end{pmatrix}, \quad c \coloneqq \begin{pmatrix} c_{11} \\ \vdots \\ c_{k_1 1} \\ c_{21} \\ \vdots \\ c_{k_1 k_2} \end{pmatrix}, \quad u \coloneqq \begin{pmatrix} v_1 \\ \vdots \\ v_{k_1} \\ -w_1 \\ \vdots \\ -w_{k2} \end{pmatrix}.$$

Then we can reformulate the linear optimization problem (LP) and the corresponding dual problem as

$$\begin{aligned} &\min_{t \in \mathbb{R}^{k_1 \cdot k_2}_{\geq 0}} c^\top t && \max_{z \in \mathbb{R}^{k_1+k_2}} u^\top z \\ &\text{s.t. } At = u, && \text{s.t. } z^\top A \leq c^\top. \end{aligned}$$

For an introduction to the duality of linear programming, we refer to Nickel et al. (2022, Kapitel 1.8). The restrictions of the dual problem read as

$$z_i - z_{k_1+j} \leq |x_i - y_j| \quad \forall i \leq k_1, \quad \forall j \leq k_2,$$

the objective function extends to

$$\max_{z \in \mathbb{R}^{k_1+k_2}} \sum_{i=1}^{k_1} v_i z_i - \sum_{j=1}^{k_2} w_j z_{k_1+j}.$$

Let $h \colon \mathcal{X} \to \mathbb{R}$ be a function such that

$$\begin{aligned} z_1 = h(x_1), ..., z_{k_1} = h(x_{k_1}), \\ z_{k_1+1} = h(y_1), ..., z_{k_1+k_2} = h(y_{k_2}). \end{aligned}$$

Then the dual problem can be stated as

$$\begin{aligned} &\max_h \mathbb{E}_{\substack{X \sim \mu \\ Y \sim \nu}}[h(X) - h(Y)] \\ &\text{s.t. } h(x_1) - h(y_1) \leq |x_1 - y_1| \end{aligned}$$

$$\vdots$$
$$h(x_{k_1}) - h(y_{k_2}) \leq |x_{k_1} - y_{k_2}|.$$

This is exactly the Lipschitz condition of (2.15) on the relevant set of points in the discrete case.

Closed form For $d = 1$ the Wasserstein-1 distance has the following closed form representation.

Lemma 2.2. *Let $\mathcal{X} = \mathbb{R}$. Further let F, G be the cumulative distribution functions of μ, ν and F^{-1}, G^{-1} their quantile functions. Then*

$$\mathsf{W}_1(\mu, \nu) = \int_{-\infty}^{\infty} |F(x) - G(x)| \, \mathrm{d}x = \int_0^1 \left|F^{-1}(y) - G^{-1}(y)\right| \, \mathrm{d}y.$$

Proof. For the first equality, we refer to Santambrogio (2015, Theorem 2.9). For the second equality, we note that

$$\begin{aligned}\{(x, y) \in \mathbb{R}^2 \colon \min(F(x), G(x)) \leq y \leq \max(F(x), G(x))\} \\ = \{(x, y) \in \mathbb{R}^2 \colon \min(F^{-1}(y), G^{-1}(y)) \leq x \leq \max(F^{-1}(y), G^{-1}(y))\}\end{aligned}$$

and thus the definition of a cumulative distribution function implies

$$\begin{aligned}\int_{-\infty}^{\infty} |F(x) - G(x)| \, \mathrm{d}x &= \int_0^1 \int_{\min(F(x), G(x))}^{\max(F(x), G(x))} \mathrm{d}y \, \mathrm{d}x \\ &= \int_{\infty}^{\infty} \int_{\min(F^{-1}(y), G^{-1}(y))}^{\max(F^{-1}(y), G^{-1}(y))} \mathrm{d}x \, \mathrm{d}y \\ &= \int_0^1 \left|F^{-1}(y) - G^{-1}(y)\right| \, \mathrm{d}y. \qquad \square\end{aligned}$$

For $d \geq 2$ there is, to the author's best knowledge, no closed form of the Wasserstein-1 distance. However, the closed form in Lemma 2.2 gave rise to the sliced Wasserstein distance (Bonnotte, 2013). For a moment, this distance was thought to be equivalent (Bayraktar & Guo, 2021), however, the proof turned out to be false in case of $d \geq 2$ (Bayraktar & Guo, 2024).

2.2.2. Total variation distance

Next, we are going to introduce the total variation distance between probability measures, which behaves differently compared to the Wasserstein distance.

Definition 2.3. *(Total variation) Let μ, ν probability measures on $\mathcal{X}$ and $\mathcal{B}_{\mathcal{X}}$ be the Borel σ-algebra on $\mathcal{X}$. The total variation distance between μ and ν is defined as*

$$\mathsf{TV}(\mu, \nu) \coloneqq \sup_{A \in \mathcal{B}_{\mathcal{X}}} |\mu(A) - \nu(A)|.$$

Note that the properties of a metric can easily be checked. As apparent from the definition, the total variation distance takes values in $[0, 1]$. Additionally, if μ and ν are singular measures, then

$\mathsf{TV}(\mu,\nu) = 1$. If the metric d in Definition 2.1 is replaced by the function $\mathbb{1}_{x\neq y}$, then, as noted by Villani (2008, p. 972),

$$\mathsf{TV}(\mu,\nu) = \inf_{\pi\in\Pi(\mu,\nu)} \int_{\mathcal{X}} \mathbb{1}_{\{x\neq y\}} \,\mathrm{d}\pi(x,y), \tag{2.16}$$

where again $\Pi(\mu,\nu)$ denotes the set of all joint probability measures π on $\mathcal{X}$ with marginals μ and ν. This already gives a intuition in what sense the total variation distance behaves differently than the Wasserstein distance: while (2.16) can only detect non-coinciding areas of mass, (2.14) incorporates differences quantitatively.
In case μ and ν have densities p_μ and p_ν with respect to the Lebesgue measure, Scheffé's lemma shows that

$$\mathsf{TV}(\mu,\nu) = \frac{1}{2}\int |p_\mu(x) - p_\nu(x)| \,\mathrm{d}x. \tag{2.17}$$

2.2.3. Kullback-Leibler divergence and the Jensen-Shannon divergence

Not all distances between probability measures are metrics. One of the most famous examples is the Kullback-Leibler divergence.

Definition 2.4. *(Kullback-Leibler divergence) Let μ and ν be probability measures on $\mathcal{X}$. The Kullback-Leibler divergence of μ from ν is defined as*

$$\mathsf{KL}(\mu \mid \nu) \coloneqq \begin{cases} \int \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right) \mathrm{d}\mu, & \text{if } \mu \ll \nu, \\ +\infty, & \text{otherwise,} \end{cases}$$

where $\frac{d\mu}{d\nu}$ denotes the Radon-Nikodym density of μ with respect to ν.

From the definition we see that the Kullback-Leibler divergence is not symmetric, does not satisfy the triangle inequality and is hence not a metric. The Kullback-Leibler divergence can be connected to the classical maximum likelihood estimation. Assume we observe n i.i.d. samples $X_1, ..., X_n$ from a distribution μ on $\mathcal{X}$. Further assume we have a parametric family $\{p_\vartheta \mid \vartheta \in \Theta\}$, where Θ is a parameter space which is such that the maximal argument exists in the following definition. The maximum likelihood estimator $\hat{\vartheta}$ is chosen such that

$$\hat{\vartheta} \in \arg\max_{\vartheta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \log(p_\vartheta(X_i)). \tag{2.18}$$

Denote the distribution associated to an element of the parametric family by μ_ϑ. Assume that $\{p_\vartheta \mid \vartheta \in \Theta\}$ is such that the Kullback-Leibler divergence of μ from μ_ϑ is finite for all $\vartheta \in \Theta$. Then

$$\mathsf{KL}(\mu \mid \mu_\vartheta) = \int \log\Big(\frac{p_\mu(x)}{p_\vartheta(x)}\Big) p_\mu(x) \,\mathrm{d}x = \int \log(p_\mu(x)) p_\mu(x) \,\mathrm{d}x - \mathbb{E}_{X\sim\mu}[\log(p_\vartheta(X))].$$

The last term is the population counterpart of (2.18). Further, we can rearrange

$$\mathbb{E}_{X\sim\mu}[\log(p_\vartheta(X))] = \int \log(p_\mu(x)) p_\mu(x) \,\mathrm{d}x - \mathsf{KL}(\mu \mid \mu_\vartheta).$$

As the first term on the right hand side is independent of ϑ, the following two sets are, in case of existence, the same

$$\arg\max_{\vartheta\in\Theta} \mathbb{E}_{X\sim\mu}[\log(p_\vartheta(X))] = \arg\min_{\vartheta\in\Theta} \mathsf{KL}(\mu \mid \mu_\vartheta).$$

Albeit the Kullback-Leibler divergence itself is not symmetric, we can construct a symmetric version.

Definition 2.5. *(Jensen-Shannon divergence) Let μ and ν be probability measures on $\mathcal{X}$. The Jensen-Shannon divergence between μ and ν is defined as*

$$\mathsf{JS}(\mu,\nu) \coloneqq \frac{1}{2}\mathrm{KL}\Big(\mu\Big|\,\frac{\mu+\nu}{2}\Big) + \frac{1}{2}\mathrm{KL}\Big(\nu\Big|\,\frac{\mu+\nu}{2}\Big).$$

The Jensen-Shannon divergence takes values in $[0,\log(2)]$, which in case μ and ν admit densities with respect to the Lebesgue measure follows from $\frac{p_\mu(x)}{p_\mu(x)+p_\nu(x)} \leq 1$ for all $x \in \mathcal{X}$. This maximal value is attained if μ and ν are singular, since

$$\begin{aligned}\mathsf{JS}(\mu,\nu) &= \frac{1}{2}\mathrm{KL}\Big(\mu\Big|\,\frac{\mu}{2}\Big) + \frac{1}{2}\mathrm{KL}\Big(\nu\Big|\,\frac{\nu}{2}\Big)\\ &= \frac{1}{2}\int \log(2)\,\mathrm{d}\mu + \frac{1}{2}\int \log(2)\,\mathrm{d}\nu = \log(2).\end{aligned} \tag{2.19}$$

Additionally, the square root of the Jensen-Shannon divergence is a metric (Endres & Schindelin, 2003, p. 1859).

2.2.4. Illustrative comparison of the distances

In this section, we illustratively compare the metrics defined in Section 2.2.1, Section 2.2.2 and Section 2.2.3 in case $\mathcal{X} = \mathbb{R}$. For a first illustration, we plot

$$x \mapsto d(\mathcal{U}[0,1], \mathcal{U}[x,x+1]), \quad x \in [0,2], d \in \{\mathsf{W}_1, \mathsf{TV}, \mathsf{JS}\}.$$

Using the quantile function representation of the Wasserstein-1 distance from Lemma 2.2, the evaluations are straightforward and summarized in Figure 2.1.
For $x > 1$, neither TV nor JS captures the distance between the support of the two uniform distributions.
While for $x \leq 1$ the behavior of all three distances is the same, this cannot be generalized as the second illustration shows in this situation. From comparing Lemma 2.2 to (2.17) and Definition 2.5, we see that the Wasserstein-1 distance operates on the level of differences in the cumulative distribution function, while the total variation distance and the Jensen-Shannon distance operate on the density level. In case of equal support, this can be seen directly when considering a kinked adjustment of the first example. To this end, we define the following adaptation of a triangle wave with slope ± 2

$$\Lambda(t) = (-1)^{\lfloor t\rfloor}\Big(1 - 2\Big|(t-\lfloor t\rfloor) - \frac{1}{2}\Big|\Big).$$

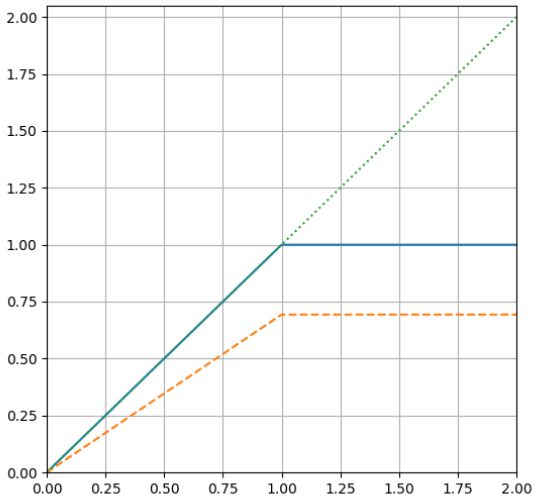


Figure 2.1.: Different behavior evaluating $x \mapsto d(\mathcal{U}[0,1], \mathcal{U}[x, x+1])$ for $d \in \{\mathsf{W}, \mathsf{TV}, \mathsf{JS}\}$ for different values of x. The green dots correspond to $d = \mathsf{W}_1$, the blue line correspond to $d = \mathsf{TV}$ and the orange dashed line corresponds to $d = \mathsf{JS}$.

Then define for $n \in \mathbb{N}$ and $c \in \left(0, \frac{1}{2(n+1)}\right]$ the functions

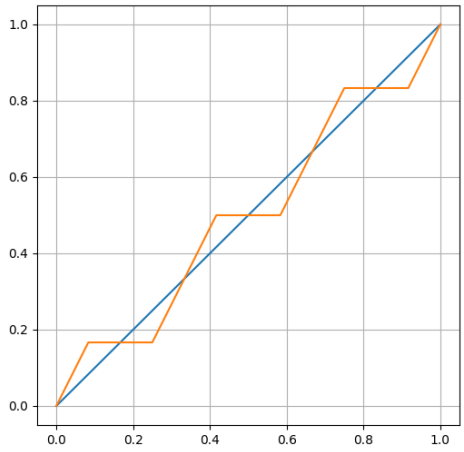
$$F_n(x) = \begin{cases} 0, & x \leq 0, \\ x + c \cdot \Lambda((n+1)x), & 0 < x < 1, \\ 1, & x \geq 1, \end{cases} \quad \text{and} \quad F(x) = \begin{cases} 0, & x \leq 0, \\ x, & 0 < x < 1, \\ 1, & x \geq 1. \end{cases}$$

From the properties of F_n we conclude that there is a distribution μ_n such that F_n is its cumulative distribution function. The function F is the cumulative distribution function of $\mathcal{U}[0,1]$. Figure 2.2 shows the two functions for example values of n.
For the Wasserstein-1 distance between the two distributions, we calculate

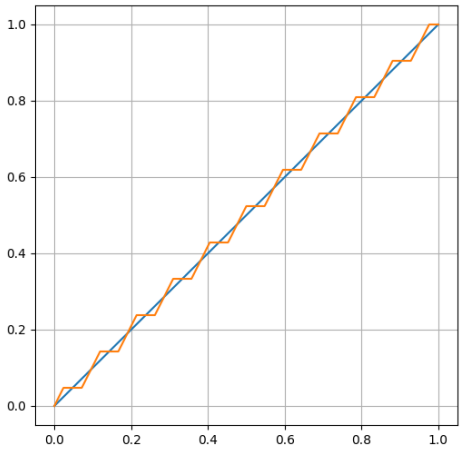
$$\begin{aligned} \mathsf{W}_1(\mu_n, \mathcal{U}[0,1]) &= \int_{-\infty}^{\infty} |F_n(x) - F(x)| \, \mathrm{d}x = c \int_0^1 |\Lambda((n+1)x)| \, \mathrm{d}x \\ &= \frac{c}{n+1} \int_0^{n+1} |\Lambda(t)| \, \mathrm{d}t = \frac{c}{2} \leq \frac{c}{4(n+1)}. \end{aligned}$$

For $n \to \infty$, clearly $\mathsf{W}_1(\mu_n, \mathcal{U}[0,1]) \to 0$, in line with the weak convergence of the two distributions.
For the total variation norm, we first calculate the densities, which exist everywhere except for the set of kink points. Thus for $k \in \{0, ..., n+1\}$ and $x \in \left(\frac{k}{n+1}, \frac{k+\frac{1}{2}}{n+1}\right) \cup \left(\frac{k+\frac{1}{2}}{n+1}, \frac{k+1}{n+1}\right)$,

$$f_n(x) \coloneqq F_n'(x) = \begin{cases} 1 + 2c(-1)^k(n+1), & x \in \left(\frac{k}{n+1}, \frac{k+\frac{1}{2}}{n+1}\right), \\ 1 - 2c(-1)^k(n+1), & x \in \left(\frac{k+\frac{1}{2}}{n+1}, \frac{k+1}{n+1}\right). \end{cases}$$



(a) $n = 5$, $c = \frac{1}{2(n+1)}$



(b) $n = 20$, $c = \frac{1}{2(n+1)}$

Figure 2.2.: F (blue) and F_n (orange) for different values of n.

For the total variation distance, we obtain for fixed $c = \frac{m}{2(n+1)}, m \in (0, 1]$,

$$\mathsf{TV}(\mu_n, \mathcal{U}[0,1]) = \frac{1}{2}\int_0^1 |f_n(x) - 1| \, \mathrm{d}x = c(n+1) = \frac{m}{2}.$$

Hence, $\mathsf{TV}(\mu_n, \mathcal{U}[0,1])$ is independent of n and does not converge to 0 for $n \to \infty$. The comparison between the Wasserstein-1 distance and the Jensen-Shannon divergence was part of the author's masters thesis. Due to examination regulations, it cannot be presented in this thesis. The calculations are more complex but follow the same line. Unsurprisingly, the Jensen-Shannon divergence shows the same behavior as the total variation norm.
This second example shows that even for distributions with the same support, a careful choice of the evaluation distance is of high importance.

2.3. Proper scoring rules

In probabilistic forecasting, a different, though not entirely unrelated, approach to evaluating distributions is taken and well established in the literature, see (Gneiting & Raftery, 2007).
Let $\mathcal{P}$ be a convex set of probability distributions on $\mathcal{X}$, that is for any $\mu_1, \mu_2 \in \mathcal{P}$ and any $\lambda \in [0, 1]$

$$\lambda\mu_1 + (1-\lambda)\mu_2 \in \mathcal{P}.$$

Then a *proper scoring rule* is a measurable function that is quasi-integrable in the second argument with respect to all distributions in $\mathcal{P}$ and

$$S \colon \mathcal{P} \times \mathcal{X} \to \overline{\mathbb{R}}, \quad \text{such that} \quad \mathbb{E}_{X\sim\mu}[S(\mu, X)] \leq \mathbb{E}_{X\sim\mu}[S(\nu, X)] \text{ for all } \mu, \nu \in \mathcal{P}.$$

Here $\overline{\mathbb{R}}$ is the extended real line $[-\infty, \infty]$. For a definition of quasi-integrable functions, we

refer to Elstrodt (2018, Definition 3.2). A *strictly proper scoring rule* is a proper scoring rule where the inequality holds in a strict sense for all $\mu \neq \nu$. In order to enable comparisons to the previous chapter, we define proper scoring rule via the distribution. Note that in the literature the definition via the cumulative distribution function is also common. Additionally, the orientation is frequently reversed, specifically in Gneiting & Raftery (2007).
Next, we assume that a distribution $\mu \in \mathcal{P}$ is approximated by a distribution $\nu \in \mathcal{P}$ and the approximation is evaluated using the expectation of a proper scoring rule. We can decompose the expected score

$$\mathbb{E}_{X\sim\mu}[S(\nu, X)] = \mathbb{E}_{X\sim\mu}[S(\mu, X)] + \mathbb{E}_{X\sim\mu}[S(\nu, X)] - \mathbb{E}_{X\sim\mu}[S(\mu, X)].$$

We refer to $\mu \mapsto \mathbb{E}_{X\sim\mu}[S(\mu, X)]$ as the *entropy* function and to $(\mu, \nu) \mapsto \mathbb{E}_{X\sim\mu}[S(\nu, X)] - \mathbb{E}_{X\sim\mu}[S(\mu, X)]$ as the *divergence* function. The entropy function depends only on μ, the choice of ν is irrelevant. The divergence function captures the difference between μ and ν and is therefore in line with the distances presented in Section 2.2.
In the following, we are going to present proper scoring rules that are either used for comparison to Section 2.2 or of interest in Chapter 6. Due to that, we restrict this presentation of scoring rules to the continuous case, but the concept is also applicable to categorical variables, see the aforementioned Gneiting & Raftery (2007, Section 3).
In case $\mathcal{X} = \mathbb{R}$, let F be the cumulative distribution function and p be the density with respect to the Lebesgue measure of a distribution $\mu \in \mathcal{P}$.
The *logarithmic score*, dating back to Good (1952, Section 8), is defined for x such that $p(x) > 0$ as

$$S_{\log}(\mu, x) = -\log(p(x)).$$

The divergence function is the Kullback-Leibler divergence, introduced in Definition 2.4. The logarithmic score is strictly proper relative to the class of distributions that admit a Lebesgue density (Gneiting & Raftery, 2007, Section 4.1).
Another score defined via the density is the *Hyvärinen score*, defined implicitly by Hyvärinen (2005). For the definition we assume the density p is twice continuously differentiable and define

$$\mathrm{S}_H(\mu, y) = 2\frac{p''(y)}{p(y)} - \left(\frac{p'(y)}{p(y)}\right)^2.$$

The divergence function for two distributions μ and ν admissible to the Hyvärinen score is given by

$$\mathbb{E}_{X\sim\mu}[S_H(\nu, X)] - \mathbb{E}_{X\sim\mu}[S_H(\mu, X)] = \int \left(\frac{q'(y)}{q(y)} - \frac{p'(y)}{p(y)}\right)^2 q(y)\, \mathrm{d}y,$$

where q is the density of ν with respect to the Lebesgue measure. The Hyvärinen score is proper relative to a class of distributions called *valid*, for the definition we refer to Ehm & Gneiting (2009, Definition 3.1). The divergence function of the Hyvärinen score will appear in context of diffusion models in Section 5.1.2.

The *continuous ranked probability score* (CRPS) is defined as

$$\mathrm{CRPS}(\mu, x) = \int (F(y) - \mathbb{1}\{y \geq x\})^2 \mathrm{d}y = \mathbb{E}_{X\sim\mu}[|X - x|] - \frac{1}{2}\mathbb{E}_{X,X'\overset{i.i.d}{\sim}\mu}[|X - X'|].$$

The CRPS is strictly proper with respect to the class of probability distributions with finite first moment, see Gneiting & Raftery (2007) who give further references for the above equality. The divergence function of the CRPS is given by

$$\mathbb{E}_{X\sim\mu}[\mathrm{CRPS}(\nu, X)] - \mathbb{E}_{X\sim\mu}[\mathrm{CRPS}(\mu, X)] = \int (F(y) - G(y))^2 \, \mathrm{d}y.$$

Thus, the divergence function of the CRPS is another example of distances between one-dimensional probability measures based on integral norms: in Section 2.2 we already saw the Wasserstein-1 distance as the L_1 distance between the cumulative distribution functions and the total variation distance as the scaled L_1 distance between the densities. In Section 2.2.4 we illustrated that distances using integral norms based on densities are too strong to capture weak convergence.

In Chapter 6, we will not limit ourselves to the univariate case, but also consider the case where $\dim(\mathcal{X}) > 1$. In the higher-dimensional case the natural extension of the CRPS is the *energy score*

$$\mathrm{ES}(\mu, x) \coloneqq \mathbb{E}_{X\sim\mu}[|X - x|^\beta] - \frac{1}{2}\mathbb{E}_{X,X'\overset{iid}{\sim}\mu}\big[|X - X'|^\beta\big],$$

where $\beta \in (0, 2)$. The associated divergence function is given by

$$d_{\mathrm{ES}}(\mu, \nu) = \mathbb{E}_{Y\sim\mu, Z\sim\nu}[|Y - Z|^\beta] - \frac{1}{2}\mathbb{E}_{Y,Y'\overset{\mathrm{iid}}{\sim}\mu}[|Y - Y'|^\beta] - \frac{1}{2}\mathbb{E}_{Z,Z'\overset{\mathrm{iid}}{\sim}\nu}[|Z - Z'|^\beta].$$

This is the square root of the energy distance (Székely, 2003) multiplied with the factor 2. By Gneiting & Raftery (2007, Theorem 4), the energy score is strictly proper with respect to the class of probability distributions such that $\mathbb{E}_{X\sim\mu}[|X|^\beta] < \infty$.

Gneiting & Raftery (2007, Section 5.3) have generalized this to the *Fourier score*

$$\mathrm{FS}(\mu, y) = \|\varphi_\mu - e^{i\langle\cdot, y\rangle}\|_\gamma^2, \quad \text{with} \quad \|\psi\|_\gamma^2 := \int_{\mathbb{R}^d} \frac{|\psi(u)|^2}{\|u\|^\gamma} \, \mathrm{d}u, \tag{2.20}$$

where φ_μ is the characteristic function corresponding to μ and $\gamma \geq 0$ as defined in Section 2.1.2. The associated divergence function of the Fourier score is given by

$$d_{\mathrm{FS}}(\mu, \nu) = \int \frac{|\varphi_\mu(u) - \varphi_\nu(u)|^2}{\|u\|^\gamma} \, \mathrm{d}u.$$

For $\gamma = 0$, we get the L_2 distance between the characteristic functions. In case $\gamma = d + \beta$ we obtain the energy score (Székely, 2003, Proposition 2).

2.4. ReLU networks and approximation properties

In the following chapters, we frequently use approximation results on feedforward ReLU networks. In each of the proofs, we will isolate an approximation error of the form

$$\inf_{f_{\text{approx}} \in \mathcal{M}} \|f - f_{\text{approx}}\|_\infty, \tag{2.21}$$

where f is some function intrinsic to the model and $\mathcal{M}$ is a set of functions that serve as candidates for the best approximation of f. In all cases, we are going to need to ensure that the functions in $\mathcal{M}$ satisfy some smoothness assumption.

Theoretically, we could choose an arbitrary set of measurable functions for the set $\mathcal{M}$ that satisfies the smoothness assumption. The proofs are structured such that they can be easily adapted to arbitrary approximation results that provide quantitative bounds on (2.21). In view of (2.21), the set $\mathcal{M}$ should be chosen as large as possible. The state-of-the-art method to employ a large function class is of course the use of neural networks. For a mathematical proof, we additionally need some kind of structure in the set $\mathcal{M}$. We use fully connected feedforward ReLU networks, which enjoy nice approximation properties, see for example Yarotsky (2017), Gühring et al. (2020), Kohler & Langer (2021), Schmidt-Hieber (2020), Suzuki (2019).

In this chapter, we are going to define feedforward ReLU networks precisely, present the result used in Chapter 5 and Chapter 6 and prove the result needed in Chapter 3.

To fix the notation we give a general definition of feedforward neural networks. Let $d, L, N_1, ..., N_L \in \mathbb{N}$. A function $f_{\mathsf{NN}} \colon \mathbb{R}^d \to \mathbb{R}$ is a neural network with L layers and $N_1 + \cdots + N_L$ neurons if it results for an argument $x \in \mathbb{R}^d$ from the following scheme:

$$\begin{aligned} x_0 &:= x, \\ x_l &:= \sigma\left(A_l x_{l-1} + b_l\right), \quad \text{for } l = 1, \dots L-1, \\ f_{\mathsf{NN}}(x) = x_L &:= A_L x_{L-1} + b_L, \end{aligned} \tag{2.22}$$

where for $l \in \{1, ..., L\}$, $A_l \in \mathbb{R}^{N_l \times N_{l-1}}$, $b_l \in \mathbb{R}^{N_l}$ and $\sigma \colon \mathbb{R} \to \mathbb{R}$ is an element-wise applied arbitrary activation function. The number of nonzero weights of the matrices A_ℓ and b_ℓ is given by $\sum_{j=1}^{L} \left(|A_j|_{\ell^0} + |b_j|_{\ell^0}\right)$. In the theoretical results, we focus on the ReLU activation function $\sigma(x) = \max(0, x)$. When talking about the size of a network, we refer to a property that concerns the number of layers, the number of nonzero weight and the number of neurons jointly.

In this section, we denote the number of layers with L. This should not be confused with the Lipschitz constant L omnipresent in Chapter 3.

In practice, there are many more advanced architectures of neural networks. A quite simple example that still aligns with the notation of the feedforward neural network can be seen in the architecture behind Figure 1.1 presented in Appendix A: the generator consists of layers with different activation functions. Another example are deviations from the scheme (2.22).

In order to maintain at least some accordance with the theoretical results, all implementations

in this thesis are restricted to feedforward networks.

2.4.1. Approximation in $W^{s,p}$-norms

In Chapter 5 and Chapter 6 we need to control the Lipschitz constant of the neural network used for the vector field in the Flow Matching model. Additionally, in both chapters, we want to approximate a function that is in C^∞. We recall the following theorem from Gühring et al. (2020).

Theorem 2.6. *(Gühring et al., 2020, Corollary 4.2) Let $d \in \mathbb{N}, n \in \mathbb{N}_{\geq 2}, 1 \leq p \leq \infty, B > 0$, and $0 \leq s \leq 1$. Further, let $f \in W^{n,p}\big((0,1)^d\big)$ and assume $\|f\|_{W^{n,p}((0,1)^d)} \leq B$.*
For any $\varepsilon \in (0, 1/2)$, there is a ReLU neural network $f_{\mathsf{NN},\varepsilon}$ with no more than $\lceil c \cdot \log_2\big(\varepsilon^{-n/(n-s)}\big)\rceil$ layers, $\lceil c \cdot \varepsilon^{-d/(n-s)} \cdot \log_2^2\big(\varepsilon^{-n/(n-s)}\big)\rceil$ nonzero weights and $\lceil c \cdot \varepsilon^{-d/(n-s)} \cdot \log_2^2\big(\varepsilon^{-n/(n-s)}\big)\rceil$ neurons, where $c = c(d, n, p, B, s)$ is a constant, such that

$$\|f_{\mathsf{NN},\varepsilon} - f\|_{W^{s,p}((0,1)^d)} \leq \varepsilon.$$

Note that Gühring et al. (2020) refers to the networks following the scheme (2.22) as standard neural networks. In their main theorem (Gühring et al., 2020, Theorem 4.1), they use networks with skip-connections. This is an example for the previous mentioned deviations from the scheme (2.22). In (2.22), only connections between neighboring layers are feasible. In networks with skip-connections, the activation function is applied to all previous layers, see Gühring et al. (2020, Definition 2.1). As the number of neurons and nonzero weights is of the same magnitude, we sometimes omit the number of neurons in the following chapters.

2.4.2. Approximation in $\mathcal{H}^\alpha$-norms

In Chapter 3, the underlying function f is only Lipschitz continuous. Hence it does not satisfy the smoothness assumption needed for Theorem 2.6. Additionally, we cannot use uniform approximation results, since we need to control at least the Hölder smoothness of the function. Thus, we derive our own approximation result suited for the setting in Chapter 3.

Theorem 2.7. *Let $K, B > 0$, and $0 < \alpha < 1$. Then there are constants B and $c(d, K, \alpha, B)$ with the following properties: For any $\varepsilon \in (0, 1/2)$ and any $f \in \mathrm{Lip}(K, B, (0,1)^d)$, there is a ReLU neural network $f_{\mathsf{NN},\varepsilon}$ with no more than $\lceil c \log_2(\varepsilon^{-\frac{1}{1-\alpha}})\rceil$ layers, $\lceil c\varepsilon^{-\frac{d}{1-\alpha}} \log_2^2(\varepsilon^{-\frac{1}{1-\alpha}})\rceil$ nonzero weights and $\lceil c\varepsilon^{-\frac{d}{1-\alpha}} (\log_2^2(\varepsilon^{-\frac{1}{1-\alpha}}) \vee \log_2(\varepsilon^{-\frac{1}{1-\alpha}}))\rceil$ neurons such that*

$$\big\|f_{\mathsf{NN},\varepsilon} - f\big\|_\infty \leq \varepsilon \qquad \text{and} \qquad f_{\mathsf{NN},\varepsilon} \in \mathcal{H}^\alpha\big(\max(K, 2B) + \varepsilon\big).$$

Compared to Theorem 2.6, we can see that Theorem 2.7 is an extension to the approximation of less smooth functions. While Theorem 2.6 requires $n - s \geq 1$, we allow for a smaller gap between the smoothness of the approximated function and the smoothness of the approximation. For the limit case $\alpha \to 1$, we see that the size of the network diverges for every $\varepsilon \in (0, 1/2)$.

2.5. Functional and concentration inequalities

We focus on the setting needed in Section 5.5, which specifically assumes the continuous setting with μ having a density p_μ with respect to the Lebesgue measure. First we recall some classical concentration inequalities. Afterwards we are going to look at connections between functional inequalities and further concentration properties of distributions.

2.5.1. Concentration inequalities

Of fundamental importance is Markov's inequality.

Theorem 2.8 (Markov's inequality)**.** *Let μ be a distribution on $\mathbb{R}$ and $h\colon \mathbb{R} \to [0,\infty)$ a nondecreasing function. Then*

$$h(a)\mathbb{P}_{X\sim\mu}(X \geq a) \leq \mathbb{E}_{X\sim\mu}[h(X)].$$

Proof. We have that

$$h(a)\mathbb{P}(X \geq a) = \int h(a)\mathbb{1}_{\{x\geq a\}}p_\mu(x)\,\mathrm{d}x \leq \int h(x)\mathbb{1}_{\{x\geq a\}}p_\mu(x)\,\mathrm{d}x \leq \mathbb{E}[h(X)]. \qquad \square$$

If we choose $h(x) = e^{tx}$ for $t > 0$ in Theorem 2.8, we recover the general form of the Chernoff inequality:

$$\mathbb{P}(X \geq a) = \mathbb{P}(e^{tX} \geq e^{ta}) \leq \mathbb{E}[e^{tX}]e^{-ta}.$$

This implies

$$\mathbb{P}(X \geq a) \leq \inf_{t>0} \mathbb{E}[e^{tX}]e^{-ta}.$$

In a statistical context, very classical problems are high-probability bounds on the difference between the (empirical) mean of an i.i.d. sample and the expected value of the underlying distribution, i.e. bounds on

$$\mathbb{P}\Big(\Big|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}[X_i]\Big| > a\Big). \tag{2.23}$$

In case that the underlying distribution is bounded, the following version of a Bernstein inequality provides a concentration inequality.

Theorem 2.9 (Bernstein inequality)**.** *(Vershynin, 2018, Special case of Theorem 2.9.5) Let $X_1, \ldots, X_n$ be i.i.d. random variables with $\mathbb{E}[X_i] = 0$, $\mathrm{Var}(X_i) = \sigma^2$ and $|X_i - \mathbb{E}[X_i]| \leq K$ for all i. Then, for every $t \geq 0$, we have*

$$\mathbb{P}\Big(\Big|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}[X_i]\Big| \geq t\Big) \leq 2\exp\Big(-\frac{nt^2}{2(\sigma^2 + Kt/3)}\Big).$$

A subsequent question arises when considering the setting (2.23) but applied to random variables that have been transformed by functions from some set of functions $\mathcal{G}$. To this end, we first define the covering number of a set of functions.

Definition 2.10. *Let $\mathcal{G}$ be a set of bounded functions of $\mathcal{X}$. For $\tau > 0$, define the covering number of $\mathcal{G}$ with respect to the supremum norm as*

$$\mathcal{N}(\tau, \mathcal{G}, \|\cdot\|_\infty) \coloneqq \min\{N \in \mathbb{N} \,|\exists f_1, \ldots, f_N \in \mathcal{G} \text{ such that } \forall f \in \mathcal{G}, \\ \exists i \in \{1, \ldots, N\} : \|f - f_i\|_\infty \leq \tau\}$$

Using the covering number as a complexity measure of $\mathcal{G}$, we recall the following result.

Theorem 2.11. *(Chen et al., 2023b, Lemma 15) Let $\mathcal{G}$ be a class of functions on $\mathcal{X}$ such that $\|G\|_\infty \leq B$ for a $B > 0$. Let $X_1, \ldots, X_n \in \mathcal{X}$ be i.i.d. random variables. For any $\delta \in (0, 1), a \leq 1$, and $\tau > 0$, we have*

$$\mathbb{P}\left(\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n} g(X_i) - (1+a)\mathbb{E}[g(X)] > \frac{(1+3/a)B}{3n}\log\frac{\mathcal{N}(\tau, \mathcal{G}, \|\cdot\|_\infty)}{\delta} + (2+a)\tau\right) \leq \delta \quad \text{and}$$

$$\mathbb{P}\left(\sup_{g\in\mathcal{G}} \mathbb{E}[g(X)] - \frac{1+a}{n}\sum_{i=1}^{n} g(X_i) > \frac{(1+6/a)B}{3n}\log\frac{\mathcal{N}(\tau, \mathcal{G}, \|\cdot\|_\infty)}{\delta} + (2+a)\tau\right) \leq \delta.$$

We note that the proof of Chen et al. (2023b, Lemma 15) contains some minor typos that can easily be corrected.

2.5.2. Poincaré and log-Sobolev inequalities

We start by defining the Poincaré inequality in a probabilistic continuous setting. This concept of bounding the fluctuation of a function around its mean by the norm of the gradient is of fundamental interest in many areas.

Definition 2.12. *A distribution μ on $\mathcal{X}$ satisfies the Poincaré inequality with Poincaré constant $\rho > 0$ if for all smooth functions f such that the terms below are well-defined*

$$\mathrm{Var}_\mu(f) \coloneqq \mathbb{E}_{X\sim\mu}[(f(X) - \mathbb{E}_{X\sim\mu}[f(X)])^2] \leq \rho\mathbb{E}_{X\sim\mu}[|\nabla f(X)|^2].$$

The Poincaré constant of Gaussian distributions will be of particular importance.

Theorem 2.13. *Let $\mu = \mathcal{N}(a, \sigma^2 I_\mathcal{X})$. Then the Poincaré inequality is fulfilled with Poincaré constant σ^2.*

Proof. From Boucheron et al. (2013, Theorem 3.20) we know that $\mathcal{N}(0, I_\mathcal{X})$ satisfies the Poincaré inequality with Poincaré constant 1. Thus, via defining $g(x) \coloneqq f(\sigma x + a)$ and

$$\mathrm{Var}_\mu(g) \leq \sigma^2\mathbb{E}_{X\sim\mu}[|\nabla g(X)|^2],$$

we see that the Poincaré constant of $\mathcal{N}(a, \sigma^2 I_d)$ is σ^2. □

As apparent from Definition 2.12, the Poincaré constant controls the variance of a function using its expected gradient. A natural extension is the control of higher moments in the same setting. Therefore we are going to define the logarithmic Sobolev inequality.

First, we define the entropy of a function $f\colon \mathcal{X} \to (0,\infty)$ with respect to μ in case $\mathbb{E}_{X\sim\mu}[f(X)\log(f(X))] < \infty$ as

$$\mathrm{Ent}_\mu(f) \coloneqq \mathbb{E}_{X\sim\mu}[f(X)\log(f(X))] - \mathbb{E}_{X\sim\mu}[f(X)]\log(\mathbb{E}_{X\sim\mu}[f(X)]).$$

This definition is closely related to the Kullback-Leibler divergence, also called the relative entropy: assume $\mu \ll \nu$ in Definition 2.4 and note that

$$\mathsf{KL}(\mu \mid \nu) = \int \frac{\mathrm{d}\mu}{\mathrm{d}\nu} \log\big(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\big)\,\mathrm{d}\nu.$$

If we replace $\frac{\mathrm{d}\mu}{\mathrm{d}\nu}$ with a positive function f with $\mathbb{E}_{X\sim\mu}[f(X)\log(f(X))] < \infty$, that is normalized using its integral with respect to ν, then

$$\int \frac{f}{\int f\,\mathrm{d}\nu} \log\big(\frac{f}{\int f\,\mathrm{d}\nu}\big)\,\mathrm{d}\nu = \frac{1}{\int f\,\mathrm{d}\nu}\Big(\int f\log\big(f\big)\,\mathrm{d}\nu - \log\big(\int f\,\mathrm{d}\nu\big)\int f\,\mathrm{d}\nu\Big).$$

This coincides with $\mathrm{Ent}_\mu(f)$ up to normalization by $\int f\,\mathrm{d}\nu$, which explains the similar naming.

Definition 2.14. *A distribution μ on $\mathcal{X}$ with density p_μ satisfies the logarithmic Sobolev inequality with* log*-Sobolev constant $\lambda > 0$ if for all smooth functions f such that the terms below are well-defined*

$$\mathrm{Ent}_\mu(f^2) \le 2\lambda\mathbb{E}_{X\sim\mu}[|\nabla f(X)|^2]. \tag{2.24}$$

The relation of the logarithmic Sobolev inequality to the control over higher order moments will become apparent in the next subsection. Again, are we interested in the log-Sobolev constant of Gaussian distributions.

Theorem 2.15. *Let $\mu = \mathcal{N}(a,\sigma^2 I_{\mathcal{X}})$. Then the logarithmic Sobolev inequality is fulfilled with* log*-Sobolev constant σ^2.*

Proof. From Boucheron et al. (2013, Theorem 5.4) we know that $\mathcal{N}(0, I_{\mathcal{X}})$ satisfies the logarithmic Sobolev inequality with log-Sobolev constant 1. Then we can use the same scaling argument as in the proof of Theorem 2.13. □

Theorem 2.15 is the motivation of the factor 2 appearing on the left hand side of (2.24). As expected, the logarithmic Sobolev inequality implies the Poincaré inequality.

Theorem 2.16. *If μ satisfies a logarithmic Sobolev inequality, then it satisfies a Poincaré inequality.*

Proof. This is a special case of Bakry et al. (2013, Proposition 5.1.3). □

2.5.3. Sub-Gaussian and sub-exponential distributions

Next, we are going to look at distributions whose tails decay at least as fast as the Gaussian or the exponential distribution. We restrict the definitions to the case $\mathcal{X} = \mathbb{R}$, which is sufficient for the results needed in this thesis.

Definition 2.17.

1. *A distribution μ on $\mathbb{R}$ is called sub-Gaussian, if there is a constant $\rho > 0$ such that for $X \sim \mu$ and every $t \geq 0$*
$$\mathbb{P}_{X\sim\mu}(|X| \geq t) \leq 2\exp\big(-\frac{t^2}{\rho^2}\big).$$
2. *A distribution μ on $\mathcal{X}$ is called sub-exponential, if there is a constant $\rho > 0$ such that for $X \sim \mu$ and every $t \geq 0$*
$$\mathbb{P}_{X\sim\mu}(|X| \geq t) \leq 2\exp\big(-\frac{t}{\rho}\big).$$

In both cases, the defining properties can be characterized in several useful ways. We start with the characterizations of sub-Gaussian distributions.

Theorem 2.18. *(Vershynin, 2018, Proposition 2.5.2) Let μ be a distribution on $\mathbb{R}$ and $X \sim \mu$. The following statements are equivalent*

1. *The distribution μ is sub-Gaussian with constant $\rho > 0$.*
2. *For every $q \geq 1$ we have that*
$$(\mathbb{E}_{X\sim\mu}[|X|^q])^{\frac{1}{q}} \lesssim \rho\sqrt{q}.$$
3. *For a constant ρ_1 proportional to ρ, i.e. $\rho_1 \asymp \rho$, it holds that*
$$\mathbb{E}_{X\sim\mu}[\exp(\lambda^2 X^2)] \leq \exp(\rho_1^2\lambda^2), \quad \text{for all } \lambda \text{ such that } |\lambda| \leq \frac{1}{\rho_1}.$$
4. *For a constant ρ_2 proportional to ρ we have that*
$$\mathbb{E}_{X\sim\mu}\Big[\exp\Big(\frac{X^2}{\rho_2^2}\Big)\Big] \leq 2.$$

If additionally $\mathbb{E}[X] = 0$, then the following statement is also equivalent

5. *For a constant ρ_3 proportional to ρ it holds that*
$$\mathbb{E}[\exp(\lambda X)] \leq \exp(\rho_3^2\lambda^2), \quad \text{for all } \lambda \in \mathbb{R}.$$

Furthermore, we can define a norm on the space of sub-Gaussian distributions.

Theorem 2.19. *Let μ be a sub-Gaussian distribution on $\mathbb{R}$ and $X \sim \mu$. Define*
$$\|X\|_{\psi_2} = \inf\big\{t > 0 : \mathbb{E}[\exp\big(X^2/t^2\big)] \leq 2\big\}.$$
Then $\|\cdot\|_{\psi_2}$ is a norm on the space of sub-Gaussian distributions.

Proof. This is a special instance of an Luxemburg-Orlicz norm (Pick et al., 2012, Section 4.8) driven by the function $\psi_2(x) = e^{x^2} - 1$. □

Similarly, sub-exponential distributions can be characterized using the following properties.

Theorem 2.20. *(Vershynin, 2018, Proposition 2.7.1) Let μ be a distribution on $\mathbb{R}$ and $X \sim \mu$. The following statements are equivalent*

1. *The distribution μ is sub-exponential with constant $\rho > 0$.*
2. *For every $q \geq 1$ we have that*
$$(\mathbb{E}_{X\sim\mu}[|X|^q])^{\frac{1}{q}} \lesssim \rho q.$$
3. *For a constant ρ_1 proportional to ρ it holds that*
$$\mathbb{E}_{X\sim\mu}[\exp(\lambda|X|)] \leq \exp(\rho_1\lambda), \quad \text{for all } \lambda \text{ such that } 0 \leq \lambda \leq \frac{1}{\rho_1}.$$
4. *For a constant ρ_2 proportional to ρ we have that*
$$\mathbb{E}_{X\sim\mu}\Big[\exp\Big(\frac{|X|}{\rho_2}\Big)\Big] \leq 2.$$

If additionally $\mathbb{E}[X] = 0$, then the following statement is also equivalent

5. *For a constant ρ_3 proportional to ρ it holds that*
$$\mathbb{E}_{X\sim\mu}[\exp(\lambda X)] \leq \exp(\rho_3^2\lambda^2), \quad \text{for all } \lambda \text{ such that } |\lambda| \leq \frac{1}{\rho_3}.$$

On the space of sub-exponential distributions, we can also define a norm.

Theorem 2.21. *Let μ be a sub-exponential distribution on $\mathbb{R}$ and $X \sim \mu$. Define*

$$\|X\|_{\psi_1} = \inf\big\{t > 0 : \mathbb{E}[\exp(|X|/t)] \leq 2\big\}.$$

Then $\|\cdot\|_{\psi_1}$ is a norm on the space of sub-exponential distributions.

Proof. This is again a special instance of an Luxemburg-Orlicz norm (Pick et al., 2012, Section 4.8) driven by the function $\psi_1(x) = e^x - 1$. □

Sub-Gaussian and sub-exponential distributions are linked in several ways. Two of those links are going to be of relevance for this thesis. First, as apparent from comparing Theorem 2.18 No. 2 to Theorem 2.20 No. 2, a sub-Gaussian random variable is always sub-exponential. Second, the product of two sub-Gaussian random variables is sub-exponential.

Theorem 2.22. *(Vershynin, 2018, Proposition 2.7.6) Let μ and ν be sub-Gaussian distributions and $X \sim \mu$, $Y \sim \nu$. Then XY is sub-exponential and*

$$\|XY\|_{\psi_1} \leq \|X\|_{\psi_2}\|Y\|_{\psi_2}.$$

2.5.4. Log-concave distributions

A common assumption on the unknown distribution, especially in the analysis of diffusion models, is that it is strongly log-concave (Bruno et al., 2025; Gao & Zhu, 2025; Tang & Zhao, 2024). In

this section, we are going to introduce the Brascamp-Lieb inequality, which is one of the main reasons to impose this assumption. We start with the definition of convexity.

Definition 2.23. *Let $f : \mathcal{X} \to \mathbb{R}$ be a twice differentiable function. Then*

1. *we call f* convex *if $H_f(x) \succeq 0$ for all $x \in \mathcal{X}$,*
2. *we call f* strictly convex *if $H_f(x) \succ 0$ for all $x \in \mathcal{X}$,*
3. *we call f* strongly convex with parameter $m > 0$ *if $H_f(x) \succeq mI_d$ for all $x \in \mathcal{X}$.*

Note that requiring $H_f(x) \succ 0$ for all $x \in \mathcal{X}$ implies the classical definition via inequalities of function values of strict convexity, but is not equivalent to it.

Theorem 2.24 (Brascamp-Lieb inequality)**.** *(Brascamp & Lieb, 1976, Theorem 4.1) Let μ be a probability measure on $\mathcal{X}$ with density p_μ with respect to the Lebesgue measure and $p_\mu(x) > 0$ for all $x \in \mathcal{X}$. Further denote $h \coloneqq -\log(p_\mu)$ and assume that h is strictly convex. Then for any differentiable function $g\colon \mathcal{X} \to \mathbb{R}$*

$$\mathrm{Var}_\mu(g) \leq \mathbb{E}_{X\sim\mu}\Big[\langle \nabla g(X), (H_h(X))^{-1}\nabla g(X)\rangle\Big].$$

A distribution that satisfies the assumption of Theorem 2.24 is typically referred to as a strictly log-concave distribution. In case h is not only strictly but strongly convex with parameter m, we call the corresponding distribution strongly log-concave with parameter m. In this case we can use Theorem 2.24 to obtain an immediate bound on the coordinate variances of μ. To this end, we set $g(x) = x_i$ for $i \in \{1, ..., \dim(\mathcal{X})\}$. Then $H_h(x) - mI_d \succeq 0$ for all $x \in \mathcal{X}$, which implies $x^\top H_h(x)^{-1}x \leq \frac{1}{m}|x|^2$ for all $x \in \mathcal{X}$. Choosing e_i yields $(H_h(x)^{-1})_{ii} \leq \frac{1}{m}$ for all $x \in \mathcal{X}$. Hence

$$\mathrm{Var}_{X\sim\mu}(X_i) \leq \frac{1}{m}.$$

2.5.5. Connections

The functional inequalities of Section 2.5.2, and the properties discussed in Section 2.5.3 and Section 2.5.4 can be related to each other. First we relate log-concave distributions to the Poincaré and the logarithmic Sobolev inequality.

Theorem 2.25. *Let μ be a probability measure on $\mathcal{X}$.*

1. *If μ is* log*-concave, then it satisfies the Poincaré inequality.*
2. *If μ is strongly* log*-concave, then it satisfies the logarithmic Sobolev inequality.*

Proof.

1. See Bobkov (1999, Theorem 1.2) or Bakry et al. (2013, Theorem 4.6.3).
2. See Villani (2008, Theorem 21.2, Remark 12.4), referring to the Bakry–Emery theorem of Bakry & Émery (1985). □

The relation of the functional inequalities to the tail behavior properties is restrained to the case $\mathcal{X} = \mathbb{R}$ in this presentation. The following argument is commonly known as Herbst's argument.

Theorem 2.26. *Let μ be a distribution on $\mathbb{R}$. If μ satisfies the logarithmic Sobolev inequality with* log*-Sobolev constant λ, then the centered shift of μ is sub-Gaussian with constant $\tilde{\rho} \asymp \lambda$.*

Proof. In the entire proof we assume $X \sim \mu$. The proof follows along the lines of Ledoux (1999, p. 148). Set $f(x) = e^{\frac{\vartheta x}{2}}$ for a $\vartheta > 0$ in (2.24). Define $H(\vartheta) \coloneqq \mathbb{E}[e^{\vartheta X}]$. Then using the dominated convergence theorem

$$\operatorname{Ent}(f^2) = \mathbb{E}[\vartheta X e^{\vartheta X}] - \mathbb{E}[e^{\vartheta X}] \log\big(\mathbb{E}[e^{\vartheta X}]\big) = \vartheta H'(\vartheta) - H(\vartheta)\log(H(\vartheta)).$$

Further

$$2\lambda \mathbb{E}[|\nabla f(X)|^2] = \frac{\vartheta^2 \lambda}{2} \mathbb{E}[e^{\vartheta X}] = \frac{\vartheta^2 \lambda}{2} H(\vartheta).$$

Thus by (2.24)

$$\vartheta H'(\vartheta) - H(\vartheta)\log(H(\vartheta)) \leq \frac{\vartheta^2 \lambda}{2} H(\vartheta) \quad \iff \quad \frac{1}{\vartheta}\frac{H'(\vartheta)}{H(\vartheta)} - \frac{\log(H(\vartheta))}{\vartheta^2} \leq \frac{\lambda}{2}.$$

Now define $K(\vartheta) \coloneqq \frac{1}{\vartheta}\log(H(\vartheta))$. Then

$$K'(\vartheta) = \frac{1}{\vartheta}\frac{H'(\vartheta)}{H(\vartheta)} - \frac{\log(H(\vartheta))}{\vartheta^2} \leq \frac{\lambda}{2}.$$

As

$$\lim_{\vartheta \to 0} K(\vartheta) = \lim_{\vartheta \to 0} \frac{\mathbb{E}[X e^{\vartheta X}]}{\mathbb{E}[e^{\vartheta X}]} = \lim_{\vartheta \to 0} \frac{M'(\vartheta)}{M(\vartheta)} = \mathbb{E}[X],$$

where M is the moment generating function, we can extend the definition of K to

$$K(\vartheta) \coloneqq \begin{cases} \frac{1}{\vartheta}\log(H(\vartheta)), & \vartheta > 0, \\ \mathbb{E}[X], & \vartheta = 0. \end{cases}$$

Hence

$$K(\vartheta) = K(0) + \int_0^{\vartheta} K'(u)\,\mathrm{d}u \leq \mathbb{E}[X] + \frac{\vartheta\lambda}{2}.$$

Going back to the function H, we get

$$\begin{aligned} & \frac{1}{\vartheta}\log(H(\vartheta)) \leq \mathbb{E}[X] + \frac{\vartheta\lambda}{2} \\ \iff \quad & H(\vartheta) \leq \exp\big(\vartheta\mathbb{E}[X] + \frac{\vartheta^2\lambda}{2}\big) \\ \iff \quad & \mathbb{E}[e^{\vartheta X}] \leq \exp\big(\vartheta\mathbb{E}[X] + \frac{\vartheta^2\lambda}{2}\big). \end{aligned}$$

Now Markov's inequality yields

$$\mathbb{P}(X - \mathbb{E}[X] \geq r) = \mathbb{P}\big(\exp(\vartheta(X - \mathbb{E}[X])) \geq \exp(\vartheta r)\big) \leq \frac{\mathbb{E}[\exp(\vartheta(X - \mathbb{E}[X]))]}{\exp(\vartheta r)}$$

$$= \exp(-\vartheta r)\frac{\mathbb{E}[e^{\vartheta X}]}{e^{\vartheta \mathbb{E}[X]}} \leq \exp(-\vartheta r)\frac{\exp(\vartheta \mathbb{E}[X] + \frac{\vartheta^2 \lambda}{2})}{e^{\vartheta \mathbb{E}[X]}} = e^{\frac{\vartheta^2 \lambda}{2} - \vartheta r}.$$

Setting $f(x) = e^{-\frac{\vartheta X}{2}}$, we obtain analogously

$$\mathbb{P}(X - \mathbb{E}[X] \leq -r) \leq e^{\frac{\vartheta^2 \lambda}{2} - \vartheta r}.$$

Thus a union bound argument and setting $\vartheta = \frac{r}{\lambda}$ leads to

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq r) \leq 2e^{-\frac{r^2}{2\lambda}}. \qquad \square$$

The Poincaré inequality is weaker than the logarithmic Sobolev inequality. Thus, the implied concentration is weaker as well. As Section 5.5 does not rely on concentration implied by Poincaré inequalities, we do not elaborate the dependencies between the constants.

Theorem 2.27. *(Villani, 2008, special case of Theorem 22.30) Let μ be a distribution on $\mathbb{R}$. If μ satisfies the Poincaré inequality, then there is a constant C depending on the Poincaré constant such that*

$$\mathbb{P}(|X| \geq t) \leq 2\exp(-C\min(t, t^2)).$$

Theorem 2.27 shows that the Poincaré inequality leads to sub-Gaussian tail behavior for small t and sub-exponential tail behavior for large t. For $\mathcal{X} = \mathbb{R}$, we summarized the relations needed in this thesis in Figure 2.3.

$$\begin{array}{ccccc}
\mu \text{ strongly log concave} & \Longrightarrow & \mu \text{ satisfies LSI} & \Longrightarrow & \mu \text{ sub-Gaussian} \\
\Downarrow & & \Downarrow & & \Downarrow \\
\mu \text{ log concave} & \Longrightarrow & \mu \text{ satisfies PI} & & \mu \text{ sub-exponential}
\end{array}$$

Figure 2.3.: Overview of the relations in case of $\mathcal{X} = \mathbb{R}$ needed in this thesis. In all cases, we assume for simplicity $\mathbb{E}_{X\sim\mu}[X] = 0$. PI is short for Poincaré inequality, LSI is short for logarithmic Sobolev inequality.

2.6. Conceptual proof of oracle and related inequalities

In this section, we present a conceptual proof of convergence rate in a distribution estimation setting. To this end, let $\mathcal{P}$ be some set of distributions on $\mathcal{X}$. Assume we have an evaluation distance $d\colon \mathcal{P} \times \mathcal{P} \to \mathbb{R}_{\geq 0}$, which satisfies the triangle inequality, and a selection criterion $c\colon \mathcal{P} \times \mathcal{P} \to \mathbb{R}_{\geq 0}$. Further assume we observe n i.i.d. observations $X_1, ..., X_n$ of an unknown distribution μ^*. Based on these observations, we define the empirical distribution

$$\mu_n \coloneqq \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$$

and assume we have a class of potential distribution estimators $\mathcal{P}_\Theta \subset \mathcal{P}$ such that the minimizers below exist.

The empirical risk minimizer of the model associated with c is chosen as

$$\hat{\mu} \in \operatorname*{arg\,min}_{\mu \in \mathcal{P}_\Theta} c(\mu_n, \mu).$$

Our goal is to bound $d(\mu^*, \hat{\mu})$. Let $\tilde{\mu}$ be some reference measure and $\varepsilon_{\text{ref}} \coloneqq d(\mu^*, \tilde{\mu})$ the corresponding reference error. We assume a linear, multiplicative relation between d and c, i.e. there exists a constant $\varepsilon_{d,c} > 0$ such that

$$d(\mu, \nu) \leq \varepsilon_{d,c} c(\mu, \nu), \quad \text{for all } \mu, \nu \in \mathcal{P}.$$

Further, we assume an additive relation between $c(\tilde{\mu}, \cdot)$ and $c(\mu_n, \cdot)$, i.e. for all $\nu \in \mathcal{P}_\Theta$

$$|c(\tilde{\mu}, \nu) - c(\mu_n, \nu)| \leq \varepsilon_c^n.$$

Note that n is an index here. Of course, ε_c^n depends on the reference measure $\tilde{\mu}$. Then for any $\mu \in \mathcal{P}_\Theta$

$$\begin{aligned}
d(\mu^*, \hat{\mu}) &\leq \varepsilon_{\text{ref}} + d(\tilde{\mu}, \hat{\mu}) && \text{(triangle inequality)} \\
&\leq \varepsilon_{\text{ref}} + \varepsilon_{d,c} c(\tilde{\mu}, \hat{\mu}) && \text{(difference of criterion)} \\
&\leq \varepsilon_{\text{ref}} + \varepsilon_{d,c}(\varepsilon_c^n + c(\mu_n, \hat{\mu})) && \text{(stochastic error)} \\
&\leq \varepsilon_{\text{ref}} + \varepsilon_{d,c}(\varepsilon_c^n + c(\mu_n, \mu)) && \text{(empirical risk minimization)} \\
&\leq \varepsilon_{\text{ref}} + \varepsilon_{d,c}(2\varepsilon_c^n + c(\tilde{\mu}, \mu)). && \text{(stochastic error)}
\end{aligned}$$

As this holds for any $\mu \in \mathcal{P}_\Theta$, we conclude

$$d(\mu^*, \hat{\mu}) \leq \varepsilon_{\text{ref}} + \varepsilon_{d,c}(2\varepsilon_c^n + \min_{\mu \in \mathcal{P}_\Theta} c(\tilde{\mu}, \mu)).$$

The error $\varepsilon_{d,c}$ depends only on the relationship between the distance d and the criterion c. The error $\min_{\mu \in \mathcal{P}_\Theta} c(\tilde{\mu}, \mu)$ depends on the ability of the class $\mathcal{P}_\Theta$ to approximate $\tilde{\mu}$ in criterion c. The stochastic error ε_c^n can be controlled using concentration inequalities, such as the ones presented in Section 2.5.1. This error can be avoided when using $\tilde{\mu} = \mu_n$ as a reference distribution.
In this case, the reference error is the distance between the true distribution and the empirical distribution based on $X_1, ..., X_n$. As the empirical distribution cannot reflect the regularity of the distribution μ^*, the use of μ_n as a reference distribution comes at the cost of profiting from this smoothness in the rate.
The presented conceptual proof serves as a starting point; all models studied in this thesis will require adaptations tailored to the corresponding setting.

2.7. Proof of Theorem 2.7

To prove Theorem 2.7, we need some auxiliary results and notation.
Theorem 2.7 is very similar to Gühring et al. (2020, Theorem 4.1), which however applies only

to functions f which are at least twice (weakly) differentiable. Our proof can thus build on numerous auxiliary results and arguments from this previous work. We keep the proof structure of Gühring et al. (2020) which in turn relies on Yarotsky (2017).
Let $d, N \in \mathbb{N}$. For $m \in \{0, ..., N\}^d$, define the functions $\phi_m \colon \mathbb{R}^d \to \mathbb{R}$,

$$\phi_m(x) = \prod_{\ell=1}^{d} \psi\Big(3N\Big(x_\ell - \frac{m_\ell}{N}\Big)\Big), \quad \text{where} \quad \psi(x) = \begin{cases} 1, & |x| < 1, \\ 0, & |x| > 2, \\ 2 - |x|, & 1 \le |x| \le 2. \end{cases}$$

By definition, we have $\|\phi_m\|_\infty = 1$ for all m and

$$\operatorname{supp} \phi_m \subset \Big\{x : \sup_k |x_k - \frac{m_k}{N}| < \frac{1}{N}\Big\} =: B_{\frac{1}{N}, |\cdot|_\infty}(\frac{m}{N}). \tag{2.25}$$

Gühring et al. (2020, Lemma C.3 (iv)) have verified that $\|\phi_m\|_{W^{1,\infty}(\mathbb{R}^d)} \le cN$ for some constant $c > 0$. A direct consequence of Lemma 2.11, Lemma C.3, Lemma C.5 and Lemma C.6 by Gühring et al. (2020) is the following approximation result for the localizing functions ϕ_m via ReLU networks.

Lemma 2.28. *For any $\varepsilon \in (0, 1/2)$ and any $m \in \{0, ..., N\}^d$ there is a network ψ_ε with ReLU activation function, no more than $C_1 \log_2(\varepsilon^{-1})$ layers, no more than $C_2(N+1)^d \log_2^2(\varepsilon^{-1})$ nonzero weights and no more than $C_3(N+1)^d(\log_2^2(\varepsilon^{-1}) \vee \log_2(\varepsilon^{-1}))$ neurons such that for $k \in \{0, 1\}$*

$$\|\psi_\varepsilon - \phi_m\|_{W^{k,\infty}} \le cN^k \varepsilon,$$

where C_1, C_2, C_3 and c are constants independent of m and ε. Additionally,

$$\phi_m(x) = 0 \quad \Longrightarrow \quad \psi_\varepsilon(x) = 0,$$

and therefore $\operatorname{supp} \psi_\varepsilon \subset B_{\frac{1}{N}, |\cdot|_\infty}(\frac{m}{N})$.

Next we approximate a bounded Lipschitz function using linear combinations of the set $\{\phi_m \colon m \in \{1, ..., N\}^d\}$. The approximation error will be measured in the Hölder norm from (2.3).

Lemma 2.29. *Let $0 < \alpha < 1$. There exists a constant $C_1 > 0$ such that for any $f \in W^{1,\infty}((0,1)^d)$ there are constants $c_{f,m}$ for $m \in \{0, ..., N\}^d$ such that*

$$\Big\| f - \sum_{m \in \{0,...,N\}^d} c_{f,m} \phi_m \Big\|_{\mathcal{H}^\alpha} \le C_1 \Big(\frac{1}{N}\Big)^{1-\alpha} \|f\|_{W^{1,\infty}}.$$

The coefficients satisfy for a $C_2 > 0$

$$|c_{f,m}| \le C_2 \|\tilde{f}\|_{W^{1,\infty}(\Omega_{m,N})},$$

where $\Omega_{m,N} \coloneqq B_{\frac{1}{N}, |\cdot|_\infty}(\frac{m}{N})$ and $\tilde{f} \in W^{1,\infty}(\mathbb{R})$ is an extension of f.

Proof. Let $E \colon W^{1,\infty}((0,1)^d) \to W^{1,\infty}(\mathbb{R})$ be the continuous linear extension operator from Stein

(1970, Theorem 5) and set $\tilde{f} \coloneqq Ef$. As E is continuous there exists a $C_E > 0$ such that

$$\|\tilde{f}\|_{W^{1,\infty}(\mathbb{R}^d)} \leq C_E \|f\|_{W^{1,\infty}}.$$

Step 1 (Choice of $c_{f,m}$): For each $m \in \{0, \dots, N\}^d$ we define

$$c_{f,m} \coloneqq \int_{B_{m,N}} \tilde{f}(y)\rho(y)\ \mathrm{d}y \text{ for } B_{m,N} \coloneqq B_{\frac{3}{4N},|\cdot|}\Big(\frac{m}{N}\Big)$$

and an arbitrary cut-off function ρ supported in $B_{m,N}$, i.e.

$$\rho \in C_c^\infty(\mathbb{R}^d) \quad \text{with} \quad \rho(x) \geq 0 \ \text{ for all } x \in \mathbb{R}^d, \quad \operatorname{supp}\rho = B_{m,N} \quad \text{and} \quad \int_{\mathbb{R}^n} \rho(x)dx = 1.$$

Then

$$|c_{m,f}| = \Big| \int_{B_{m,N}} \tilde{f}(y)\rho(y)\ \mathrm{d}y \Big| \leq \|\tilde{f}\|_{\infty,\Omega_{m,N}} \int_{B_{m,N}} \rho(y)\ \mathrm{d}y = \|\tilde{f}\|_{\infty,\Omega_{m,N}} \leq C_E \|f\|_{W^{1,\infty}(\Omega_{m,N})}.$$

Step 2 (Local estimates in $\|\cdot\|_{W^{k,p}}$): The coefficients $c_{m,f}$ are the averaged Taylor polynomials in the sense of Brenner & Scott (2008, Definition 4.1.3) of order 1 averaged over $B_{m,N}$. As Gühring et al. (2020, Proof of Lemma C.4, Step 2) showed, the conditions of the Bramble-Hilbert-Lemma (Brenner & Scott, 2008, Theorem 4.3.8) are satisfied. Hence for $k \in \{0, 1\}$

$$|\tilde{f} - c_{m,f}|_{W^{k,\infty}(\Omega_{m,N})} \leq C_1 \Big(\frac{2\sqrt{d}}{N}\Big)^{1-k} |\tilde{f}|_{W^{1,\infty}(\Omega_{m;n})} \leq C_2 \Big(\frac{1}{N}\Big)^{1-k} \|\tilde{f}\|_{W^{1,\infty}(\Omega_{m;n})}.$$

Now using ϕ_m as defined above, we get

$$\Big\|\phi_m\Big(\tilde{f} - c_{f,m}\Big)\Big\|_{\infty,\Omega_{m,N}} \leq \|\phi_m\|_{\infty,\Omega_{m,N}} \cdot \Big\|\tilde{f} - c_{f,m}\Big\|_{\infty,\Omega_{m,N}} \leq C_2 \frac{1}{N} \|\tilde{f}\|_{W^{1,\infty}(\Omega_{m,N})}. \tag{2.26}$$

Due to the product inequality for weak derivatives (Gühring et al., 2020, Lemma B.6) there is a constant $C' > 0$ such that the supremum norm of the weak derivative is bounded by

$$\begin{aligned}
\Big|\phi_m\Big(\tilde{f} - c_{f,m}\Big)\Big|_{W^{1,\infty}(\Omega_{m,N})} &\leq C' |\phi_m|_{W^{1,\infty}(\Omega_{m,N})} \cdot \Big\|\tilde{f} - c_{f,m}\Big\|_{\infty,\Omega_{m,N}} \\
&\quad + C' \|\phi_m\|_{\infty,\Omega_{m,N}} \cdot |\tilde{f} - c_{f,m}|_{W^{1,\infty}(\Omega_{m,N})} \\
&\leq C' \cdot cN \cdot C_2 \frac{1}{N} \|\tilde{f}\|_{W^{1,\infty}(\Omega_{m,N})} + C' \cdot C_3 \|\tilde{f}\|_{W^{1,\infty}(\Omega_{m,N})} \\
&= C_4 \|\tilde{f}\|_{W^{1,\infty}(\Omega_{m,N})}.
\end{aligned} \tag{2.27}$$

Combining (2.26) and (2.27) we get

$$\Big\|\phi_m\Big(\tilde{f} - c_{f,m}\Big)\Big\|_{W^{1,\infty}(\Omega_{m,N})} \leq C_5 \|\tilde{f}\|_{W^{1,\infty}(\Omega_{m,N})}.$$

Step 3 (Global estimate in $\|\cdot\|_{W^{k,p}}$): As $\sum_{m\in\{0,\dots,N\}^d}\phi_m = 1$, we have that

$$\tilde{f}(x) = \sum_{m\in\{0,\dots,N\}^d} \phi_m(x)\tilde{f}(x), \quad \text{for almost every } x\in(0,1)^d.$$

As $\tilde{f}\big|_{(0,1)^d} = f$ we have for $k\in\{0,1\}$

$$\begin{aligned}\Big\|f - \sum_{m\in\{0,\dots,N\}^d}\phi_m c_{f,m}\Big\|_{W^{k,\infty}((0,1)^d)} &= \Big\|\tilde{f} - \sum_{m\in\{0,\dots,N\}^d}\phi_m c_{f,m}\Big\|_{W^{k,\infty}((0,1)^d)}\\ &= \Big\|\sum_{m\in\{0,\dots,N\}^d}\phi_m\big(\tilde{f}-c_{f,m}\big)\Big\|_{W^{k,\infty}((0,1)^d)}\\ &\le \sup_{\widetilde{m}\in\{0,\dots,N\}^d}\Big\|\sum_{m\in\{0,\dots,N\}^d}\phi_m\big(\tilde{f}-c_{f,m}\big)\Big\|_{W^{k,\infty}(\Omega_{\widetilde{m},N})},\end{aligned} \tag{2.28}$$

where the last step follows from $(0,1)^d\subset\bigcup_{\widetilde{m}\in\{0,\dots,N\}^d}\Omega_{\widetilde{m},N}$. Now we obtain for each $\widetilde{m}\in\{0,\dots,N\}^d$ using (2.25), (2.26) and (2.27)

$$\begin{aligned}\Big\|\sum_{m\in\{0,\dots,N\}^d}\phi_m\big(\tilde{f}-c_{f,m}\big)\Big\|_{W^{k,\infty}(\Omega_{\widetilde{m},N})} &\le \sup_{\substack{m\in\{0,\dots,N\}^d\\ |m-\widetilde{m}|_\infty\le 1}}\big\|\phi_m\big(\tilde{f}-c_{f,m}\big)\big\|_{W^{k,\infty}(\Omega_{\widetilde{m},N})}\\ &\le \sup_{\substack{m\in\{0,\dots,N\}^d\\ |m-\widetilde{m}|_\infty\le 1}}\big\|\phi_m\big(\tilde{f}-c_{f,m}\big)\big\|_{W^{k,\infty}(\Omega_{m,N})}\\ &\le C_6\Big(\frac{1}{N}\Big)^{1-k}\sup_{\substack{m\in\{0,\dots,N\}^d,\\ |m-\widetilde{m}|_\infty\le 1}}\|\tilde{f}\|_{W^{1,\infty}(\Omega_{m,N})}.\end{aligned}$$

Plugging this into (2.28), we obtain for $k\in\{0,1\}$

$$\begin{aligned}\Big\|f - \sum_{m\in\{0,\dots,N\}^d}\phi_m c_{f,m}\Big\|_{W^{k,\infty}((0,1)^d)} &\le C_6\Big(\frac{1}{N}\Big)^{(1-k)}\sup_{\widetilde{m}\in\{0,\dots,N\}^d}\Big(\sup_{\substack{m\in\{0,\dots,N\}^d,\\ |m-\widetilde{m}|_\infty\le 1}}\|\tilde{f}\|_{W^{1,\infty}(\Omega_{m,N})}\Big)\\ &\le C_7\Big(\frac{1}{N}\Big)^{(1-k)}\sup_{\widetilde{m}\in\{0,\dots,N\}^d}\|\tilde{f}\|_{W^{1,\infty}(\Omega_{\widetilde{m},N})}\\ &\le C_8\Big(\frac{1}{N}\Big)^{(1-k)}\|\tilde{f}\|_{W^{1,\infty}(\mathbb{R}^d)}\\ &\le C_9\Big(\frac{1}{N}\Big)^{(1-k)}\|f\|_{W^{1,\infty}((0,1)^d)}.\end{aligned} \tag{2.29}$$

Step 4 (Interpolation): Define the linear operators $T_0\colon W^{1,\infty}((0,1)^d)\to L^\infty((0,1)^d)$, $T_\alpha\colon W^{1,\infty}((0,1)^d)\to\mathcal{H}^\alpha((0,1)^d)$ and $T_1\colon W^{1,\infty}((0,1)^d)\to W^{1,\infty}((0,1)^d)$ via

$$T_k(f) \coloneqq f - \sum_{m\in\{0,\dots,N\}^d}\phi_m c_{f,m}, \quad k\in\{0,\alpha,1\}.$$

Note that the linearity follows from the definition of the constants $c_{f,m}$. Using Lunardi (2018,

Theorem 1.6), for the nontrivial interpolation couple see Lunardi (2018, p. 11 f.), leads to

$$\|T_\alpha\| \leq \|T_0\|^{1-\alpha}\|T_1\|^\alpha.$$

Note that $\|\cdot\|_{\mathcal{H}^\alpha}$ is equivalent to $\|\cdot\|_{W^{s,\infty}(\Omega)}$ in Gühring et al. (2020). Using (2.29) we conclude

$$\Big\|f - \sum_{m\in\{0,\dots,N\}^d} \phi_m c_{f,m}\Big\|_{\mathcal{H}^\alpha} \leq C_{10}\Big(\frac{1}{N}\Big)^{1-\alpha}\|f\|_{W^{1,\infty}}. \qquad \square$$

Now we want to approximate the function $\sum_{m\in\{0,\dots,N\}^d} c_{f,m}\phi_m$ in Hölder norm using a ReLU network.

Lemma 2.30. *For any $\varepsilon \in (0,1/2)$ there is a neural network $f_{\mathsf{NN},\varepsilon}$ with ReLU activation function such that for $(c_{f,m})_m$ from Lemma 2.29, there is a constant $C > 0$ such that*

$$\Big\|\sum_{m\in\{0,\dots,N\}^d} \phi_m c_{f,m} - f_{\mathsf{NN},\varepsilon}\Big\|_{\mathcal{H}^\alpha} \leq C\|f\|_{W^{1,\infty}} N^\alpha \varepsilon,$$

the number of layers of $f_{\mathsf{NN},\varepsilon}$ is at most $\lceil C_1 \log_2(\varepsilon^{-1})\rceil$, the number of nonzero weights is at most $\lceil C_2(N+1)^d \log_2^2(\varepsilon^{-1})\rceil$ and the number of neurons is at most $\lceil C_3(N+1)^d(\log_2^2(\varepsilon^{-1})\vee\log_2(\varepsilon^{-1}))\rceil$, with C_1, C_2 and C_3 from Lemma 2.28.

Proof. From Lemma 2.28 we know that there are neural networks $\psi_{\varepsilon,m}$ with at most $\lceil C_1 \log_2(\varepsilon^{-1})\rceil$ layers, $\lceil C_2(N+1)^d \log_2^2(\varepsilon^{-1})\rceil$ nonzero weights and $\lceil C_3(N+1)^d(\log_2^2(\varepsilon^{-1}) \vee \log_2(\varepsilon^{-1}))\rceil$ neurons that approximate ϕ_m such that for $k \in \{0,1\}$

$$\|\phi_m - \psi_{\varepsilon,m}\|_{W^{k,\infty}} \leq c' N^k \varepsilon.$$

Now we parallelize these networks and multiply with the coefficients $c_{f,m}$ afterwards. Hereby, we construct a network $f_{\mathsf{NN},\varepsilon}$ with $1 + \lceil C_1 \log_2(\varepsilon^{-1})\rceil$ layers, $N^d + \lceil C_2(N+1)^d \log_2^2(\varepsilon^{-1})\rceil$ nonzero weights and $1 + \lceil C_3(N+1)^d(\log_2^2(\varepsilon^{-1}) \vee \log_2(\varepsilon^{-1}))\rceil$ neurons such that

$$f_{\mathsf{NN},\varepsilon} = \sum_{m\in\{0,\dots,N\}^d} c_{f,m}\psi_{\varepsilon,m}. \tag{2.30}$$

For each $m \in \{0, ..., N\}^d$ denote $\Omega_{m,N} = B_{\frac{1}{N},|\cdot|_\infty}(\frac{m}{N})$ as above. For $k \in \{0, 1\}$ we get

$$\begin{aligned}
\Big\|f_{\mathsf{NN},\varepsilon} - \sum_{m\in\{0,\dots,N\}^d} c_{f,m}\phi_m\Big\|_{W^{k,\infty}((0,1)^d)} &= \Big\| \sum_{m\in\{0,\dots,N\}^d} c_{f,m}(\psi_{\varepsilon,m} - \phi_m)\Big\|_{W^{k,\infty}((0,1)^d)} \\
&\le \sup_{\widetilde{m}\in\{0,\dots,N\}^d} \Big\| \sum_{m\in\{0,\dots,N\}^d} c_{f,m}(\psi_{\varepsilon,m} - \phi_m)\Big\|_{W^{k,\infty}(\Omega_{\widetilde{m},N}\cap(0,1)^d)} \\
&\le 3^d \sup_{\widetilde{m}\in\{0,\dots,N\}^d} \sup_{m\in\{0,\dots,N\}^d} \Big\|c_{f,m}(\psi_{\varepsilon,m} - \phi_m)\Big\|_{W^{k,\infty}(\Omega_{\widetilde{m},N}\cap(0,1)^d)} \\
&\le 3^d \sup_{\widetilde{m}\in\{0,\dots,N\}^d} \sup_{m\in\{0,\dots,N\}^d} |c_{f,m}|\Big\|(\psi_{\varepsilon,m} - \phi_m)\Big\|_{W^{k,\infty}(\Omega_{\widetilde{m},N}\cap(0,1)^d)} \\
&\le 3^d \sup_{\widetilde{m}\in\{0,\dots,N\}^d} \sup_{m\in\{0,\dots,N\}^d} \|\tilde{f}\|_{W^{1,\infty}(\Omega_{m,N})}\|\psi_{\varepsilon,m} - \phi_m\|_{W^{k,\infty}(\Omega_{\widetilde{m},N}\cap(0,1)^d)} \\
&\le CN^k\varepsilon\|f\|_{W^{1,\infty}}.
\end{aligned}$$

The second to last inequality follows from the fact that on $\Omega_{\widetilde{m},N}$ is within the support of ϕ_m only for $|m - \widetilde{m}|_\infty \le 1$. The last inequality follows from (2.30) and the continuity of the extension operator, see Stein (1970, Theorem 5). As in Step 4 of Lemma 2.29, we conclude using Lunardi (2018, Theorem 1.6)

$$\Big\|f_{\mathsf{NN},\varepsilon} - \sum_{m\in\{0,\dots,N\}^d} \phi_m c_{f,m}\Big\|_{\mathcal{H}^\alpha} \le CN^\alpha\varepsilon\|f\|_{W^{1,\infty}}. \qquad \square$$

Now we are ready to prove Theorem 2.7.

Proof of Theorem 2.7. Combining Lemma 2.29 and Lemma 2.30 with $\|f\|_{W^{1,\infty}} \le B$ yields for a constant $C > 0$ for any $\tilde{\varepsilon} \in (0, 1/2)$ that

$$\begin{aligned}
\|f - f_{\mathsf{NN},\tilde{\varepsilon}}\|_{\mathcal{H}^\alpha} &\le \Big\|f - \sum_{m\in\{0,\dots,N\}^d} c_{f,m}\phi_m\Big\|_{\mathcal{H}^\alpha} + \Big\| \sum_{m\in\{0,\dots,N\}^d} c_{f,m}\phi_m - f_{\mathsf{NN},\tilde{\varepsilon}}\Big\|_{\mathcal{H}^\alpha} \\
&\le CB\Big(\Big(\frac{1}{N}\Big)^{1-\alpha} + N^\alpha\tilde{\varepsilon}\Big), \qquad (2.31)
\end{aligned}$$

where $\tilde{\varepsilon}$ determines the approximation accuracy in Lemma 2.30. For

$$N := \left\lceil \Big(\frac{\varepsilon}{2CB}\Big)^{-1/(1-\alpha)} \right\rceil,$$

we get for the first term in (2.31)

$$\Big(\frac{1}{N}\Big)^{1-\alpha} \le \frac{\varepsilon}{2CB}.$$

Choosing

$$\tilde{\varepsilon} = \frac{\varepsilon}{2CB}\Big(\Big(\frac{\varepsilon}{2CB}\Big)^{-\frac{1}{1-\alpha}} + 1\Big)^{-\alpha} \qquad (2.32)$$

leads to

$$\|f - f_{\mathsf{NN},\tilde{\varepsilon}}\|_{\mathcal{H}^\alpha} \le \varepsilon.$$

From Lemma 2.28 we know that there is a ReLU network with no more than $1 + \lceil C_1 \log_2(\tilde{\varepsilon}^{-1})\rceil$

layers, $N^d + \lceil C_2(N+1)^d \log_2^2(\tilde{\varepsilon}^{-1}) \rceil$ nonzero weights and $1 + \lceil C_3(N+1)^d (\log_2^2(\tilde{\varepsilon}^{-1}) \vee \log_2(\tilde{\varepsilon}^{-1})) \rceil$ neurons with the required properties. Inserting (2.32) and assuming $CB > \frac{1}{2}$ yields

$$\log_2(\tilde{\varepsilon}^{-1}) \leq \log_2 \Big(\frac{2CB}{\varepsilon} 2^\alpha \Big(\frac{\varepsilon}{2CB} \Big)^{-\frac{\alpha}{1-\alpha}} \Big) \leq C' \log_2(\varepsilon^{-\frac{1}{1-\alpha}}).$$

Thus there are C', C'' and C''' such that the ReLU network has no more than $1 + \lceil C' \log_2(\varepsilon^{-\frac{1}{1-\alpha}}) \rceil$ layers, $\lceil C'' \varepsilon^{-\frac{d}{1-\alpha}} \log_2^2(\varepsilon^{-\frac{1}{1-\alpha}}) \rceil$ nonzero weights and $1 + \lceil C''' \varepsilon^{-\frac{d}{1-\alpha}} (\log_2^2(\varepsilon^{-\frac{1}{1-\alpha}}) \vee \log_2(\varepsilon^{-\frac{1}{1-\alpha}})) \rceil$ neurons. Taking the largest constant yields the first part of the result.
Since $f \in \mathrm{Lip}(K, B) \subseteq \mathcal{H}^\alpha(\Gamma)$ for $\Gamma = \max(K, 2B)$, we conclude

$$\|f_{\mathsf{NN},\tilde{\varepsilon}}\|_{\mathcal{H}^\alpha} \leq \|f\|_{\mathcal{H}^\alpha} + \|f_{\mathsf{NN},\tilde{\varepsilon}} - f\|_{\mathcal{H}^\alpha} \leq \Gamma + \varepsilon. \qquad \square$$

Chapter 3

Generative adversarial networks

The first method we are going to analyze are *Generative Adversarial Networks* (GANs) introduced by Goodfellow et al. (2014). As already indicated in the introduction, GANs have attracted much attention in the 2010s, initially due to impressive results in the creation of photorealistic images. Meanwhile, the areas of application have expanded far beyond this, and GANs serve as a prototypical example of generative models. To ensure easy readability, we shortly recall the definitions from the introduction.

The *Vanilla GAN* as constructed by Goodfellow et al. (2014) relies on the minimax game

$$\inf_{G\in\mathcal{G}} \sup_{D\in\mathcal{D}} \mathbb{E}\big[\log D(X) + \log\big(1 - D(G(Z))\big)\big], \tag{3.1}$$

to learn an unknown distribution $\mathbb{P}^*$ of the random variable X. The generator G chosen from a set $\mathcal{G}$, applied to the latent random variable Z aims to mimic the distribution of X as closely as possible. The discriminator D, chosen from a set $\mathcal{D}$, has to distinguish between real and fake samples.

Generalizations of the underlying Jensen-Shannon distance have led to various extensions of the original GAN, such as f-GANs (Nowozin et al., 2016). More famously, *Wasserstein GANs* (Arjovsky et al., 2017), characterized by

$$\inf_{G\in\mathcal{G}} \sup_{W\in\mathrm{Lip}(1)} \mathbb{E}\big[W(X) - W(G(Z))\big], \tag{3.2}$$

are obtained by replacing the Jensen-Shannon divergence by the Kantorovich dual of the Wasserstein distance. This approach can be generalized using *Integral Probability Metrics* (Müller, 1997).

In contrast to Wasserstein GANs, Vanilla GANs and the Jensen-Shannon divergence have been studied less extensively, and fundamental questions have not been settled. In particular, all statistical results for Vanilla GANs require the same dimension of the latent space and the target space which is in stark contrast to common practice. By (2.19), the Jensen-Shannon divergence between singular measures is maximal. This leads to the algorithmic drawback of Vanilla GANs highlighted by Arjovsky & Bottou (2017) is that an arbitrarily large discriminator class prevents the generator from learning. Thus, using neural networks as a discriminator class must be advantageous compared to the set of all measurable functions. This empirical fact is

supported by the numerical results of Farnia & Tse (2018) who impose a Lipschitz constraint on the discriminator class. In the following, we broaden the theoretical boundaries of Vanilla GANs to cope with the empirical evidence. To this end, we replace the Jensen-Shannon framework with a Wasserstein perspective.

Related work The existence and uniqueness of the optimal generator for Vanilla GANs is shown by Biau et al. (2020) under the condition that the class $\mathcal{G}$ is convex and compact. They also study the asymptotic properties of Vanilla GANs. Puchkin et al. (2024) have shown a non-asymptotic rate of convergence in the Jensen-Shannon divergence for Vanilla GANs with neural networks under the assumption that the density of $\mathbb{P}^*$ exists and that the generator functions are continuously differentiable.
In practice, however, the ReLU is commonly used (Aggarwal, 2018, p. 13). The resulting neural network generates continuous piecewise linear functions. Therefore, the convergence rate of Puchkin et al. (2024) combined with Belomestny et al. (2023) is not applicable to this class of functions.
The statistical analysis of Wasserstein GANs is much better understood. Biau et al. (2021) have studied optimization and asymptotic properties. Liang (2021) has shown error decompositions with respect to the Kullback-Leibler divergence, the Hellinger distance and the Wasserstein distance. The case where the unknown distribution lies on a low-dimensional manifold is considered in Schreuder et al. (2021) as well as Tang & Yang (2023). The latter also derived minimax rates in a more general setting using the Hölder metric. Assuming that the density function of $\mathbb{P}^*$ exists, Liang (2017) has shown a rate of convergence in Wasserstein distance with ReLU activation function and a factor growing exponentially in the depth of the network. Theoretical results including sampling the latent distribution in addition to dimension reduction have been derived by Huang et al. (2022), who have also shown a rate of convergence in a slightly more general Hölder setting using ReLU networks whose Lipschitz constant grows exponentially in the depth. A rate of convergence using the total variation metric and leaky ReLU networks has been shown in Liang (2021).
Convergence rates with respect to the Wasserstein distance have been studied by Chen et al. (2020) and Lee et al. (2025). Up to a logarithmic factor, optimal rates in the Hölder metric were obtained by Stéphanovitch et al. (2024) using smooth networks. In a similar setting, Chakraborty & Bartlett (2025) discussed several methods for dimension reduction. Recently, Suh & Cheng (2024) have reviewed the theoretical advances in Wasserstein GANs.
Ensuring Lipschitz continuity of the discriminator class is the essential property of Wasserstein GANs. Lipschitz-constrained neural networks and their empirical success are subject of ongoing research (Khromov & Singh, 2024). In context of GANs see Than & Vu (2021). Implementations of the Lipschitz constrained discriminator have evolved from weight clipping (Arjovsky et al., 2017) to penalizing the objective function (Gulrajani et al., 2017; Wei et al., 2018; Zhou et al., 2019; Petzka et al., 2018; Miyato et al., 2018; Asokan & Seelamantula, 2023), which heuristically leads to networks with bounded Lipschitz constants. Farnia & Tse (2018) use an objective function that combines Wasserstein and Vanilla GANs.

Own contribution This chapter aims to bridge the gap in theoretical analysis between Vanilla GANs and Wasserstein GANs while addressing the theoretical limitations of the former ones. By imposing a Lipschitz condition on the discriminator class in (3.1), we recover Wasserstein GAN-like behavior. As a main result, we can derive an oracle inequality for the Wasserstein distance between the true data generating distribution and its Vanilla GAN estimate. In particular, this allows us to transfer key features, such as dimension reduction, known from the statistical analysis of Wasserstein GANs. We show that the statistical error of the modified Vanilla GAN depends only on the dimension of the latent space, independent of the potentially much larger dimension of the feature space $\mathcal{X}$. Thus, Vanilla GANs can avoid the curse of dimensionality. Such properties are well known from practice, see for example Figure 1.1, but cannot be verified by the classical Jensen-Shannon analysis. On the other hand the derived rate of convergence for the Vanilla GAN is slower than for Wasserstein GANs which is in line with the empirical advantage of Wasserstein GANs.

Afterwards we consider the most relevant case where the classes $\mathcal{G}$ and $\mathcal{D}$ are parameterized by neural networks. Using our previous results, we derive an oracle inequality that depends on the network approximation errors for the best possible generator and the optimal Lipschitz discriminator. To bound the approximation error of the discriminator, we replace the Lipschitz constraint on the networks with a less restrictive Hölder constraint. This enables the use of the approximation result shown in Theorem 2.7. As a result, we obtain the rate of convergence $n^{-\alpha/2d^*}, \alpha \in (0,1)$, with latent space dimension $d^* \geq 2$ for sufficiently large classes of networks. Additionally, our approximation theorem allows for an explicit bound on the discriminator approximation error for Wasserstein-type GANs, which achieve the rate $n^{-\alpha/d^*}, \alpha \in (0,1)$.

We use a simple illustrative example to assess the practical implications of our theoretical results. This example allows us to quantify the rate depending on the number of observations, the dimension reduction property, and the stabilizing effect of a Lipschitz-constrained discriminator class.

3.1. The Vanilla GAN distance

Let us first fix two notations that are unusual or specific to this chapter. For ease of notation we abbreviate for $x \in (0,\infty)$

$$[x]^{1;1/2} := \max\{x, \sqrt{x}\}. \tag{3.3}$$

When referring to function spaces, we omit the domain Ω in in this chapter if $\Omega = (0,1)^d$.

In this chapter, we assume to observe i.i.d. samples $X_1, ..., X_n \sim \mathbb{P}^*$ with values in $\mathcal{X} := (0,1)^d$. On another space $\mathcal{Z} := (0,1)^{d^*}$, called the *latent* space, we choose a latent distribution $\mathbb{U}$. Unless otherwise specified, $X \sim \mathbb{P}^*$ and $Z \sim \mathbb{U}$. We further assume that $\mathbb{P}^*$ and $\mathbb{U}$ have finite first moments. Throughout, the generator class $\mathcal{G}$ is a nonempty set of measurable functions from $\mathcal{Z}$ to $\mathcal{X}$.

Typically the discriminator class consists of functions mapping to $\mathbb{R}$ concatenated to a sigmoid function that maps into $(0,1)$ to account for the classification task. This is especially the case for standard classification networks. The most common sigmoid function used for this purpose

is the logistic function $x \mapsto (1+e^{-x})^{-1}$, which we fix throughout. Together with a shift by $\log 4$, we can rewrite the Vanilla GAN optimization problem (3.1) as

$$\inf_{G \in \mathcal{G}} \mathsf{V}_{\mathcal{W}}(\mathbb{P}^*, \mathbb{P}^{G(Z)}) \tag{3.4}$$

in terms of the *Vanilla GAN distance* between probability measures $\mathbb{P}$ and $\mathbb{Q}$ on $\mathcal{X}$

$$\mathsf{V}_{\mathcal{W}}(\mathbb{P}, \mathbb{Q}) \coloneqq \sup_{W \in \mathcal{W}} \mathbb{E}_{\substack{X \sim \mathbb{P} \\ Y \sim \mathbb{Q}}} \Big[-\log\Big(\frac{1+e^{-W(X)}}{2}\Big) - \log\Big(\frac{1+e^{W(Y)}}{2}\Big)\Big], \tag{3.5}$$

where $\mathcal{W}$ is a set of measurable functions $W \colon \mathcal{X} \to \mathbb{R}$. As long as $0 \in \mathcal{W}$, we have that $\mathsf{V}_{\mathcal{W}} \geq 0$. To choose the generator $\hat{G}_n$ as the empirical risk minimizer, the unknown distribution $\mathbb{P}^*$ in (3.4) must be replaced by the empirical distribution $\mathbb{P}_n$ based on the observations $X_1, ..., X_n$. In practice, the expectation with respect to $Z \sim \mathbb{U}$ is replaced by an empirical mean too, which we omit for the sake of simplicity. Along Huang et al. (2022), the next and all subsequent results easily extend to the corresponding setting.

The following error bound in terms of the Vanilla GAN distance provides an error decomposition for the empirical risk minimizer of the Vanilla GAN. To this end, define $\mathrm{Lip}(1) \circ \mathcal{W}$ as the set of all concatenations $f \circ g$, where $f \in \mathrm{Lip}(1)$ and $g \in \mathcal{W}$.

Lemma 3.1. *Assume that $\mathcal{G}$ is chosen such that a minimum exists. Let $\mathcal{W}$ be symmetric, that is, $W \in \mathcal{W}$ implies $-W \in \mathcal{W}$. For*

$$\hat{G}_n \in \operatorname*{arg\,min}_{G \in \mathcal{G}} \mathsf{V}_{\mathcal{W}}(\mathbb{P}_n, \mathbb{P}^{G(Z)}) \tag{3.6}$$

we have that

$$\mathsf{V}_{\mathcal{W}}(\mathbb{P}^*, \mathbb{P}^{\hat{G}_n(Z)}) \leq \min_{G \in \mathcal{G}} \mathsf{V}_{\mathcal{W}}(\mathbb{P}^*, \mathbb{P}^{G(Z)}) + 2 \sup_{W \in \mathrm{Lip}(1) \circ \mathcal{W}} \frac{1}{n} \sum_{i=1}^{n} \big(W(X_i) - \mathbb{E}[W(X)]\big). \tag{3.7}$$

The first term in (3.7) is the error due to the approximation capabilities of the class $\mathcal{G}$. The second term refers to the stochastic error due to the amount of training data. As $\mathcal{W}$ is symmetric, the stochastic error is non-negative. Both error terms depend on the discriminator class $\mathcal{W}$. Large discriminator classes lead to finer discrimination between different probability distributions and thus to a larger approximation error term. Similarly, the stochastic error term will increase with the size of $\mathcal{W}$. The cost of small classes $\mathcal{W}$ is a less informative loss function on the left side of (3.7).

If $\mathcal{W}$ is the set of all measurable functions, the analysis by Goodfellow et al. (2014, Theorem 1) shows that the Vanilla GAN distance is equivalent to the Jensen-Shannon distance. Arjovsky & Bottou (2017) have elaborated on the theoretical and practical disadvantages of this case. As already emphasized, the Jensen-Shannon divergence is not compatible with high-dimensional settings because it cannot distinguish between different singular measures. Therefore, we need a weaker distance and thus restrict $\mathcal{W}$.

The key insight of Wasserstein GANs (3.2) is that this particular drawback of the Jensen-Shannon

distance can be solved by the Wasserstein distance, as introduced in Section 2.2.1. In this chapter, we need a slightly adapted version of the Wasserstein distance. Instead of taking the supremum over all Lipschitz 1 functions in (2.15), we only consider Lipschitz 1 functions W such that $W(0) = 0$. Since this adaptation only leads to the addition of a constant, which cancels in the objective function, we conclude that

$$\mathsf{W}_1(\mathbb{P}, \mathbb{Q}) = \sup_{\substack{W \in \mathrm{Lip}(1) \\ W(0)=0}} \mathbb{E}_{\substack{X \sim \mathbb{P} \\ Y \sim \mathbb{Q}}}[W(X) - W(Y)]. \tag{3.8}$$

Bounds for weaker metrics, such as the Kolmogorov or Levy metric, can be easily derived from the bounds in the Wasserstein metric under weak conditions, see e.g. Gibbs & Su (2002).

Therefore, we choose $\mathcal{W} = \mathrm{Lip}(L)$ for some $L \geq 1$ in Lemma 3.1. In this case the following result shows that the existence of an empirical risk minimizer is guaranteed as soon as $\mathcal{G}$ is compact.

Lemma 3.2. *Assume $\mathcal{G}$ is compact with respect to the supremum norm. The map $T\colon \mathcal{G} \to \mathbb{R}_{\geq 0}$, $T(G) \coloneqq V_{\mathrm{Lip}(L)}(\mathbb{P}_n, \mathbb{P}^{G(z)})$ is continuous and $\arg\min_{\mathcal{G}} V_{\mathrm{Lip}(L)}(\mathbb{P}^*, \mathbb{P}^{G(z)})$ is nonempty.*

Hence, we throughout assume the following:

Assumption 3.3. *$\mathcal{G}$ is compact with respect to the supremum norm.*

In the context of neural networks the compactness assumption is satisfied for all practically relevant implementations. Furthermore, it should be noted that the aforementioned assumption is only required for the use of the minimizing argument.

3.2. Relation between Vanilla GAN and Wasserstein distance

Our subsequent analysis builds on the following equivalence result between the Vanilla GAN distance and the Wasserstein distance with an additional L_2-penalty term on the discriminator.

Theorem 3.4. *For $L > 2$ and $B > 0$ we have for probability measures $\mathbb{P}$ and $\mathbb{Q}$ on $\mathcal{X}$*

$$\begin{aligned}
&\sup_{\substack{W \in \mathrm{Lip}(1, B') \\ W(\cdot) > -\log(2-2/L)}} \Big\{ \mathbb{E}_{\substack{X \sim \mathbb{P} \\ Y \sim \mathbb{Q}}}[W(X) - W(Y)] - \frac{L(L-1)}{2} \mathbb{E}_{X \sim \mathbb{Q}}[W(X)^2] \Big\} \\
&\qquad \leq \mathsf{V}_{\mathrm{Lip}(L,B)}(\mathbb{P}, \mathbb{Q}) \\
&\qquad \leq \sup_{\substack{W \in \mathrm{Lip}(L, B) \\ W(\cdot) > -\log(2)}} \Big\{ \mathbb{E}_{\substack{X \sim \mathbb{P} \\ Y \sim \mathbb{Q}}}[W(X) - W(Y)] - \frac{e^B}{(2e^B - 1)^2} \mathbb{E}_{X \sim \mathbb{Q}}[W(X)^2] \Big\},
\end{aligned}$$

where $B' = \log((1 + e^B)/2)$.

Note that, using the function $g\colon (-\infty, \log(2 - 2/L)) \to \mathbb{R},\ g(x) = -\log(2e^{-x} - 1)$, we obtain lower and upper bounds with a penalty term depending on $\mathbb{E}[W(Y)^2]$ instead of $\mathbb{E}[W(X)^2]$. Theorem 3.4 reveals that the Vanilla GAN distance is indeed compatible with the Wasserstein distance and will allow us to prove rates of convergence of the Vanilla GAN with respect to the

Wasserstein distance. In doing so, we need to investigate the consequences of the penalty term. An upper bound without the penalty term and independent of B can be shown as in the proof of Theorem 3.5. For the lower bound, a similiar improvement cannot be expected in general as indicated in Example 3.6. However, the restriction to $\mathrm{Lip}(1, B')$ has far less severe consequences than the corresponding restriction in the upper bound.
We can deduce from Theorem 3.4 that the Vanilla GAN distance is bounded from above and below by the Wasserstein distance or the squared Wasserstein distance, respectively. In the following, we equip $\mathcal{X}$ with the p-norm $|\cdot|_p$ for $1 \leq p \leq \infty$.

Theorem 3.5. *Let $L > 2$ and $B \in [1, \infty]$. For probability measures $\mathbb{P}$ and $\mathbb{Q}$ on $\mathcal{X}$ we have*

$$\min\left(c_1 \mathsf{W}_1(\mathbb{P}, \mathbb{Q}), c_2 \mathsf{W}_1(\mathbb{P}, \mathbb{Q})^2\right) \leq \mathsf{V}_{\mathrm{Lip}(L,B)}(\mathbb{P}, \mathbb{Q}) \leq L \mathsf{W}_1(\mathbb{P}, \mathbb{Q}),$$

where $c_1 = \frac{1}{2}\frac{\log(2-2/L)}{d^{1/p}}$ and $c_2 = \frac{1}{2d^{2/p}}\frac{1}{L(L-1)}$, setting $1/p = 0$ if $p = \infty$.

The assumption $L > 2$ is not very restrictive. In practically relevant cases, such as neural network discriminators, the Lipschitz constant is typically quite large. A higher Lipschitz constraint on the discriminator will subsequently result in a less stringent constraint on the neural network. However, an arbitrarily large Lipschitz constant is also undesirable, as the upper bound grows linearly in L.
More importantly, we observe a gap between $\mathsf{W}_1(\mathbb{P}, \mathbb{Q})^2$ in the lower bound and $\mathsf{W}_1(\mathbb{P}, \mathbb{Q})$ in upper bound when $\mathsf{W}_1(\mathbb{P}, \mathbb{Q}) < 1$ which is a consequence of the penalty term in Theorem 3.4. The following example indicates that this loss is unavoidable, by restricting the discriminator class to a subset of $\mathrm{Lip}(L)$.

Example 3.6. *For $\varepsilon, \gamma > 0, \gamma + \varepsilon < 1$ let $\mathbb{P} = \frac{1}{2}(\delta_\gamma + \delta_{\gamma+\varepsilon})$ and $\mathbb{Q} = \frac{1}{2}(\delta_0 + \delta_\varepsilon)$. The Wasserstein distance is then given by*

$$\mathsf{W}_1(\mathbb{P}, \mathbb{Q}) = \gamma.$$

We consider the Vanilla GAN distance using L-Lipschitz affine linear functions as discriminator, $\mathsf{V}_{a\cdot+b}(\mathbb{P}, \mathbb{Q})$, with $a, b \in \mathbb{R}$ and $|a| \leq L$. Note that the class of affine linear functions can be represented by one layer ReLU neural networks. The optimal b can be calculated explicitly, the optimal a can be determined numerically. Using the optimal slope a and b we obtain for $\gamma < \varepsilon$, $\varepsilon = \frac{1}{4}$ and $a > 16$

$$\frac{\mathsf{W}_1(\mathbb{P}, \mathbb{Q})^2}{2} \leq \mathsf{V}_{a\cdot+b}(\mathbb{P}, \mathbb{Q}) \leq a \cdot \mathsf{W}_1(\mathbb{P}, \mathbb{Q})^2.$$

If $\gamma \geq \varepsilon$, then the optimal a is $a = L$ and

$$\log(2) \cdot \mathsf{W}_1(\mathbb{P}, \mathbb{Q}) \leq \mathsf{V}_{a\cdot+b}(\mathbb{P}, \mathbb{Q}) \leq a \cdot \mathsf{W}_1(\mathbb{P}, \mathbb{Q}).$$

Wasserstein GANs, where the generator is chosen as the empirical risk minimizer of the Wasserstein distance (3.8), achieve optimal convergence rates up to logarithmic factors with respect to the Wasserstein distance as proved by Stéphanovitch et al. (2024). In view of Theorem 3.5 we cannot hope that Vanilla GANs achieve the same rate even if we use a Lipschitz discriminator class. This is in line with the better performance of Wasserstein GANs in practice.

However, Theorem 3.5 allows us to study the behavior of Vanilla GANs in settings where the dimension of the latent space is smaller than the dimension of the sample space, a setting that is excluded in all previous works on convergence rates for Vanilla GANs.

3.3. Oracle inequalities for Vanilla GANs

Our aim is to bound the Wasserstein distance between the unknown distribution $\mathbb{P}^*$ and the generated distribution $\mathbb{P}^{\hat{G}_n(Z)}$ using the empirical risk minimizer $\hat{G}_n$ of the Vanilla GAN. The following oracle inequality shows that imposing a Lipschitz constraint on the discriminator class does circumvent the theoretical limitations of Vanilla GANs which is caused by the Jensen-Shannon distance. Recall the notation introduced in (3.3).

Theorem 3.7. *Let $L > 2$ and $B \in [1, \infty]$. For the empirical risk minimizer $\hat{G}_n$ from (3.6) with $\mathcal{W} = \mathrm{Lip}(L, B)$ we have*

$$\mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{\hat{G}_n(Z)}) \leq c\big[\inf_{G \in \mathcal{G}} \mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{G(Z)})\big]^{1;1/2} + (1+c)[\mathsf{W}_1(\mathbb{P}_n, \mathbb{P}^*)]^{1;1/2}, \tag{3.9}$$

for some constant $c > 0$ depending on d, p and L.

Note that the discriminator class $\mathrm{Lip}(L, B)$ admits no finite-dimensional parameterization and is therefore not feasible in practice. We will return to this issue in Section 3.4. The terms in (3.9) can be interpreted analogously to the interpretation of the bound in Lemma 3.1, but here we have an oracle inequality with respect to the Wasserstein distance. The first term is the approximation error. It is large when $\mathcal{G}$ is not flexible enough to provide a good approximation of $\mathbb{P}^*$ by $\mathbb{P}^{G(Z)}$ for some $G \in \mathcal{G}$. The second term refers to the stochastic error. With a growing number of observations the empirical measure $\mathbb{P}_n$ converges to $\mathbb{P}^*$ in Wasserstein distance, see Dudley (1969), and thus the stochastic error converges to zero.
Within the framework of the conceptual proof in Section 2.6, the evaluation distance is the Wasserstein-1 distance, and the optimization criterion c is the Vanilla GAN distance. In Theorem 3.7, we chose the empirical measure as the reference measure, which results in $\varepsilon_c^n = 0$. Theorem 3.5 revealed that the relation between the evaluation distance and the optimization criterion is not linear, resulting in the mixed dependency.
Together with the bounds on $\mathsf{W}_1(\mathbb{P}_n, \mathbb{P}^*)$ by Schreuder (2020) we conclude the following:

Corollary 3.8. *Let $L > 2, B \in [1, \infty]$. The empirical risk minimizer $\hat{G}_n$ from (3.6) with $\mathcal{W} = \mathrm{Lip}(L, B)$ satisfies for some constant $c > 0$ depending on d, p and L that*

$$\begin{aligned}\mathbb{E}[\mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{\hat{G}_n(Z)})] \leq \inf_{G^*\colon \mathcal{Z} \to \mathcal{X}} \Big\{ c[\mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{G^*(Z)})]^{1;1/2} + c[\inf_{G \in \mathcal{G}} \|G - G^*\|_\infty]^{1;1/2} \Big\} \\ + c \begin{cases} n^{-1/2d}, & d > 2, \\ n^{-1/4}(\log n)^{1/2}, & d = 2, \\ n^{-1/4}, & d = 1, \end{cases}\end{aligned}$$

where the infimum is taken over all Borel measurable functions $G^*\colon \mathcal{Z} \to \mathcal{X}$.

If there is some G^* such that $\mathbb{P}^* = \mathbb{P}^{G^*(Z)}$, which is commonly assumed in the GAN literature, see e.g. Stéphanovitch et al. (2024), the first term vanishes and the approximation error is bounded by $\inf_{G\in\mathcal{G}}[\|G - G^*\|_\infty^{1;1/2}]$. In the bound of the stochastic error we observe the curse of dimensionality: For large dimensions d the rate of convergence $n^{-1/2d}$ deteriorates.
To allow for a dimension reduction setting, we adopt the miss-specified setting from Vardanyan et al. (2024, p. 5). In this scenario we can conclude statistical guarantees for Vanilla GANs that are comparable to the results obtained for Wasserstein GANs by Schreuder et al. (2021, Theorem 2). In view of Theorem 3.4 we expect a slower rate of convergence compared to Wasserstein GANs.

Theorem 3.9. *Let $L > 2, B \in [1, \infty]$ and $M > 0$. The empirical risk minimizer $\hat{G}_n$ from (3.6) with $\mathcal{W} = \mathrm{Lip}(L, B)$ satisfies*

$$\mathbb{E}[\mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{\hat{G}_n(Z)})] \leq \inf_{G^*\in\mathrm{Lip}(M,\mathcal{Z})} \left\{ c[\mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{G^*(Z)})]^{1;1/2} + c[\inf_{G\in\mathcal{G}} \|G - G^*\|_\infty]^{1;1/2} \right\}$$
$$+ c \begin{cases} n^{-1/2d^*}, & d^* > 2, \\ n^{-1/4}(\log n)^{1/2}, & d^* = 2, \\ n^{-1/4}, & d^* = 1, \end{cases}$$

for some constant c depending d^*, d, p, L and M.

The Wasserstein distance $\mathsf{W}_1(\mathbb{P}^{G^*(Z)}, \mathbb{P}^*)$ now includes an error due to the dimension reduction while the stochastic error is determined by the potentially much smaller dimension $d^* < d$ of the latent space. Compared to Corollary 3.8, the only price for this improvement is the additional Lipschitz restriction on G^*. We observe a trade-off in the choice of d^*, since large latent dimensions reduce the approximation error for $\mathbb{P}^*$, but increase the stochastic error term. Additionally, there is a trade-off in M. A larger constant M results in a smaller value of $W_1(\mathbb{P}^*, \mathbb{P}^{G^*(Z)})$, but increases the constant c. If the unknown distribution $\mathbb{P}^*$ is supported on a lower-dimensional subspace and there exists a $G^* \in \mathrm{Lip}(M, \mathcal{Z})$ such that $\mathbb{P}^{G^*(Z)} = \mathbb{P}^*$, then the rate of convergence is solely determined by the dimension d^* of $\mathcal{Z}$. This is true for the smallest possible d^* for which a perfect G^* exists, as well as any $d^{**} > d^*$.
In many applications, the smallest possible d^* is unknown. Theorem 3.9 covers both over- and underestimation of the true dimension of the lower-dimensional subspace. If the choice of d^* is too small, then $\mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{G^*(Z)})$ might not converge to 0, but the stochastic error still converges with the smaller rate d^*. If d^* is selected to be larger than the dimension of the lower-dimensional subspace, then there could be a $G^* \in \mathrm{Lip}(M, \mathcal{Z})$ such that $\mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{G^*(Z)}) = 0$, but the stochastic rates converges only with rate d^*. In the special case that a function $G^* \in \mathrm{Lip}(M, \mathcal{Z})$ exists such that $\mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{G^*(Z)}) = 0$ the rate $n^{-1/2d^*}$ is slower than the rate n^{-1/d^*} obtained for the Wasserstein GAN by Schreuder et al. (2021). This is in line with Theorem 3.5 and Example 3.6. However, Theorem 3.9 reveals why Vanilla GANs do perform well in high dimensions in the setting of an unknown distribution on a lower-dimensional manifold. This phenomenon could not be explained in previous work on Vanilla GANs. Puchkin et al. (2024) and Biau et al. (2020) both obtain rates in the Jensen-Shannon distance.

3.4. Rates of convergence for Vanilla GANs in Wasserstein distance

In practice, both $\mathcal{G}$ and $\mathcal{D}$ are sets of neural networks. Our conditions on the generator class $\mathcal{G}$ are compactness and good approximation properties of some G^* which is chosen such that $\mathbb{P}^{G^*(Z)}$ mimics $\mathbb{P}^*$. Since neural networks have a finite number of weights, and the absolute value of those weights is typically bounded, the compactness assumption is usually satisfied and neural networks enjoy excellent approximation properties, c.f. DeVore et al. (2021).
The situation is more challenging for the discriminator class. So far, $\mathcal{D}$ was chosen as the set of Lipschitz functions concatenated to the logistic function. The Lipschitz property is crucial for proof of Theorem 3.7 and thus for all subsequent results.
Controlling the Lipschitz constant while preserving the approximation properties is an area of ongoing research and is far from trivial. Without further restrictions on the class of feedforward networks, the Lipschitz constant would be a term that depends exponentially on the size of the network, see Liang (2017, Theorem 3.2). Bounding the Lipschitz constant of a neural network is a problem that arises naturally in the implementation of Wasserstein GANs. Arjovsky et al. (2017) use weight clipping to ensure Lipschitz continuity. Later, other approaches such as gradient penalization (Gulrajani et al., 2017; Wei et al., 2018; Zhou et al., 2019), Lipschitz penalization (Petzka et al., 2018), or spectral penalization (Miyato et al., 2018) were introduced and have achieved improved performance in practice.
To extend the theory from the previous section to neural network discriminator classes, we first generalize Theorem 3.7 from $\mathcal{W} = \mathrm{Lip}(L, B)$ to subsets $\mathcal{W} \subseteq \mathrm{Lip}(L, B)$. As a result there is an additional approximation error term that accounts for the smaller discriminator class.

Theorem 3.10. *Let $L > 2, B \in [1, \infty]$. The empirical risk minimizer $\hat{G}_n$ from (3.6) with $\mathcal{W} \subseteq \mathrm{Lip}(L, B)$ satisfies*

$$\mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{\hat{G}_n(Z)}) \leq c\big[\inf_{G\in\mathcal{G}} \mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{G(Z)})\big]^{1;1/2} + c\big[\inf_{W'\in\mathcal{W}} \sup_{W\in\mathrm{Lip}(L,B)} \|W - W'\|_\infty\big]^{1;1/2} + c\big[\mathsf{W}_1(\mathbb{P}_n, \mathbb{P}^*)\big]^{1;1/2},$$

for some constant $c > 0$ depending on d, p and L.

The approximation error of the discriminator depends on the supremum norm bound B of the functions in $\mathcal{W}$. While the statement remains true for $B = \infty$, when approximating the set $\mathrm{Lip}(L, B)$, this bound will be essential. To apply this result, we must ensure that the Lipschitz constant of a set of neural networks $\mathcal{W}$ is uniformly bounded by some constant L. Adding penalties to the objective function of the optimization problem does not guarantee a fixed bound on the Lipschitz constant. Approaches such as bounds on the spectral or row-sum norm of matrices in feedforward neural networks ensure a bound on the Lipschitz constant, but lead to a loss of expressiveness when considering ReLU networks, even in very simple cases such as the absolute value, see Huster et al. (2019) and Anil et al. (2019). On the other hand, Eckstein (2020) has shown that one-layer L Lipschitz networks are dense (with respect to the uniform norm) in

the set of all L Lipschitz functions on bounded domains. While this implies that the discriminant approximation error converges to zero for growing network architectures, the density statement does not lead to a rate of convergence that depends on the size of the network.

Anil et al. (2019), motivated by Chernodub & Nowicki (2016), have introduced an adapted activation function, Group Sort, which leads to significantly improved approximation properties of the resulting networks. They show that networks using the Group Sort activation function are dense in the set of Lipschitz functions, but there is no quantitative approximation result. A discussion of the use of Group Sort in the context of Wasserstein GANs can be found in Biau et al. (2021).

To overcome this problem, we would like to approximate not only the optimal discriminating Lipschitz function from the Wasserstein optimization problem in the uniform norm, but also its (weak) derivative. This would allow us to keep the Lipschitz norm of the approximating neural network bounded. For networks with regular activation functions Belomestny et al. (2023) have studied the simultaneous approximation of smooth functions and their derivatives. Gühring et al. (2020) have focused on ReLU networks and have derived quantitative approximation bounds in higher order Hölder and Sobolev spaces. As an intrinsic insight from approximation theory, the regularity of the function being approximated must exceed the regularity order of the norm used to derive approximation bounds. Therefore, we cannot expect to obtain quantitative approximation results for ReLU networks in Lipschitz norm without assuming the continuous differentiability of the approximated function.

Unfortunately, the maximizing function of the Wasserstein optimization problem is in general just Lipschitz continuous. Since we cannot increase the regularity of the target function, we instead relax the Lipschitz assumption of the discriminator in Theorem 3.11 to α-Hölder continuity for $\alpha \in (0,1)$. This generalization in the context of Wasserstein GANs has recently been discussed by Stéphanovitch et al. (2024). Recall the definition of the Hölder ball from (2.4).

Theorem 3.11. *Let $L > 2, B \in [1,\infty)$, $\Gamma > \max(L, 2B)$ and $M > 0$. The empirical risk minimizer $\hat{G}_n$ from (3.6) with $\mathcal{W} \subseteq \mathcal{H}^\alpha(\Gamma)$ satisfies*

$$\begin{aligned}\mathbb{E}[\mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{\hat{G}_n(Z)})] \leq \inf_{G^* \in \mathrm{Lip}(M,\mathcal{Z})} \Big\{ c\big[\inf_{G\in\mathcal{G}} \|G^* - G\|_\infty^\alpha\big]^{1;1/2} + c\big[\mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{G^*(Z)})^\alpha\big]^{1;1/2} \Big\} \\ + c\big[\inf_{W\in\mathcal{W}} \sup_{W^* \in \mathrm{Lip}(L,B)} \|W - W^*\|_\infty\big]^{1;1/2} \\ + c \begin{cases} n^{-\alpha/2d^*}, & 2\alpha < d^*, \\ n^{-1/4}(\log n)^{1/2}, & 2\alpha = d^*, \\ n^{-1/4}, & 2\alpha > d^*, \end{cases}\end{aligned}$$

for some constant c depending on d^*, d, p, L, M and Γ.

The lower bound on the Hölder constant of the discriminator class $\mathcal{W}$ is not overly restrictive when employing neural networks for this function class. Since c is increasing in Γ, it is advantageous to control the value of Γ.

It remains to show that there are ReLU networks that satisfy the assumptions of Theorem 3.11.

To this end, we use the approximation result obtained in Theorem 2.7.
Combining Theorem 3.11 and Theorem 2.7 with a standard approximation result for the generator approximation error, such as Yarotsky (2017, Theorem 1), leads to a rate of convergence. The networks in $\mathcal{G}$ approximating the function $G^* \in \mathrm{Lip}(M, \mathcal{Z})$ are only required to be measurable without any additional smoothness assumption.

Corollary 3.12. *For $0 < \alpha < 1$, $\Gamma > 5$, $M > 0$, $d^* > 2\alpha$ and $n > 2^{\frac{2d^*}{\alpha}}$ choose $\mathcal{G}$ as the set of ReLU networks with at most $\lceil c \cdot \log(n) \rceil$ layers, $\lceil c \cdot n \log(n) \rceil$ nonzero weights and $\lceil c \cdot n \log(n) \rceil$ neurons and $\mathcal{W}'$ as the set of ReLU networks with at most $\lceil c \cdot \log(n) \rceil$ layers, $\lceil c \cdot n^{\frac{\alpha}{2(1-\alpha)}} \log^2(n) \rceil$ nonzero weights and $\lceil c \cdot n^{\frac{\alpha}{2(1-\alpha)}} \log^2(n) \rceil$ neurons, where c is a constant depending on d, d^*, Γ, M and α. Then the empirical risk minimizer $\hat{G}_n$ from (3.6) with $\mathcal{W} = \mathcal{W}' \cap \mathcal{H}^\alpha(\Gamma)$ satisfies*

$$\mathbb{E}\big[\mathrm{W}_1\big(\mathbb{P}^*, \mathbb{P}^{\hat{G}_n(Z)}\big)\big] \leq c \cdot n^{-\alpha/2d^*} + c\big[\inf_{G^* \in \mathrm{Lip}(M,\mathcal{Z})} \mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{G^*(Z)})^\alpha\big]^{1;1/2}.$$

From Theorem 2.7 we know that the set $\mathcal{W}' \cap \mathcal{H}^\alpha(\Gamma)$ of ReLU networks of finite width and depth is nonempty. In practice, this corresponds to a discriminator network with a controlled Hölder constant. On a bounded domain, any Lipschitz function is a Hölder function. Corollary 3.12 shows that Vanilla GANs with a Hölder regular discriminator class are theoretically advantageous. The Hölder parameter α can be chosen arbitrarily close to one. On the one hand this reveals why a Lipschitz regularization as implemented for Wasserstein GANs also improves the Vanilla GAN. An empirical confirmation can be found in Zhou et al. (2019) and Section 3.6. On the other hand the corollary then requires more neurons in the discriminator than in the generator class which coincides with common practice.
The width of the generator networks in Corollary 3.12 can be improved by replacing $G^* \in \mathrm{Lip}(M, \mathcal{Z})$ with $G^* \in \mathrm{C}^{n-1}(\mathcal{Z}), n \in \mathbb{N}$, whose $(n-1)$-th derivative is Lipschitz continuous with Lipschitz constant M. Once more, this results in a trade-off, as $\big[\mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{G^*(Z)})^\alpha\big]^{1;1/2}$ increases when G^* is selected from a smaller set of functions.

3.5. Wasserstein-type GAN with ReLU network discriminator

The same analysis can be applied to Wasserstein-type GANs. The constraint on the Hölder constant can be weakened, as we do not need Theorem 3.5. Note that this does not impact the rate, but the constant. Define the Wasserstein-type distance with discriminator class $\mathcal{W}$ as

$$\mathsf{W}_{\mathcal{W}}(\mathbb{P}, \mathbb{Q}) = \sup_{W \in \mathcal{W}} \mathbb{E}_{\substack{X \sim \mathbb{P} \\ Y \sim \mathbb{Q}}}[W(X) - W(Y)].$$

The following theorem shows that by using Hölder continuous ReLU networks as the discriminator class, Wasserstein-type GANs can avoid the curse of dimensionality. Furthermore, this avoids the difficulties arising from the Lipschitz assumption of the neural network, as pointed out by Huang et al. (2022).

Theorem 3.13. *For $0 < \alpha < 1$, $\Gamma > 1, M > 0$ and $d > 2\alpha$ and $n > 2^{\frac{d}{\alpha}}$ choose $\mathcal{G}$ as the set of ReLU networks with at most $\lceil c \cdot \log(n) \rceil$ layers, $\lceil c \cdot n \log(n) \rceil$ nonzero weights and $\lceil c \cdot n \log(n) \rceil$*

neurons and $\mathcal{W}'$ *as the set of ReLU networks with at most* $\lceil c \cdot \log(n) \rceil$ *layers,* $\lceil c \cdot n^{\frac{\alpha}{(1-\alpha)}} \log^2(n) \rceil$ *nonzero weights and* $\lceil c \cdot n^{\frac{\alpha}{(1-\alpha)}} \log^2(n) \rceil$ *neurons, where* c *is a constant depending on* d, d^*, Γ, M *and* α*. The empirical risk minimizer with* $\mathcal{W} = \mathcal{W}' \cap \mathcal{H}^\alpha(\Gamma)$*,*

$$\hat{G}_n \in \operatorname*{argmin}_{G \in \mathcal{G}} \mathsf{W}_{\mathcal{W}}(\mathbb{P}_n, \mathbb{P}^{G(Z)}),$$

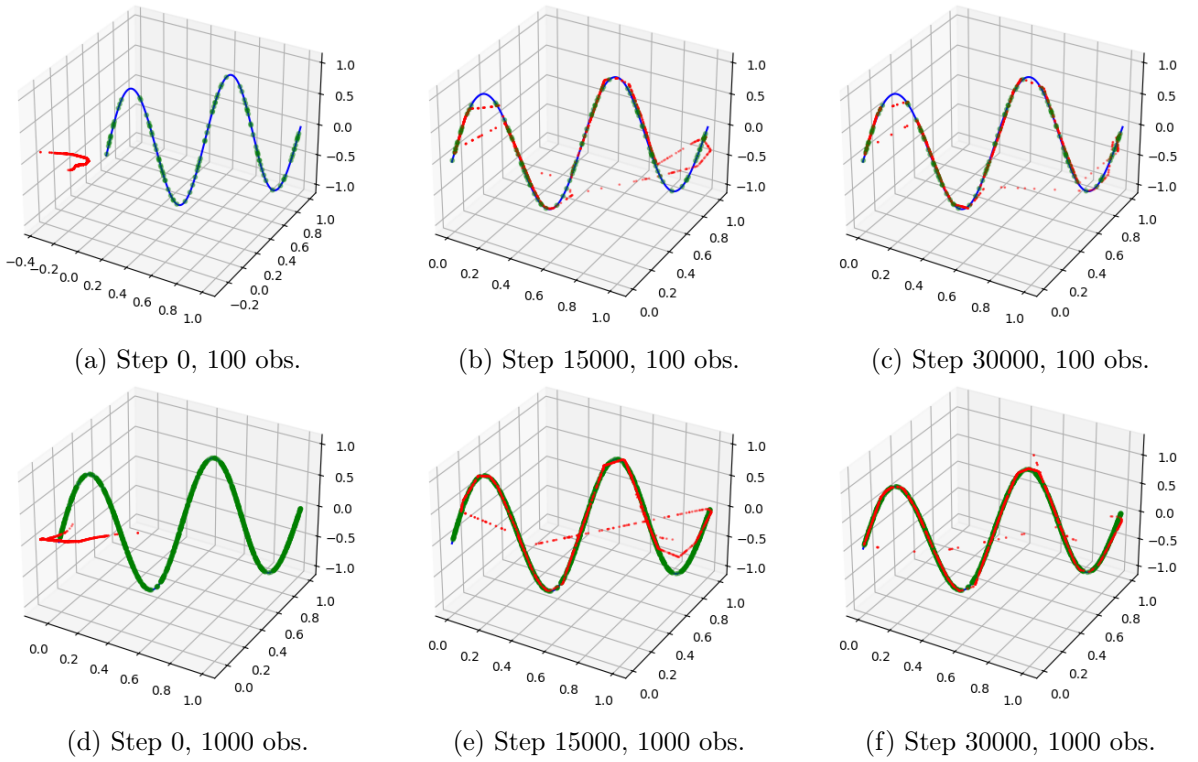
satisfies

$$\mathbb{E}[\mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{\hat{G}_n(Z)})] \leq c \cdot n^{-\frac{\alpha}{d^*}} + \inf_{G^* \in \mathrm{Lip}(M, \mathcal{Z})} \mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{G^*(Z)}).$$

Compared to Corollary 3.12 the rate improves to $n^{-\alpha/d^*}$ for any $\alpha < 1$. The number of observations necessary for the theorem to hold, the size of the discriminator network and the lower bound for Γ decrease. Note that Γ does not effect the rate, but the constants. In case there exists a G^* such that $\mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{G^*(Z)}) = 0$, this upper bound coincides with the lower bound in Tang & Yang (2023, Theorem 1) up to an arbitrary small polynomial factor.
Our rate does not depend exponentially on the number of layers like the results of Liang (2017), Huang et al. (2022) and we use non-smooth simple ReLU networks compared to smooth ReQU (using $\max(0, x)^2$ as activation function in (2.22)) networks in Stéphanovitch et al. (2024) or group sort networks in Biau et al. (2021).

3.6. Simulation

The results in Section 3.4 and Section 3.5 were obtained under the assumption that the discriminator class consists of Lipschitz networks. In the context of image generation, these findings align with the results of Zhou et al. (2019), Miyato et al. (2018), Kodali et al. (2017), and Fedus et al. (2017). Furthermore, Fedus et al. (2017) demonstrated in a two-dimensional experiment that a Vanilla GAN with a gradient penalty (and, consequently, a lower Lipschitz constant) can be effective in scenarios where the measures $\mathbb{P}^*$ and $\mathbb{P}^{\hat{G}(Z)}$ are singular.
This section presents a transparent and accessible example that confirms our theoretical findings and especially demonstrates how imposing a Lipschitz constant on the discriminator stabilizes the Vanilla GAN. Additionally, it demonstrates the capacity of the Vanilla GAN to detect a lower-dimensional manifold. In order to monitor rates of convergence, it is necessary to at least approximately evaluate $\mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{\hat{G}_n(Z)})$. Therefore, we study the numerical performance of the Vanilla GAN in a simulation setting where the true data distribution is known by construction. In this work, the Wasserstein distance is employed as the metric for measuring the rate of convergence. In practice, the Wasserstein distance is only computable in the one-dimensional case. To investigate multivariate distributions, we approximated the Wasserstein distance by averaging the Wasserstein distance on the marginals.
In order to model the distribution $\mathbb{P}^*$ of a lower-dimensional manifold, we employed a one-dimensional uniform distribution on the graph of the function $x \mapsto \sin(4\pi x)$ on the diagonal of the two-dimensional unit cube, resulting in a three-dimensional distribution. For the latent distribution, we used the one-dimensional standard Gaussian distribution. Consequently, the dimensions of the lower-dimensional manifold and the latent space are identical.



(a) Step 0, 100 obs. (b) Step 15000, 100 obs. (c) Step 30000, 100 obs.

(d) Step 0, 1000 obs. (e) Step 15000, 1000 obs. (f) Step 30000, 1000 obs.

Figure 3.1.: Training of Vanilla GAN with weight clip using 100 observations (first row) or 1000 observations (second row). Red dots show 1000 generated samples, green dots show the observations used for the training. The blue line is the one-dimensional manifold.

For the discriminator, we used a neural network with four layers of width 128 concatenated to a sigmoid function. For the generator, we used a neural network with three layers of width 64. In order to preserve as much alignment as possible with the theoretical result, we used plain ReLU activations. Each training consisted of 30000 training iterations. We used the Adam optimizer (Kingma & Ba, 2014) with parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and a learning rate of $\gamma = 0.0005$. When the number of observations exceeded 512, we used minibatches of that size in each iteration. We updated generator and discriminator alternating.

Three snapshots of the training of the Vanilla GAN for samples sizes $n = 100$ and $n = 1000$ are given in Figure 3.1, respectively. The difference between Figure 3.1c and Figure 3.1f is solely due to the number of observations. The observations in Figure 3.1a to Figure 3.1c cover the manifold to a smaller extent than the observations in Figure 3.1d to Figure 3.1f. This corresponds to a larger stochastic error.

To maintain the Lipschitz constant within a controllable range, we implemented the simple weight clipping mechanism of Arjovsky et al. (2017), limiting each weight to a value of 0.5. It is important to note that the network used in the unclipped case is also Lipschitz continuous, however, we do not have control over this Lipschitz constant. Given the width and depth parameters used in this study, it is evident that the Lipschitz constant of the clipped network remains relatively high and is considerably distinct from the theoretical value typically employed in Wasserstein GANs. However, a smaller Lipschitz constant requires an adjustment to the learning rate. Otherwise the weights are likely to remain at their maximum absolute value. This

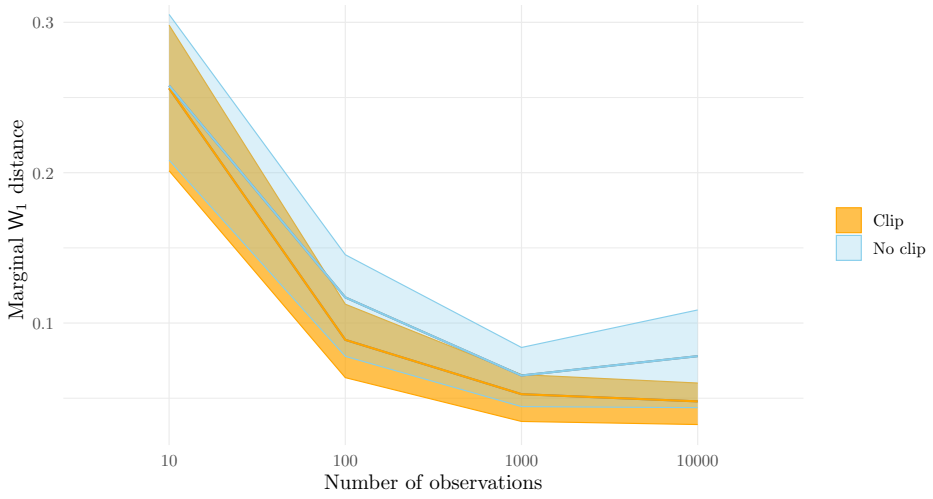


Figure 3.2.: Marginal W_1 distance depending on number of observations. Thick line shows the average over 50 independent runs, ribbons show the first to third quartile.

affects the experiment in several other ways. To ensure a fair and accurate comparison between the clipped and unclipped scenarios, we kept the learning rate consistent.

The results are summarized in Figure 3.2. As predicted by our theory, the averaged marginal Wasserstein distance between the generated distribution and the true data distribution decays approximately as $n^{-1/2}$ for $n \in \{10, 100, 1000\}$. While we see a clear improvement with 10000 observations, the additional gain is limited by the optimization error, since the manifold is already densely covered for 1000 observations.

It is apparent that controlling the Lipschitz constant overall stabilizes the training process, resulting in less variability in the results. In certain cases, the GAN without weight clipping can achieve the same level of effectiveness. This does not negate the outcome. Since the discriminator without clipped weights is still Lipschitz continuous (with a large Lipschitz constant), the theoretical limitations of Vanilla GANs without restricted discriminator classes do not directly translate to practice. This, combined with the finite nature of the implementations, ultimately resulted in the empirical success of these models. The variability between different simulation runs is described by the first to the third quartile in Figure 3.2 which again confirms a more stable behavior of the clipped algorithm.

Figure 3.3 demonstrates the high degree of precision with which the generated samples concentrate on the low-dimensional support of the true data distribution. Our experiments show that this concentration holds true across all sample sizes and can be observed in both the clipped and unclipped case. However, a high concentration does not necessarily indicate that the generated distribution is an accurate imitation of the unknown distribution with respect to the Wasserstein distance. Consequently, Figure 3.3 is only informative in conjunction with Figure 3.2.

Additionally, we investigated the use of a space $\mathbb{U}$ of the same dimension as the ambient space. Our observations indicated that the Vanilla GAN is still capable of identifying the lower-dimensional subspace with reasonable efficacy.

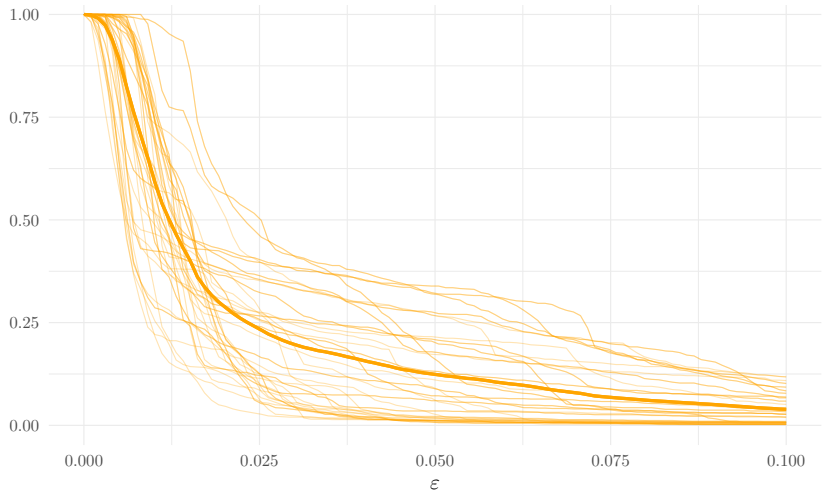


Figure 3.3.: Percentage of generated samples with Euclidean distance to manifold greater than ε using 1000 observations and a discriminator with 0.5 clip. Transparent lines show the individual runs, thick line shows the average over 50 runs.

3.7. Proofs

3.7.1. Proofs of Section 3.1

Proof of Lemma 3.1. Let $X \sim \mathbb{P}^*, \hat{X} \sim \mathbb{P}_n$ and $Z \sim \mathbb{U}$. The symmetry of $\mathcal{W}$ and the Lipschitz continuity of $x \mapsto \log(1+e^{-x})$ yields for any $G \in \mathcal{G}$

$$
\begin{aligned}
&\mathsf{V}_{\mathcal{W}}(\mathbb{P}^*, \mathbb{P}^{\hat{G}_n}(Z)) \\
&= \sup_{W\in\mathcal{W}} \mathbb{E}\Big[-\log\Big(\frac{1+e^{-W(X)}}{2}\Big) + \log\Big(\frac{1+e^{-W(\hat{X})}}{2}\Big) \\
&\qquad\qquad - \log\Big(\frac{1+e^{-W(\hat{X})}}{2}\Big) - \log\Big(\frac{1+e^{W(\hat{G}_n(Z))}}{2}\Big)\Big] \\
&\le \sup_{W\in\mathcal{W}} \mathbb{E}\big[-\log(1+e^{-W(X)}) + \log(1+e^{-W(\hat{X})})\big] + \mathsf{V}_{\mathcal{W}}(\mathbb{P}_n, \mathbb{P}^{\hat{G}_n(Z)}) \\
&\le \sup_{W\in\mathcal{W}} \mathbb{E}\big[-\log(1+e^{-W(X)}) + \log(1+e^{-W(\hat{X})})\big] + \mathsf{V}_{\mathcal{W}}(\mathbb{P}_n, \mathbb{P}^{G(Z)}) \\
&= \sup_{W\in\mathcal{W}} \mathbb{E}[-\log(1+e^{-W(X)}) + \log(1+e^{-W(\hat{X})})] \\
&\quad + \sup_{W\in\mathcal{W}} \mathbb{E}[-\log(1+e^{-W(\hat{X})}) + \log(1+e^{-W(X)})] + \mathsf{V}_{\mathcal{W}}(\mathbb{P}^*, \mathbb{P}^{G(Z)}) \\
&\le 2 \sup_{W\in\mathrm{Lip}(1)\circ\mathcal{W}} \mathbb{E}[W(X) - W(\hat{X})] + \mathsf{V}_{\mathcal{W}}(\mathbb{P}^*, \mathbb{P}^{G(Z)}).
\end{aligned}
$$

The bound for $\hat{G}_n$ from (3.6) follows since $G \in \mathcal{G}$ was arbitrary. □

Proof of Lemma 3.2. Let $(G_n)_{n\in\mathbb{N}} \in \mathcal{G}$ be a sequence that converges to $G \in \mathcal{G}$. If $\mathsf{V}_{\mathrm{Lip}(L)}(\mathbb{P}^*, \mathbb{P}^{G(Z)}) \ge \mathsf{V}_{\mathrm{Lip}(L)}(\mathbb{P}^*, \mathbb{P}^{G_n(Z)})$, then

$$
\mathsf{V}_{\mathrm{Lip}(L)}(\mathbb{P}^*, \mathbb{P}^{G(Z)}) - \mathsf{V}_{\mathrm{Lip}(L)}(\mathbb{P}^*, \mathbb{P}^{G_n(Z)})
$$

$$\begin{aligned}
&\leq \sup_{W\in\mathrm{Lip}(L)} \mathbb{E}\Big[\log\Big(\frac{1+e^{-W(G_n(Z))}}{2}\Big) - \log\Big(\frac{1+e^{-W(G(Z))}}{2}\Big)\Big] \\
&\leq \sup_{W\in\mathrm{Lip}(L)} \mathbb{E}\Big[W(G_n(Z)) - W(G(Z))\Big] \\
&\leq L\|G_n - G\|_\infty.
\end{aligned}$$

The case $\mathsf{V}_{\mathrm{Lip}(L)}(\mathbb{P}^*, \mathbb{P}^{G(Z)}) < \mathsf{V}_{\mathrm{Lip}(L)}(\mathbb{P}^*, \mathbb{P}^{G_n(Z)})$ can be bounded analogously. Therefore, T is continuous and there is at least one minimizer if $\mathcal{G}$ is compact. □

3.7.2. Proofs of Section 3.2

Before we prove the main results from Section 3.2 we require an auxiliary lemma, whose proof can be found in Section 3.7.6:

Lemma 3.14. *For $X \sim \mathbb{P}$ and $Y \sim \mathbb{Q}$ and an arbitrary set of measurable functions $\mathcal{W}$ we have that*

$$\mathsf{V}_{\mathcal{W}}(\mathbb{P}, \mathbb{Q}) \leq \sup_{W\in\mathcal{W}} \mathbb{E}[-\log\big(1+e^{-W(X)}\big) + \log\big(1+e^{-W(Y)}\big)].$$

Proof of Theorem 3.4. Defining

$$\psi\colon \mathbb{R}\to\mathbb{R},\ \psi(x) \coloneqq -\log\Big(\frac{1+e^{-x}}{2}\Big),$$

we can rewrite

$$\mathsf{V}_{\mathrm{Lip}(L,B)}(\mathbb{P}, \mathbb{Q}) = \sup_{W\in\mathrm{Lip}(L,B)} \mathbb{E}[\psi(W(X)) + \psi(-W(Y))].$$

The function $f\colon [-\log(2-2/L), \infty) \to \mathbb{R},\ \ f(x) = \log(2e^x - 1)$ is bijective and Lipschitz continuous with Lipschitz constant L and satisfies $\psi(-f(x)) = x$ for all $x \geq -\log(2-2/L)$. Therefore, we obtain a lower bound

$$\begin{aligned}
\mathsf{V}_{\mathrm{Lip}(L,B)}(\mathbb{P}, \mathbb{Q}) &\geq \sup_{\substack{W\in\mathrm{Lip}(1,\log((1+e^B)/2)) \\ W(\cdot)\geq -\log(2-2/L)}} \mathbb{E}[\psi(f(W(X))) + \psi(-f(W(Y)))] \\
&= \sup_{\substack{W\in\mathrm{Lip}(1,B') \\ W(\cdot)\geq -\log(2-2/L)}} \mathbb{E}[\psi(f(W(X))) - W(Y)].
\end{aligned}$$

Since $f^{-1} \in \mathrm{Lip}(1, \mathbb{R})$, we can estimate $\mathsf{V}_{\mathrm{Lip}(L,B)}$ from above by

$$\begin{aligned}
\mathsf{V}_{\mathrm{Lip}(L,B)}(\mathbb{P}, \mathbb{Q}) &= \sup_{W\in\mathrm{Lip}(L,B)} \mathbb{E}[\psi(f(f^{-1}(W(X)))) + \psi(-f(f^{-1}(W(Y))))] \\
&\leq \sup_{\substack{W\in\mathrm{Lip}(L,B) \\ W(\cdot)>-\log(2)}} \mathbb{E}[\psi(f(W(X))) + \psi(-f(W(Y)))] \\
&= \sup_{\substack{W\in\mathrm{Lip}(L,B) \\ W(\cdot)>-\log(2)}} \mathbb{E}[\psi(f(W(X))) - W(Y)].
\end{aligned}$$

A Taylor approximation at zero of the function $\psi \circ f(x) = \log(2 - e^{-x})$ yields that for every

$x \in (-\log(2), \infty)$ there exists a ξ between x and 0 such that

$$\psi \circ f(x) = x - \frac{e^\xi}{(2e^\xi - 1)^2} x^2.$$

For the lower bound, we thus conclude

$$\mathsf{V}_{\mathrm{Lip}(L,B)}(\mathbb{P}, \mathbb{Q}) \geq \sup_{\substack{W \in \mathrm{Lip}(1,B') \\ W(\cdot) \geq -\log(2-2/L)}} \mathbb{E}[W(X) - W(Y)] - \frac{L(L-1)}{2} \mathbb{E}[W(X)^2].$$

For the upper bound, we get

$$\mathsf{V}_{\mathrm{Lip}(L,B)}(\mathbb{P}, \mathbb{Q}) \leq \sup_{\substack{W \in \mathrm{Lip}(L,B) \\ W(\cdot) > -\log(2)}} \mathbb{E}[W(X) - W(Y)] - \frac{e^B}{(2e^B - 1)^2} \mathbb{E}[W(X)^2]. \qquad \square$$

To prove Theorem 3.5, we again need an auxiliary lemma:

Lemma 3.15. *For $X \sim \mathbb{P}$ and $Y \sim \mathbb{Q}$ and an arbitrary set of measurable functions $\mathcal{W}$ we have that*

$$\mathsf{V}_{\mathcal{W}}(\mathbb{P}, \mathbb{Q}) \leq \sup_{W \in \mathcal{W}} \mathbb{E}[-\log\left(1 + e^{-W(X)}\right) + \log\left(1 + e^{-W(Y)}\right)].$$

Proof of Theorem 3.5. We prove the lower bound first. Theorem 3.4 yields

$$\begin{aligned}
\mathsf{V}_{\mathrm{Lip}(L,B)}(\mathbb{P}, \mathbb{Q}) &\geq \sup_{\substack{W \in \mathrm{Lip}(1,B') \\ W(\cdot) > -\log(2-2/L)}} \mathbb{E}[W(X) - W(Y)] - \frac{L(L-1)}{2} \mathbb{E}[W(X)^2] \\
&\geq \sup_{W \in \mathrm{Lip}(1, \log(2-2/L))} \mathbb{E}[W(X) - W(Y)] - \frac{L(L-1)}{2} \mathbb{E}[W(X)^2].
\end{aligned}$$

Let $W^* \in \operatorname*{arg\,max}_{W \in \mathrm{Lip}(1,\log(2-2/L))} \mathbb{E}[W(X) - W(Y)]$. This element exists by Villani (2008, Theorem 5.10 (iii)). Then $\delta W^* \in \mathrm{Lip}(1, \log(2 - 2/L))$ for all $\delta \in (0, 1]$ and we can conclude

$$\begin{aligned}
&\sup_{W \in \mathrm{Lip}(1,\log(2-2/L))} \mathbb{E}[W(X) - W(Y)] - \frac{L(L-1)}{2} \mathbb{E}[W(X)^2] \\
&\qquad \geq \sup_{\delta \in (0,1]} \left\{ \mathbb{E}[\delta W^*(X) - \delta W^*(Y)] - \frac{L(L-1)}{2} \mathbb{E}[(\delta W^*(X))^2] \right\} \\
&\qquad = \sup_{\delta \in (0,1]} \left\{ \delta \mathbb{E}[W^*(X) - W^*(Y)] - \delta^2 \frac{L(L-1)}{2} \mathbb{E}[(W^*(X))^2] \right\},
\end{aligned}$$

which is independent of B. In case $\Delta := \mathbb{E}[W^*(X) - W^*(Y)] < L(L-1)\mathbb{E}[W^*(X)^2]$ we have for $\delta = \frac{\Delta}{\mathbb{E}[W^*(X)^2] L(L-1)} \in (0, 1)$

$$\begin{aligned}
&\sup_{W \in \mathrm{Lip}(1,\log(2-2/L))} \mathbb{E}[W(X) - W(Y)] - \frac{L(L-1)}{2} \mathbb{E}[W(X)^2] \\
&\qquad \geq \frac{\Delta^2}{\mathbb{E}[W^*(X)^2] L(L-1)} - \frac{\Delta^2}{2\mathbb{E}[W^*(X)^2] L(L-1)} \\
&\qquad = \frac{\Delta^2}{2\mathbb{E}[W^*(X)^2] L(L-1)}
\end{aligned}$$

$$\geq \frac{\Delta^2}{2\log(2-2/L)^2 L(L-1)},$$

where we used $|W^*(x)| \leq \log(2-2/L)$ in the last step. In case $\Delta \geq L(L-1)\mathbb{E}[W^*(X)^2]$ we obtain

$$\mathbb{E}[W^*(X) - W^*(Y)] - \frac{L(L-1)}{2}\mathbb{E}[W^*(X)^2] \geq \frac{1}{2}\mathbb{E}[W^*(X) - W^*(Y)].$$

Using the boundedness of $[0,1]^d$, we get

$$\begin{aligned}\Delta &= \sup_{W \in \mathrm{Lip}(1,\log(2-2/L))} \mathbb{E}[W(X) - W(Y)] \\ &\geq \sup_{W \in \mathrm{Lip}(\log(2-2/L)d^{-1/p},\infty)} \mathbb{E}[W(X) - W(Y)] \\ &= \frac{\log(2-2/L)}{d^{1/p}} \mathsf{W}_1(\mathbb{P},\mathbb{Q}).\end{aligned}$$

Hence we can conclude the claimed lower bound for

$$c_1 = \frac{1}{2}\frac{\log(2-2/L)}{d^{\frac{1}{p}}}, \quad c_2 = \frac{1}{2d^{\frac{2}{p}}L(L-1)}.$$

For the upper bound we use Lemma 3.15 with $\mathcal{W} = \mathrm{Lip}(L)$. Since for $W \in \mathrm{Lip}(L)$ the function $-\log\left(1+e^{-W(\cdot)}\right) \in \mathrm{Lip}(L)$ we conclude

$$\begin{aligned}\mathsf{V}_{\mathrm{Lip}(L,B)}(\mathbb{P},\mathbb{Q}) &\leq \sup_{W \in \mathrm{Lip}(L)} \mathbb{E}[\psi(W(X)) + \psi(W(Y))] \\ &\leq \sup_{W \in \mathrm{Lip}(L)} \mathbb{E}[-\log\left(1+e^{-W(X)}\right) + \log\left(1+e^{-W(Y)}\right)] \\ &\leq \sup_{W \in \mathrm{Lip}(L)} \mathbb{E}[W(X) - W(Y)] \\ &= L \sup_{W \in \mathrm{Lip}(1)} \mathbb{E}[W(X) - W(Y)].\end{aligned}$$

□

Proof of Example 3.6. For the Wasserstein distance we get $\mathsf{W}_1(\mathbb{P},\mathbb{Q}) = \gamma$. The Vanilla GAN distance using all Lipschitz L affine functions as discriminator yields in this example $\mathsf{V}_{a\cdot+b}(\mathbb{P},\mathbb{Q}) = \max_{\substack{a,b\in\mathbb{R},\\ |a|\leq L}} f(a,b)$ for

$$f(a,b) \coloneqq \frac{1}{2}\big(-\log\left(1+e^{-a\gamma-b}\right) - \log\left(1+e^{-a(\gamma+\varepsilon)-b}\right) - \log\left(1+e^{b}\right) - \log\left(1+e^{a\varepsilon+b}\right)\big) + \log(4).$$

Standard calculus yields for fixed a the unique maximizer $b^* = -\frac{a(\varepsilon+\gamma)}{2}$ and

$$f(a,b^*) = -\log\left(1+e^{-\frac{a(\varepsilon+\gamma)}{2}}\right) - \log\left(1+e^{\frac{a(\varepsilon-\gamma)}{2}}\right) + \log(4).$$

Since

$$\frac{\partial}{\partial a} f(a,b^*) = \frac{\varepsilon+\gamma}{2(e^{\frac{a(\gamma+\varepsilon)}{2}}+1)} - \frac{\varepsilon-\gamma}{2(e^{-\frac{a(\varepsilon-\gamma)}{2}}+1)},$$

for $\varepsilon \leq \gamma$, the maximizing a is maximal $a^* = L$. This coincides with the intuitive choice: as the

support of $\mathbb{P}^X$ and the support of $\mathbb{P}^Y$ can be separated by a single point on $\mathbb{R}$, we expect the optimal discriminator to be affine linear. Standard calculus yields the linear upper and lower bound for $\varepsilon = \frac{1}{4}$. For $\varepsilon > \gamma$, the unrestricted maximizing a^* solves the equation

$$(\varepsilon - \gamma)e^{\frac{a^*(\varepsilon+\gamma)}{2}} - (\varepsilon + \gamma)e^{-\frac{a^*(\varepsilon-\gamma)}{2}} = 2\gamma.$$

While there is no closed form solution, a numerical approximation (for $\varepsilon = \frac{1}{4}$) yields for $\gamma < \varepsilon$ and $L > 16$ such that a^* is feasible

$$\frac{\mathsf{W}_1(\mathbb{P}^X, \mathbb{P}^Y)^2}{2} \leq \mathsf{V}_{a\cdot+b}(\mathbb{P}^X, \mathbb{P}^Y) \leq a \cdot \mathsf{W}_1(\mathbb{P}^X, \mathbb{P}^Y)^2. \qquad \square$$

3.7.3. Proofs of Section 3.3

Proof of Theorem 3.7. Using Theorem 3.5 and the triangle inequality for the Wasserstein distance, we deduce for every $G \in \mathcal{G}$ and $c = \max(c_1^{-1}, c_2^{-1/2})$ that

$$\begin{aligned}
\mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{\hat{G}_n(Z)}) &\leq \mathsf{W}_1(\mathbb{P}^*, \mathbb{P}_n) + \mathsf{W}_1(\mathbb{P}_n, \mathbb{P}^{\hat{G}_n(Z)}) \\
&\leq \mathsf{W}_1(\mathbb{P}^*, \mathbb{P}_n) + c\big[\mathsf{V}_{\mathrm{Lip}(L,B)}(\mathbb{P}_n, \mathbb{P}^{\hat{G}_n(Z)})\big]^{1;1/2} \\
&\leq \mathsf{W}_1(\mathbb{P}^*, \mathbb{P}_n) + c\big[\mathsf{V}_{\mathrm{Lip}(L,B)}(\mathbb{P}_n, \mathbb{P}^{G(Z)})\big]^{1;1/2} \\
&\leq \mathsf{W}_1(\mathbb{P}^*, \mathbb{P}_n) + cL\big[\mathsf{W}_1(\mathbb{P}_n, \mathbb{P}^{G(Z)})\big]^{1;1/2} \\
&\leq (1 + cL)\big[\mathsf{W}_1(\mathbb{P}^*, \mathbb{P}_n)\big]^{1;1/2} + cL\big[\mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{G(Z)})\big]^{1;1/2}.
\end{aligned}$$

As $G \in \mathcal{G}$ was arbitrary, we can choose the infimum over $\mathcal{G}$. $\square$

Proof of Corollary 3.8. For every measurable $G^*\colon \mathcal{Z} \to \mathcal{X}$ and any $G \in \mathcal{G}$ we have

$$\begin{aligned}
\mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{\hat{G}_n(Z)}) &\leq \mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{G^*(Z)}) + \mathsf{W}_1(\mathbb{P}^{G^*(Z)}, \mathbb{P}^{G(Z)}) \\
&= \mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{G^*(Z)}) + \sup_{W \in \mathrm{Lip}(1)} \mathbb{E}[W(G^*(Z)) - W(G(Z))] \\
&\leq \mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{G^*(Z)}) + \mathbb{E}[|G^*(Z) - G(Z)|_p] \\
&\leq \mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{G^*(Z)}) + \|G^* - G\|_\infty.
\end{aligned}$$

Since G^* was arbitrary, Theorem 3.7 yields for some constant c

$$\begin{aligned}
\mathbb{E}[\mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{G(Z)})] \leq& c \cdot \mathbb{E}[\max(\sqrt{\mathsf{W}_1(\mathbb{P}_n, \mathbb{P}^*)}, \mathsf{W}_1(\mathbb{P}_n, \mathbb{P}^*))] \\
&+ c \cdot \inf_{G^*\colon \mathcal{Z} \to \mathcal{X}} \Big\{[\mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{G^*(Z)})]^{1;1/2} + [\inf_{G \in \mathcal{G}} \|G - G^*\|_\infty^{1;1/2}]\Big\}.
\end{aligned}$$

Here the infimum can be used as we can increase the constant c multiplied with both terms by an arbitrary small $\varepsilon > 0$ to account for the possibly infinitesimal smaller value. Using Jensen's inequality, we can bound the stochastic error term by

$$\begin{aligned}
\mathbb{E}[\max(\sqrt{\mathsf{W}_1(\mathbb{P}_n, \mathbb{P}^*)}, \mathsf{W}_1(\mathbb{P}_n, \mathbb{P}^*))] &\leq \mathbb{E}[\sqrt{\mathsf{W}_1(\mathbb{P}_n, \mathbb{P}^*)}] + \mathbb{E}[\mathsf{W}_1(\mathbb{P}_n, \mathbb{P}^*)] \\
&\leq \sqrt{\mathbb{E}[\mathsf{W}_1(\mathbb{P}_n, \mathbb{P}^*)]} + \mathbb{E}[\mathsf{W}_1(\mathbb{P}_n, \mathbb{P}^*)].
\end{aligned}$$

From Schreuder (2020, Theorem 4) we know

$$\mathbb{E}[\mathsf{W}_1(\mathbb{P}^*, \mathbb{P}_n)] \leq c' \begin{cases} n^{-1/d}, & d > 2, \\ n^{-1/2} \log(n), & d = 2, \\ n^{-1/2}, & d = 1. \end{cases}$$

where c depends only on d. Since $(\log n)/\sqrt{n} \leq 1$, we conclude

$$\sqrt{\mathbb{E}[\mathsf{W}_1(\mathbb{P}_n, \mathbb{P}^*)]} + \mathbb{E}[\mathsf{W}_1(\mathbb{P}_n, \mathbb{P}^*)] \leq 2c' \begin{cases} n^{-1/2d}, & d > 2, \\ n^{-1/4} (\log n)^{1/2}, & d = 2, \\ n^{-1/4}, & d = 1. \end{cases}$$

This finishes the proof. □

Proof of Theorem 3.9. With the same reasoning as in the proof of Corollary 3.8, there exists some c such that for any measurable $G^* \colon \mathcal{Z} \to \mathcal{X}$ and any $G \in \mathcal{G}$

$$\begin{aligned} \mathbb{E}[\mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{G(Z)})] \leq c\big(& \sqrt{\mathbb{E}[\mathsf{W}_1(\mathbb{P}_n, \mathbb{P}^*)]} + \mathbb{E}[\mathsf{W}_1(\mathbb{P}_n, \mathbb{P}^*)] \\ & + [\mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{G^*(Z)})]^{1;1/2} + [\inf_{G \in \mathcal{G}} \|G^* - G\|_\infty]^{1,1/2} \big). \end{aligned}$$

By the triangle inequality

$$\mathsf{W}_1(\mathbb{P}_n, \mathbb{P}^*) \leq \mathsf{W}_1(\mathbb{P}_n, \mathbb{P}^{G^*(Z)}) + \mathsf{W}_1(\mathbb{P}^{G^*(Z)}, \mathbb{P}^*).$$

Let $Z_i \sim \mathbb{U}$ be i.i.d. random variables and denote the corresponding empirical measure by $\mathbb{U}_n$. For $G^* \in \mathrm{Lip}(M, \mathcal{Z})$ we can then bound the first term by

$$\begin{aligned} \mathsf{W}_1(\mathbb{P}_n, \mathbb{P}^{G^*(Z)}) &= \sup_{W \in \mathrm{Lip}(1)} \frac{1}{n} \sum_{i=1}^n W(X_i) - \mathbb{E}[W \circ G^*(Z)] \\ &\leq \sup_{W \in \mathrm{Lip}(1)} \frac{1}{n} \sum_{i=1}^n |W(X_i) - W \circ G^*(Z_i)| \qquad (3.10) \\ &\quad + \sup_{W \in \mathrm{Lip}(1)} \frac{1}{n} \sum_{i=1}^n W \circ G^*(Z_i) - \mathbb{E}[W \circ G^*(Z)] \\ &\leq \frac{1}{n} \sum_{i=1}^n |X_i - G^*(Z_i)|_p + \sup_{f \in \mathrm{Lip}(M)} \frac{1}{n} \sum_{i=1}^n f(Z_i) - \mathbb{E}[f(Z)] \qquad (3.11) \\ &= \frac{1}{n} \sum_{i=1}^n |X_i - G^*(Z_i)|_p + M \cdot \mathsf{W}_1(\mathbb{U}_n, \mathbb{U}). \end{aligned}$$

Hence,

$$\mathbb{E}\big[\mathsf{W}_1(\mathbb{P}_n, \mathbb{P}^{G^*(Z)})\big] \leq \frac{1}{n} \sum_{i=1}^n \mathbb{E}[|X_i - G^*(Z_i)|_p] + M \cdot \mathbb{E}[W_1(\mathbb{U}_n, \mathbb{U})].$$

Note that $\mathbb{E}[|X_i - G^*(Z_i)|_p] = \mathsf{W}_1(\mathbb{P}^{G^*(Z)}, \mathbb{P}^*)$ by the duality formula of W_1 used in this work, see

Villani (2008, Definition 6.2 and Remark 6.5). For $\mathbb{E}[W_1(\mathbb{U}_n, \mathbb{U})]$, we can exploit the convergence rate for the empirical distribution as in Corollary 3.8, but now in the d^*-dimensional latent space $\mathcal{Z}$. Therefore, there exists a c' such that

$$\sqrt{\mathbb{E}[\mathsf{W}_1(\mathbb{P}_n, \mathbb{P}^*)]} + \mathbb{E}[\mathsf{W}_1(\mathbb{P}_n, \mathbb{P}^*)] \leq c'[\mathsf{W}_1(\mathbb{P}^{G^*(Z)}, \mathbb{P}^*)]^{1;1/2} + c' \begin{cases} n^{-1/2d^*}, & d^* > 2, \\ n^{-1/4}(\log n)^{1/2}, & d^* = 2, \\ n^{-1/4}, & d^* = 1. \end{cases}$$

This concludes the proof. □

3.7.4. Proofs of Section 3.4

Proof of Theorem 3.10. First, we verify that for any two nonempty sets $\mathcal{W}_1$ and $\mathcal{W}_2$ we have

$$\mathsf{V}_{\mathcal{W}_1}(\mathbb{P}, \mathbb{Q}) \leq \mathsf{V}_{\mathcal{W}_2}(\mathbb{P}, \mathbb{Q}) + 2 \inf_{W \in \mathcal{W}_2} \sup_{W^* \in \mathcal{W}_1} \|W - W^*\|_\infty. \tag{3.12}$$

Indeed, the difference $\mathsf{V}_{\mathcal{W}_1}(\mathbb{P}, \mathbb{Q}) - \mathsf{V}_{\mathcal{W}_2}(\mathbb{P}, \mathbb{Q})$ is bounded by

$$\begin{aligned}
&\inf_{W \in \mathcal{W}_2} \sup_{W^* \in \mathcal{W}_1} \Big\{ \mathbb{E}\Big[-\log\Big(\frac{1 + e^{-W^*(X)}}{2}\Big) - \log\Big(\frac{1 + e^{W^*(Y)}}{2}\Big)\Big] \\
&\qquad\qquad - \mathbb{E}\Big[-\log\Big(\frac{1 + e^{-W(X)}}{2}\Big) - \log\Big(\frac{1 + e^{W(Y)}}{2}\Big)\Big] \Big\} \\
\leq &\inf_{W \in \mathcal{W}_2} \sup_{W^* \in \mathcal{W}_1} \Big\{ \mathbb{E}\Big[\Big| -\log\Big(\frac{1 + e^{-W^*(X)}}{2}\Big) + \log\Big(\frac{1 + e^{-W(X)}}{2}\Big)\Big|\Big] \\
&\qquad\qquad + \mathbb{E}\Big[\Big| -\log\Big(\frac{1 + e^{W^*(Y)}}{2}\Big) + \log\Big(\frac{1 + e^{W(Y)}}{2}\Big)\Big|\Big] \Big\} \\
&\leq \inf_{W \in \mathcal{W}_2} \sup_{W^* \in \mathcal{W}_1} \big\{ \mathbb{E}[|W^*(X) - W(X)|] + \mathbb{E}[|W^*(Y) - W(Y)|] \big\} \\
&\leq 2 \inf_{W \in \mathcal{W}_2} \sup_{W^* \in \mathcal{W}_1} \|W^* - W\|_\infty,
\end{aligned}$$

due to Lipschitz continuity of $x \mapsto -\log((1 + e^x)/2)$. From (3.12) we deduce for $\mathcal{W} \subset \mathrm{Lip}(L, B)$

$$\mathsf{V}_{\mathrm{Lip}(L,B)}(\mathbb{P}, \mathbb{Q}) \leq \mathsf{V}_{\mathcal{W}}(\mathbb{P}, \mathbb{Q}) + 2 \inf_{W' \in \mathcal{W}} \sup_{W \in \mathrm{Lip}(L,B)} \|W - W'\|_\infty.$$

We abbreviate $\Delta_{\mathcal{W}} := \inf_{W' \in \mathcal{W}} \sup_{W \in \mathrm{Lip}(L,B)} \|W - W'\|_\infty$. Now we can proceed as in Theorem 3.7. In particular, it is sufficient to bound $\mathsf{W}_1(\mathbb{P}_n, \mathbb{P}^{\hat{G}_n(Z)})$. Due to Theorem 3.5 there is some constant $c > 0$ such that for every $G \in \mathcal{G}$

$$\begin{aligned}
\mathsf{W}_1(\mathbb{P}_n, \mathbb{P}^{\hat{G}_n(Z)}) &\leq c[\mathsf{V}_{\mathrm{Lip}(L,B)}(\mathbb{P}_n, \mathbb{P}^{\hat{G}_n(Z)})]^{1;1/2} \\
&\leq c[\mathsf{V}_{\mathcal{W}}(\mathbb{P}_n, \mathbb{P}^{\hat{G}_n(Z)}) + 2\Delta_{\mathcal{W}}]^{1;1/2} \\
&\leq c[\mathsf{V}_{\mathcal{W}}(\mathbb{P}_n, \mathbb{P}^{\hat{G}_n(Z)})]^{1;1/2} + 2c[\Delta_{\mathcal{W}}]^{1;1/2} \\
&\leq c[\mathsf{V}_{\mathcal{W}}(\mathbb{P}_n, \mathbb{P}^{G(Z)})]^{1;1/2} + 2c[\Delta_{\mathcal{W}}]^{1;1/2}.
\end{aligned}$$

Because $\mathsf{V}_{\mathcal{W}}(\mathbb{P}_n, \mathbb{P}^{G(Z)}) \le \mathsf{V}_{\mathrm{Lip}(L,B)}(\mathbb{P}_n, \mathbb{P}^{G(Z)})$ due to $\mathcal{W} \subset \mathrm{Lip}(L,B)$, the rest of the proof is identical to the proof of Theorem 3.7. □

Proof of Theorem 3.11. Since $\mathsf{W}_1(\mathbb{P}_n, \mathbb{P}^*)$ can be estimated as in Theorem 3.9, we only need to bound $\mathsf{W}_1(\mathbb{P}_n, \mathbb{P}^{\hat{G}_n(Z)})$. For $\Gamma > \max(L, 2B)$, we have $\mathrm{Lip}(L,B) \subset \mathcal{H}^\alpha(\Gamma)$, $\alpha \in (0,1)$, and the assumptions of Theorem 3.5 are satisfied. Therefore for every $\alpha \in (0,1)$

$$\mathsf{W}_1(\mathbb{P}_n, \mathbb{P}^{\hat{G}_n(Z)}) \le c\big[\mathsf{V}_{\mathrm{Lip}(L,B)}(\mathbb{P}_n, \mathbb{P}^{\hat{G}_n(Z)})\big]^{1;1/2} \le c\big[\mathsf{V}_{\mathcal{H}^\alpha(\Gamma)}(\mathbb{P}_n, \mathbb{P}^{\hat{G}_n(Z)})\big]^{1;1/2}.$$

Now, (3.12) yields

$$\mathsf{V}_{\mathcal{H}^\alpha(\Gamma)}(\mathbb{P}_n, \mathbb{P}^{\hat{G}_n(Z)}) \le \mathsf{V}_{\mathcal{W}}(\mathbb{P}_n, \mathbb{P}^{\hat{G}_n(Z)}) + 2\Delta_{\mathcal{W}} \qquad \text{for} \qquad \Delta_{\mathcal{W}} := \inf_{W \in \mathcal{W}} \sup_{W^* \in \mathcal{H}^\alpha(\Gamma)} \|W^* - W\|_\infty.$$

Using that $\hat{G}_n$ is the empirical risk minimizer and $\mathcal{W} \subseteq \mathcal{H}^\alpha(\Gamma)$, we thus have

$$\begin{aligned}
\mathsf{W}_1(\mathbb{P}_n, \mathbb{P}^{\hat{G}_n(Z)}) &\le c\big[\mathsf{V}_{\mathcal{W}}(\mathbb{P}_n, \mathbb{P}^{\hat{G}_n(Z)})\big]^{1;1/2} + c[\Delta_{\mathcal{W}}]^{1;1/2} \\
&\le c\big[\mathsf{V}_{\mathcal{W}}(\mathbb{P}_n, \mathbb{P}^{G(Z)})\big]^{1;1/2} + c[\Delta_{\mathcal{W}}]^{1;1/2} \\
&\le c\big[\mathsf{V}_{\mathcal{H}^\alpha(\Gamma)}(\mathbb{P}_n, \mathbb{P}^{G(Z)})\big]^{1;1/2} + c[\Delta_{\mathcal{W}}]^{1;1/2}.
\end{aligned}$$

To bound the first term, we apply Lemma 3.15 and $\{-\log(1 + e^{-W(\cdot)}) \mid W \in \mathcal{H}^\alpha(\Gamma)\} \subset \mathcal{H}^\alpha(\Gamma)$ to obtain

$$\begin{aligned}
\mathsf{V}_{\mathcal{H}^\alpha(\Gamma)}(\mathbb{P}_n, \mathbb{P}^{G(Z)}) &\le \sup_{W \in \mathcal{H}^\alpha(\Gamma)} \mathbb{E}_{\hat{X} \sim \mathbb{P}_n}\big[-\log\big(1 + e^{-W(\hat{X})}\big) + \log\big(1 + e^{-W(G(Z))}\big)\big] \\
&\le \sup_{W \in \mathcal{H}^\alpha(\Gamma)} \mathbb{E}_{\hat{X} \sim \mathbb{P}_n}[W(\hat{X}) - W(G(Z))] \qquad (3.13) \\
&\le \sup_{W \in \mathcal{H}^\alpha(\Gamma)} \mathbb{E}_{\hat{X} \sim \mathbb{P}_n}[W(\hat{X}) - W(X)] + \sup_{W \in \mathcal{H}^\alpha(\Gamma)} \mathbb{E}[W(X) - W(G(Z))].
\end{aligned}$$

For the second term we have by Hölder continuity, Jensen's inequality and the duality formula of W_1 as used in the proof of Theorem 3.9 that

$$\begin{aligned}
\sup_{W \in \mathcal{H}^\alpha(\Gamma)} \mathbb{E}[W(X) - W(G(Z))] &\le \sup_{W \in \mathcal{H}^\alpha(\Gamma)} \mathbb{E}[|W(X) - W(G^*(Z))|] \\
&\quad + \sup_{W \in \mathcal{H}^\alpha(\Gamma)} \mathbb{E}[|W(G^*(Z)) - W(G(Z))|] \\
&\le \Gamma \mathbb{E}[|X - G^*(Z)|_p^\alpha] + \Gamma \|G^* - G\|_\infty^\alpha \\
&\le \Gamma \mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{G^*(Z)})^\alpha + \Gamma \|G^* - G\|_\infty^\alpha.
\end{aligned}$$

Hence, we have for any $G \in \mathcal{G}$ and any measurable $G^* \colon \mathcal{Z} \to \mathcal{X}$ for some constant $c > 0$

$$\begin{aligned}
\mathsf{W}_1(\mathbb{P}_n, \mathbb{P}^{\hat{G}_n(Z)}) \le{}& c\big[\sup_{W \in \mathcal{H}^\alpha(\Gamma)} \mathbb{E}_{\hat{X} \sim \mathbb{P}_n}[W(\hat{X}) - W(X)]\big]^{1;1/2} \\
&+ c\big[\mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{G^*(Z)})^\alpha + \|G^* - G\|_\infty^\alpha\big]^{1;1/2} + c[\Delta_{\mathcal{W}}]^{1;1/2}.
\end{aligned}$$

For the remaining stochastic error term, we first note that

$$\begin{aligned}\sup_{W\in\mathcal{H}^\alpha(\Gamma)} \mathbb{E}_{\hat{X}\sim\mathbb{P}_n}[W(\hat{X})-W(X)] &\le \sup_{W\in\mathcal{H}^\alpha(\Gamma)} \mathbb{E}_{X_n\sim\mathbb{P}_n}[W(X_n)-W(G^*(Z))] \\ &\quad + \sup_{W\in\mathcal{H}^\alpha(\Gamma)} \mathbb{E}[W(G^*(Z))-W(X)] \\ &\le \sup_{W\in\mathcal{H}^\alpha(\Gamma)} \mathbb{E}_{X_n\sim\mathbb{P}_n}[W(X_n)-W(G^*(Z))] + \Gamma \mathsf{W}_1(\mathbb{P}^*,\mathbb{P}^{G^*(Z)})^\alpha\end{aligned}$$

and as in (3.11) together with Schreuder (2020, Theorem 4) we obtain

$$\begin{aligned}\mathbb{E}\Big[\sup_{W\in\mathcal{H}^\alpha(\Gamma)} \mathbb{E}_{X_n\sim\mathbb{P}_n}[W(X_n)-W(G^*(Z))]\Big] &\le \mathbb{E}\Big[\sup_{W\in\mathcal{H}^\alpha(\Gamma)} |X-G^*(Z)|_p^\alpha\Big] \\ &\quad + \mathbb{E}\Big[\sup_{f\in\mathcal{H}^\alpha(M\cdot\Gamma)} \frac{1}{n}\sum_{i=1}^n f(Z_i)-\mathbb{E}[f(Z)]\Big] \\ &\le c\mathsf{W}_1(\mathbb{P}^*,\mathbb{P}^{G^*(Z)})^\alpha + c\begin{cases} n^{-\alpha/d^*}, & 2\alpha<d^*, \\ n^{-1/2}\ln(n), & 2\alpha=d^*, \\ n^{-1/2}, & 2\alpha>d^*. \end{cases}\end{aligned}$$

For the expectation of the first term we use Jensen's inequality

$$\mathbb{E}[|X_i-G^*(Z_i)|_p^\alpha] \le \mathbb{E}[|X_i-G^*(Z_i)|_p]^\alpha = \mathsf{W}_1(\mathbb{P}^*,\mathbb{P}^{G^*(Z)})^\alpha. \qquad \square$$

3.7.5. Proof of Section 3.5

Proof of Theorem 3.13. First we note that for $\Gamma > 1$, there is an $L > 0$, such that there is a $B > 0$ with $2B < \Gamma - 1$ and with $\hat{X} \sim \mathbb{P}^*$

$$\sup_{W\in\mathrm{Lip}(L)} \mathbb{E}[W(\hat{X})-W(\hat{G}_n(Z))] = \sup_{W\in\mathrm{Lip}(L,2B)} \mathbb{E}[W(\hat{X})-W(\hat{G}_n(Z))].$$

This $L > 0$ exists as $[0,1]^d$ is bounded and adding a constant to any function $W \in \mathrm{Lip}(L)$ will not change the value of $\mathbb{E}[W(\hat{X}) - W(\hat{G}_n(Z))]$.
Then we get for every $G \in \mathcal{G}$ with the same reasoning as in the proof of Theorem 3.10

$$\begin{aligned}\mathsf{W}_1(\mathbb{P}^*,\mathbb{P}^{\hat{G}_n(Z)}) &\le \mathsf{W}_1(\mathbb{P}^*,\mathbb{P}_n) + \mathsf{W}_1(\mathbb{P}_n,\mathbb{P}^{\hat{G}_n(Z)}) \\ &= \mathsf{W}_1(\mathbb{P}^*,\mathbb{P}_n) + \frac{1}{L}\mathsf{W}_L(\mathbb{P}_n,\mathbb{P}^{\hat{G}_n(Z)}) \\ &\le \mathsf{W}_1(\mathbb{P}^*,\mathbb{P}_n) + \frac{1}{L}\mathsf{W}_{\mathcal{W}}(\mathbb{P}_n,\mathbb{P}^{\hat{G}_n(Z)}) + \frac{2}{L}\inf_{W\in\mathcal{W}}\sup_{W'\in\mathrm{Lip}(L,2B)}\|W-W'\|_\infty \\ &\le \mathsf{W}_1(\mathbb{P}^*,\mathbb{P}_n) + \frac{1}{L}\mathsf{W}_{\mathcal{W}}(\mathbb{P}_n,\mathbb{P}^{G(Z)}) + \frac{2}{L}\inf_{W\in\mathcal{W}}\sup_{W'\in\mathrm{Lip}(L,2B)}\|W-W'\|_\infty \\ &\le \mathsf{W}_1(\mathbb{P}^*,\mathbb{P}_n) + \frac{1}{L}\mathsf{W}_{\mathcal{H}^\alpha}(\mathbb{P}_n,\mathbb{P}^{G(Z)}) + \frac{2}{L}\inf_{W\in\mathcal{W}}\sup_{W'\in\mathrm{Lip}(L,2B)}\|W-W'\|_\infty.\end{aligned}$$

The bound on $\mathsf{W}_{\mathcal{H}^\alpha}(\mathbb{P}_n,\mathbb{P}^{G(Z)})$ depending on the intrinsic dimension d^* was already derived

in Theorem 3.11 (starting with Equation (3.13)). The bound on $\mathsf{W}_1(\mathbb{P}^*, \mathbb{P}_n)$ depending on the intrinsic dimension d^* was already derived in Corollary 3.8. □

3.7.6. Additional proofs of Section 3.7

Proof of Lemma 3.14. Since

$$\log(1+e^x) + \log(1+e^{-x}) \geq \log(4) \quad \text{for all } x \in \mathbb{R},$$

we can bound

$$\begin{aligned}
&\sup_{W\in\mathcal{W}} \mathbb{E}\Big[-\log\big(1+e^{-W(X)}\big) - \log\big(1+e^{W(Y)}\big)\Big] \\
&\quad = \sup_{W\in\mathcal{W}} \mathbb{E}[-\log\big(1+e^{-W(X)}\big) + \log\big(1+e^{-W(Y)}\big) - \log\big(1+e^{-W(Y)}\big) - \log\big(1+e^{W(Y)}\big)] \\
&\quad \leq \sup_{W\in\mathcal{W}} \mathbb{E}[-\log\big(1+e^{-W(X)}\big) + \log\big(1+e^{-W(Y)}\big)] \\
&\qquad - \inf_{W\in\mathcal{W}} \mathbb{E}[\log\big(1+e^{-W(Y)}\big) + \log\big(1+e^{W(Y)}\big)] \\
&\quad \leq \sup_{W\in\mathcal{W}} \mathbb{E}[-\log\big(1+e^{-W(X)}\big) + \log\big(1+e^{-W(Y)}\big)] - \log(4).
\end{aligned}$$

□

Proof of Lemma 3.15. Since

$$\log(1+e^x) + \log(1+e^{-x}) \geq \log(4) \quad \text{for all } x \in \mathbb{R},$$

we can bound

$$\begin{aligned}
&\sup_{W\in\mathcal{W}} \mathbb{E}\Big[-\log\big(1+e^{-W(X)}\big) - \log\big(1+e^{W(Y)}\big)\Big] \\
&\quad = \sup_{W\in\mathcal{W}} \mathbb{E}[-\log\big(1+e^{-W(X)}\big) + \log\big(1+e^{-W(Y)}\big) - \log\big(1+e^{-W(Y)}\big) - \log\big(1+e^{W(Y)}\big)] \\
&\quad \leq \sup_{W\in\mathcal{W}} \mathbb{E}[-\log\big(1+e^{-W(X)}\big) + \log\big(1+e^{-W(Y)}\big)] \\
&\qquad\qquad - \inf_{W\in\mathcal{W}} \mathbb{E}[\log\big(1+e^{-W(Y)}\big) + \log\big(1+e^{W(Y)}\big)] \\
&\quad \leq \sup_{W\in\mathcal{W}} \mathbb{E}[-\log\big(1+e^{-W(X)}\big) + \log\big(1+e^{-W(Y)}\big)] - \log(4).
\end{aligned}$$

□

Chapter 4

Kernel density estimation

Of course, distribution estimation is not a novelity of generative models. In case the distribution admits a density with respect to the Lebesgue measure, the kernel density or Parzen-Rosenblatt estimator (Rosenblatt, 1956; Parzen, 1962) is the classical method for estimating a smooth density. Given an i.i.d. sample $X_1, ..., X_n$ of a distribution $\mathbb{P}^*$ on $\mathbb{R}^d$, $d \in \mathbb{N}$ and a kernel function $K\colon \mathbb{R}^d \to \mathbb{R}$ which is itself a density of a distribution $\mathbb{U}$ on $\mathbb{R}^d$, we define the kernel density estimator (KDE) as

$$p_n(x) \coloneqq \frac{1}{n \cdot h^d} \sum_{i=1}^{n} K\big(\frac{x - X_i}{h}\big),$$

where $h > 0$ is a bandwidth parameter. The assumption that K is a density guarantees that p_n is a density. In the literature, kernels that are not densities are also discussed, see for example Tsybakov (2009), but we will limit the possible kernels to densities. The reason for this will become apparent in Chapter 5.
Typically, the KDE is analyzed using the expectation of the L_1 distance, $\|p_n - p^*\|_1$, or the mean squared error (Tsybakov, 2009; Devroye & Lugosi, 2012). For the former, we known from Scheffé's lemma

$$\|p_n - p^*\|_1 = 2\mathsf{TV}(\mathbb{P}^*, \mathbb{P}^{\mathrm{KDE}}),$$

where $\mathbb{P}^{\mathrm{KDE}}$ is the distribution corresponding to the density p_n. The mean squared error is defined as

$$\mathrm{MSE}(\mathbb{P}^*, \mathbb{P}^{\mathrm{KDE}}) \coloneqq \int (p_n(x) - p^*(x))^2 \, \mathrm{d}x.$$

Hence, both errors operate on the level of densities. To incorporate the KDE in our analysis in Chapter 5, we need to bound the error in the Wasserstein-1 distance. This will be the goal of this chapter.

Related work For an overview of classical analysis in the univariate case, see Tsybakov (2009), Chapter 1.2 or Devroye & Lugosi (2012) and in the multivariate case see Scott (1992). Despite its longstanding use, kernel density estimation is still the subject of ongoing research, for example in density estimation on unknown manifolds (Berry & Sauer, 2017; Berenfeld & Hoffmann, 2021; Divol, 2022; Wu & Wu, 2022). In particular, Divol (2022) uses the Wasserstein

metric to evaluate the performance of a kernel density estimator, however he needs to assume properties of the kernel that are not satisfied by standard choices such as the Gaussian kernel.

Own contribution In this chapter, we are going to derive a rate of convergence of the KDE in Wasserstein-1 distance. We start by showing rates for compact kernels implied by existing rates in the L_1 distance. Then we show that we can improve these rates and allow for unbounded kernels. Afterwards, we show using a smoothness condition that the KDE can attain minimax optimal rates up to logarithmic constants. The assumptions made are satisfied by the Gaussian kernel. In the end, we show that we can circumvent the curse of dimensionality in case the unknown distribution is supported on a low-dimensional linear subspace.

4.1. Rate of convergence

We start by showing that for compact kernels, existing results for the L_1-error of the KDE can be exploited.

Theorem 4.1. *Assume* $\operatorname{supp}(p^*)$ *is compact and* $p^* \in \mathcal{H}^\alpha(C)$. *Further assume that* K *is such that* $\operatorname{supp}(K)$ *is compact,*

$$\int_{\mathbb{R}^d} |K(x)|dx < \infty \quad \text{and} \quad \int_{\mathbb{R}^d} K(x)dx = 1,$$

as well as

$$\int_{\mathbb{R}^d} |K(z)|^2 dz < \infty. \tag{4.1}$$

Then there exist constants $c_1, c_2 > 0$ *such that*

$$\mathbb{E}[\mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{\mathrm{KDE}})] \leq \frac{c_1}{\sqrt{n \cdot h^d}} + c_2 \cdot C \cdot h^\alpha.$$

This leads to a rate of convergence of $n^{-\frac{\alpha}{2\alpha+d}}$ using $h \asymp n^{-\frac{1}{2\alpha+d}}$. This rate can be improved exploiting the Lipschitz 1 smoothness of the test function in the dual form of the Wasserstein-1 distance (2.15). Additionally, the assumption that K is compact makes Theorem 4.1 invalid in case of standard kernels such as the Gaussian kernel. The following result will improve the rate for d not too small, allowing for noncompact kernels.

Theorem 4.2. *Assume* $p^* \in B^\alpha_{1,\infty}(M, \mathbb{R}^d), \alpha \in (0, 1], M \in \mathbb{R}_{>0}$. *Further assume* K *is a nonnegative* d*-dimensional kernel such that* $\int K(y)y \, \mathrm{d}y = 0$,

$$\int \max(|u|^2, |u|^{d+4}) K^2(u) \, \mathrm{d}u < \infty \quad \text{and} \quad \int |x|^{d+4} p^*(x) \, \mathrm{d}x < \infty.$$

If $h \leq 1$, *then there are* C_1 *and* C_2 *such that*

$$\mathbb{E}\big[\mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{\mathrm{KDE}})\big] \leq C_1 h^{1+\alpha} + \frac{C_2}{\sqrt{nh^d}}.$$

Choosing $h \asymp n^{-\frac{1}{2+2\alpha+d}}$ leads to the convergence rate $n^{-\frac{1+\alpha}{2+2\alpha+d}}$. For nonzero α and d big enough, this rate decays faster than $n^{-\frac{1}{d}}$, the rate of convergence of the empirical measure (Dudley, 1969; Boissard & Gouic, 2014). Compared to the rate implied by Theorem 4.1, we gain in the numerator of the exponent but lose a bit in the denominator. For d large enough, this improves the rate drastically. However, the following result shows that in case that p^* has bounded support, imposing a differentiability assumption on the kernel can improve the denominator up to the level of Theorem 4.1 while keeping the numerator at the level of Theorem 4.2. This comes at the cost of a logarithmic term.

Theorem 4.3. *Assume $d \geq 2, M \in \mathbb{R}_{>0}, p^* \in B^\alpha_{1,\infty}(M, \mathbb{R}^d), \alpha \in (0, 1]$ and $\operatorname{supp}(p^*)$ bounded. Assume K is a nonnegative d-dimensional kernel such that*

$$\int yK(y)\,\mathrm{d}y = 0, \quad \text{and} \quad \int |y|^{1+\alpha} K(y)\,\mathrm{d}y < \infty.$$

Further assume $K \in C^{\frac{d+2}{2}}$ with $\|D^k K\|_1 \leq C$ for $k \in \mathbb{N}_0^d$ such that $|k| \leq \frac{d+2}{2}$ and some $C > 0$. Then there are C_1 and C_2 such that

$$\mathbb{E}\big[\mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{\mathrm{KDE}})\big] \leq C_1 h^{1+\alpha} + \frac{C_2}{\sqrt{h^{d-2}}} \frac{\log n}{\sqrt{n}}.$$

For $h \asymp (n/\log^2 n)^{-\frac{1}{2\alpha+d}}$ we obtain $\mathbb{E}\big[\mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{\mathrm{KDE}})\big] = \mathcal{O}((n/\log^2 n)^{-\frac{1+\alpha}{2\alpha+d}})$.

The convergence rate $(n/\log^2 n)^{-\frac{1+\alpha}{2\alpha+d}}$ coincides up to the logarithmic factor with the lower bound by Niles-Weed & Berthet (2022, Theorem 3) and thus the above rate is minimax optimal up to the logarithm. Divol (2022) also considers rates of convergence for kernel density estimators. In case of $d \neq 2$, he obtains the optimal rate for distributions bounded away from zero. However, the kernels he considers must be smooth radial functions with bounded support in $(0, 1)^d$ (Divol, 2022, Condition A). Hence, his result does not apply to the Gaussian kernel.

Remark 4.4.

1. *The assumption $\|D^k K\|_1 \leq C$ for $k \in \mathbb{N}_0^d$ and $C > 0$ is satisfied by the Gaussian kernel.*
2. *The second term in the proof of Theorem 4.3 is the expected value of $\mathsf{W}_1(K_h * \mathbb{P}_n, K_h * \mathbb{P}^*)$. Results on the convergence of the smooth empirical measure have been discussed in case of the Gaussian kernel. Goldfeld et al. (2020, Proposition 1) obtain a rate in $\mathcal{O}\big(\frac{1}{\sqrt{nh^d}}\big)$ for subgaussian distributions $\mathbb{P}^*$, which coincides with the rate obtained in Theorem 4.2. Weed (2018) looks at distributions on $[-1, 1]$ and obtains a rate in $\mathcal{O}\big(\frac{1}{\sqrt{nh^d}}\big)$. Hence the results above improves these results for distributions in $B^\alpha_{1,\infty}$ on bounded support.*

4.2. Dimension reduction

The rate in Theorem 4.3 depends on the dimension d and therefore exhibits the classical curse of dimensionality. In the following, we analyze a case in which the curse of dimensionality can be circumvented and the rate depends on some intrinsic dimension $d' < d$. Assume that the

support of $\mathbb{P}^*$ is a subset of an d'-dimensional linear subspace of $\mathbb{R}^d$. Let the linear subspace V be defined by

$$V \coloneqq \{x \in \mathbb{R}^d \colon \exists y \in \mathbb{R}^{d'} \text{ s.t. } x = Ay\},$$

where A is a $\mathbb{R}^{d\times d'}$ matrix with normalized orthogonal columns. Let $p^*_{d'}$ be the density with respect to the Lebesgue measure on $\mathbb{R}^{d'}$, such that for $Y \sim p^*_{d'}$ we have that $AY \sim \mathbb{P}^*$. This setting has been studied in context of diffusions by Chen et al. (2023b) and Oko et al. (2023). In this case we can improve the rate of convergence, such that the corresponding rate depends only on the intrinsic dimension d' and not on the ambient dimension d. We focus on the case of the Gaussian kernel.

Theorem 4.5. *Assume $d' \geq 2$ and $\operatorname{supp}(p^*_{d'})$ bounded. Consider the case of the Gaussian kernel. Then there are C_1 and C_2 such that*

$$\mathbb{E}\big[\mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{\mathrm{KDE}})\big] \leq C_1 h + \frac{C_2}{\sqrt{h^{d'-2}}} \frac{\log n}{\sqrt{n}}.$$

For $h \asymp (n/\log^2 n)^{-\frac{1}{d'}}$ we obtain $\mathbb{E}\big[\mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{\mathrm{KDE}})\big] = \mathcal{O}((n/\log^2 n)^{-\frac{1}{d'}})$.

The above result is in line with Theorem 4.3 for $\alpha = 0$ and similar results for kernel density estimators for distributions on manifolds (Berenfeld & Hoffmann, 2021; Divol, 2022). Note however that in these articles, the kernel is normalized with respect to the lower dimension, i.e. they consider $h^{-d'}K(\cdot/h)$ for bandwidth $h > 0$ or the evaluation metric is restricted to the subspace. The setting we are going to study in Chapter 5 enforces us to use a result for kernels which are normalized with respect to the ambient space dimension, i.e. $h^{-d}K(\cdot/h)$.

4.3. Proofs

4.3.1. Proofs of Section 4.1

Proof of Theorem 4.1. Using the Kantorovich duality (2.15), we get

$$\begin{aligned}
\mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{\mathrm{KDE}}) &= \sup_{f\in \mathrm{Lip}(1)} \mathbb{E}_{\substack{X\sim\mathbb{P}^* \\ Y\sim\mathbb{P}^{\mathrm{KDE}}}} [f(X) - f(Y)] \\
&= \sup_{\substack{f\in \mathrm{Lip}(1) \\ f(0)=0}} \mathbb{E}_{\substack{X\sim\mathbb{P}^* \\ Y\sim\mathbb{P}^{\mathrm{KDE}}}} [f(X) - f(Y)] \\
&= \sup_{\substack{f\in \mathrm{Lip}(1) \\ f(0)=0}} \int f(x)(p^*(x) - p_n(x)) \,\mathrm{d}x \\
&\leq \sup_{\substack{f\in \mathrm{Lip}(1) \\ f(0)=0}} \sup_{x\in A_0} |f(x)| \int |p^*(x) - p_n(x)| \,\mathrm{d}x,
\end{aligned}$$

where $A_0 = \{x \in \mathbb{R}^d \colon p^*(x) \neq 0 \text{ or } p_n(x) \neq 0\}$. As p^* and K (and therefore p_1) have bounded support, there exists some $K(d) > 0$ such that $\sup_{x\in A_0} |f(x)| \leq K(d)$. By Kohler (2015, Satz

2.4) there exist constants $c_1, c_2 > 0$ such that

$$\mathbb{E} \int_{\mathbb{R}^d} |p_n(x) - p^*(x)| \, dx \leq \frac{c_1}{\sqrt{n \cdot h^d}} + c_2 \cdot h^\alpha. \tag{4.2}$$

□

Proof of Theorem 4.2. Set $\mu_n := \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$. We denote $K_h := \frac{1}{h^d} K(\frac{\cdot}{h})$. By the triangle inequality of the supremum we get

$$\begin{aligned}
\mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{\mathrm{KDE}}) &= \sup_{f \in \mathrm{Lip}(1)} \int f(x)(p^*(z) - (K_h * \mu_n)(z)) \, \mathrm{d}z \\
&\leq \sup_{f \in \mathrm{Lip}(1)} \int f(z)(p^*(z) - (K_h * p^*)(z)) \, \mathrm{d}z \\
&\quad + \sup_{f \in \mathrm{Lip}(1)} \int f(z)((K_h * p^*)(z) - (K_h * \mu_n)(z)) \, \mathrm{d}z.
\end{aligned}$$

The first term is the bias of the kernel density estimator and the second term refers to the stochastic error.

First term: Using Fubini's theorem and $\int K_h(y) \, \mathrm{d}y = 1$, we rewrite the first term as

$$\begin{aligned}
&\sup_{f \in \mathrm{Lip}(1)} \int f(z)(p^*(z) - (K_h * p^*)(z)) \, \mathrm{d}z \\
&\qquad = \sup_{f \in \mathrm{Lip}(1)} \int f(z) \Big(\int K_h(y) p^*(z) \, \mathrm{d}y - \int K_h(y) p^*(z - y) \, \mathrm{d}y \Big) \, \mathrm{d}z \\
&\qquad = \sup_{f \in \mathrm{Lip}(1)} \int K_h(y) \Big(\int f(z) p^*(z) \, \mathrm{d}z - \int f(z) p^*(z - y) \, \mathrm{d}z \Big) \, \mathrm{d}y \\
&\qquad = \sup_{f \in \mathrm{Lip}(1)} \int K_h(y) \big((f * p^*(-\cdot))(0) - (f * p^*(-\cdot))(y) \big) \, \mathrm{d}y.
\end{aligned}$$

Then, since $|p^*|_{B^\alpha_{1,\infty}} \leq M$, we have that

$$\begin{aligned}
|\nabla(f * p^*(-\cdot))(x) - \nabla(f * p^*(-\cdot))(y)| &= \Big| \int \nabla f(z) p^*(x - z) \, \mathrm{d}z - \int \nabla f(z) p^*(y - z) \, \mathrm{d}z \Big| \\
&\leq \sqrt{d} M |x - y|^\alpha,
\end{aligned}$$

where we use that $|\nabla f| \leq \sqrt{d}$ for $f \in \mathrm{Lip}(1)$. Therefore by the generalized mean value theorem, for every y there exists a ξ_y and a $C_1' > 0$ such that

$$\begin{aligned}
&\sup_{f \in \mathrm{Lip}(1)} \int K_h(y) \Big(\int f(z) p^*(z) \, \mathrm{d}z - \int f(z) p^*(z - y) \, \mathrm{d}z \Big) \, \mathrm{d}y \\
&\qquad = \sup_{f \in \mathrm{Lip}(1)} \int K_h(y) (\nabla(f * p^*(-\cdot))(\xi_y))^\top y \, \mathrm{d}y \\
&\qquad = \sup_{f \in \mathrm{Lip}(1)} \int K_h(y) (\nabla(f * p^*(-\cdot))(\xi_y) - \nabla(f * p^*(-\cdot))(0))^\top y \, \mathrm{d}y \\
&\qquad\quad + \int K_h(y) (\nabla(f * p^*(-\cdot))(0))^\top y \, \mathrm{d}y
\end{aligned}$$

$$\begin{aligned}
&\leq \sup_{f\in \mathrm{Lip}(1)} \int |K_h(y)||\nabla(f * p^*(-\cdot))(\xi_y) - \nabla(f * p^*(-\cdot))(0)||y| \,\mathrm{d}y \\
&\quad + \int K_h(y)(f * p^*(-\cdot))(0))^\top y \,\mathrm{d}y \\
&\leq C_1' \int |K_h(y)||y|^{1+\alpha} \,\mathrm{d}y + \int K_h(y)(\nabla(f * p^*(-\cdot))(0))^\top y \,\mathrm{d}y.
\end{aligned}$$

For the first term we have that

$$\int K_h(y)|y|^{1+\alpha} \,\mathrm{d}y = \frac{1}{h^d} \int K(\frac{y}{h})|y|^{1+\alpha} \,\mathrm{d}y = h^{1+\alpha} \int K(u)|u|^{1+\alpha} \,\mathrm{d}u.$$

For the last term we use the assumption $\int K_h(y)y \,\mathrm{d}y = 0$. Hence we get

$$\sup_{f\in \mathrm{Lip}(1)} \int f(z)(p^*(z) - (K_h * p^*)(z)) \,\mathrm{d}z \leq h^{1+\alpha} C_1' \int K(u)|u|^{1+\alpha} \,\mathrm{d}u.$$

By assumption the last term is finite. Setting $C_1 \coloneqq C_1' \int K(u)|u|^{1+\alpha} \,\mathrm{d}u$ yields the first term in the final bound.

Second term: To bound the expectation of (4.4) we first use that by Villani (2008, Theorem 6.15)

$$\sup_{f\in \mathrm{Lip}(1)} \int f(z)((K_h * p^*)(z) - (K_h * \mu_n)(z)) \,\mathrm{d}z \leq \int |z||(K_h * p^*)(z) - (K_h * \mu_n)(z)| \,\mathrm{d}z.$$

Then for $\rho > \frac{d}{2}$

$$\begin{aligned}
&\mathbb{E}[\int |z||(K_h * p^*)(z) - (K_h * \mu_n)(z)| \,\mathrm{d}z] \\
&\quad = \int (1+|z|)^{-\rho}|z|(1+|z|)^{\rho}\mathbb{E}\Big[\Big|\frac{1}{n}\sum_{i=1}^n K_h(z-X_i) - \mathbb{E}[K_h(z-X_i)]\Big|\Big] \,\mathrm{d}z \\
&\quad \leq \frac{1}{n}\sqrt{\int (1+|z|)^{-2\rho} \,\mathrm{d}z} \sqrt{\int |z|^2(1+|z|)^{2\rho}\mathbb{E}\Big[\Big(\sum_{i=1}^n K_h(z-X_i) - \mathbb{E}[K_h(z-X_i)]\Big)^2\Big] \,\mathrm{d}z},
\end{aligned} \tag{4.3}$$

where the equality holds by Fubini's theorem and the inequality holds by the Jensen inequality together with the Cauchy-Schwarz inequality.

Next we bound the expectation. Since $\mathbb{E}[K_h(z-X_i) - \mathbb{E}[K_h(z-X_i)]] = 0$ and $X_1, \ldots, X_n$ are i.i.d. we have that

$$\begin{aligned}
\mathbb{E}\Big[\Big(\sum_{i=1}^n K_h(z-X_i) - \mathbb{E}[K_h(z-X_i)]\Big)^2\Big] &= n \operatorname{Var}\big(K_h(z-X_1)\big) \\
&\leq n\mathbb{E}[(K_h(z-X_1))^2].
\end{aligned}$$

Therefore

$$\int |z|^2(1+|z|)^{2\rho}\mathbb{E}\Big[\Big(\sum_{i=1}^n K_h(z-X_i) - \mathbb{E}[K_h(z-X_i)]\Big)^2\Big] \,\mathrm{d}z$$

$$
\begin{aligned}
&= n \int \mathbb{E}\Big[|X_1 + hy|^2 \big(1 + |X_1 + hy|^{2\rho}\big)^2 K^2(y)\Big] \,\mathrm{d}y \\
&\lesssim \frac{n}{h^d} \int \mathbb{E}\big[|X_1 + hy|^2 + |X_1 + hy|^{2\rho+2}\big] K^2(y) \,\mathrm{d}y \\
&\lesssim \frac{n}{h^d} \int \big(\mathbb{E}\big[|X_1|^{2\rho+2}\big] + h^2|y|^2 + h^{2\rho+2}|y|^{2\rho+2}\big) K^2(y) \,\mathrm{d}y \\
&\lesssim \frac{n}{h^d}(1 + h^2 + h^{2\rho+2}) \int \max(|y|, |y|^{2\rho+2}) K^2(y) \,\mathrm{d}y.
\end{aligned}
$$

Using that $h \leq 1$ we can bound (4.3) for $\rho > \frac{d}{2}$ further by

$$
\mathbb{E}\Big[\int |z| |(K_h * p^*)(z) - (K_h * \mu_n)(z)| \,\mathrm{d}z\Big] \lesssim \frac{1}{\sqrt{nh^d}} \int \max(|y|, |y|^{2\rho+2}) K^2(y) \,\mathrm{d}y.
$$

For $p = \frac{d}{2} + 1$ the last integral is finite and we thus obtain

$$
\mathbb{E}\Big[\int |z| |(K_h * p^*)(z) - (K_h * \mu_n)(z)| \,\mathrm{d}z\Big] \leq \frac{C_2}{\sqrt{nh^d}}.
$$

In the end, we get for (4.4) that

$$
\mathbb{E}\big[\mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{\mathrm{KDE}})\big] \leq C_1 h^{1+\alpha} + \frac{C_2 \max(1, h, h^{p+1})}{\sqrt{nh^d}}. \qquad \square
$$

Proof of Theorem 4.3. As in Theorem 4.2

$$
\begin{aligned}
\mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{\mathrm{KDE}}) &= \sup_{f \in \mathrm{Lip}(1)} \int f(x)(p^*(z) - (K_h * \mu_n)(z)) \,\mathrm{d}z \\
&\leq \sup_{f \in \mathrm{Lip}(1)} \int f(z)(p^*(z) - (K_h * p^*)(z)) \,\mathrm{d}z \\
&\quad + \sup_{f \in \mathrm{Lip}(1)} \int f(z)((K_h * p^*)(z) - (K_h * \mu_n)(z)) \,\mathrm{d}z. \qquad (4.4)
\end{aligned}
$$

For the first term, we obtain as in Theorem 4.2

$$
\sup_{f \in \mathrm{Lip}(1)} \int f(z)(p^*(z) - (K_h * p^*)(z)) \,\mathrm{d}z \leq C_1 h^{1+\alpha}.
$$

Second term: We rewrite the second term as

$$
\begin{aligned}
&\sup_{f \in \mathrm{Lip}(1)} \int f(z)((K_h * p^*)(z) - (K_h * \mu_n)(z)) \,\mathrm{d}z \\
&\qquad = \sup_{f \in \mathrm{Lip}(1)} \int \int f(z) K_h(z - x)(p^*(x) - \mu_n(x)) \,\mathrm{d}x \,\mathrm{d}z \\
&\qquad = \sup_{f \in \mathrm{Lip}(1)} \int (K_h(-\cdot) * f)(x) p^*(x) \,\mathrm{d}x - \int (K_h(-\cdot) * f)(x) \mu_n(x) \,\mathrm{d}x \\
&\qquad = \sup_{\substack{g = K_h(-\cdot) * f \\ f \in \mathrm{Lip}(1)}} \int g(x)(p^*(x) - \mu_n(x)) \,\mathrm{d}x.
\end{aligned}
$$

Using Youngs inequality and the properties of convolution, we obtain for $k \in \mathbb{N}_0^d$

$$\|D^k(K_h * f)\|_\infty \leq \|D^1 f\|_\infty \|D^{k-1} K_{h(-\cdot)}\|_1 \leq 1 \cdot \|D^{k-1} K_h\|_1.$$

By assumption $\|D^{k-1}K\|_1 \leq C$ and hence $\|D^{k-1}K_h\|_1 \leq Ch^{-k+1}$. Now we can use Schreuder (2020, Theorem 4). For even d, set $k = \frac{d}{2}$. Then there exists a constant $C_2' > 0$ such that

$$\sup_{\substack{g=K_h*f \\ f\in \mathrm{Lip}(1)}} \int g(x)(p^*(x) - \mu_n(x)) \, \mathrm{d}x \leq \frac{C_2'}{h^{\frac{d}{2}-1}} n^{-1/2} \log(n).$$

For odd $d > 1$, we know that

$$\|D^{\frac{d-1}{2}}(f * K_h)\|_\infty \leq \|D^1 f\|_\infty \|D^{\frac{d-3}{2}} K_h\|_1 \leq C_2' h^{-\frac{d-3}{2}},$$

and

$$\frac{D^{\frac{d-1}{2}}(f * K_h)(x) - D^{\frac{d-1}{2}}(f * K_h)(y)}{|x-y|^{1/2}} \leq 2C_2' h^{-\frac{d-3}{2}} \leq 2C_2' h^{-\frac{d}{2}+1}.$$

Then we can use Schreuder (2020, Theorem 4) with $\alpha = \frac{d}{2}$ again. Therefore, there is a $C_2'' > 0$ such that

$$\sup_{\substack{g=K_h*f \\ f\in \mathrm{Lip}(1)}} \int g(x)(p^*(x) - \mu_n(x)) \, \mathrm{d}x \leq \frac{C_2''}{h^{\frac{d}{2}-1}} n^{-1/2} \log(n).$$

Combining both terms we conclude that there are constants C_1 and C_2 with

$$\mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{\mathrm{KDE}}) \leq C_1 h^{1+\alpha} + \frac{C_2}{h^{\frac{d}{2}-1}} n^{-1/2} \log(n). \qquad \square$$

4.3.2. Proof of Section 4.2

Proof of Theorem 4.5. Recall the definitions $\mu_n \coloneqq \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ and $K_h \coloneqq \frac{1}{h^d} K(\frac{\cdot}{h})$ from Theorem 4.3. We start again by decomposing

$$\mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{\mathrm{KDE}}) \leq \mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^* * K_h) + \mathsf{W}_1(\mathbb{P}^* * K_h, \mathbb{P}^{\mathrm{KDE}}),$$

where we identified the density K_h with the corresponding probability measure.
For the first term, we can bound

$$\mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^* * K_h) \leq \mathbb{E}_{\substack{X\sim\mathbb{P}^* \\ Z_h\sim K_h}} [|X - (X + Z_h)|] = \mathbb{E}_{Z_h\sim K_h}[|Z_h|] = h\mathbb{E}_{Z\sim K}[|Z|] \lesssim h.$$

For the second term we use the orthogonality of A to obtain

$$\begin{aligned}
&\mathsf{W}_1(\mathbb{P}^* * K_h, \mathbb{P}^{\mathrm{KDE}}) \\
&= \sup_{f\in\mathrm{Lip}(1)} \int \frac{f(z)}{(2\pi h^2)^{d/2}} \Big(\int \exp\Big(-\frac{\|z-x\|^2}{2h^2}\Big) \, \mathrm{d}\mathbb{P}^*(x) - \frac{1}{n}\sum_{i=1}^n \exp\Big(-\frac{\|z-X_i\|^2}{2h^2}\Big)\Big) \, \mathrm{d}z \\
&= \sup_{f\in\mathrm{Lip}(1)} \int \frac{f(z)}{(2\pi h^2)^{d/2}} \Big(\int \exp\Big(-\frac{\|z-AA^\top z\|^2 + \|AA^\top z - x\|^2}{2h^2}\Big) \, \mathrm{d}\mathbb{P}^*(x) \qquad (4.5)
\end{aligned}$$

$$- \frac{1}{n}\sum_{i=1}^{n} \exp\Big(- \frac{\|z - AA^\top z\|^2 + \|AA^\top z - X_i\|^2}{2h^2}\Big)\Big) \,\mathrm{d}z. \tag{4.6}$$

By assumption, there is for every $x \in V$ a $y \in \mathbb{R}^{d'}$ such that $x = Ay$. As the support of of $\mathbb{P}^*$ is subset of V, for every X_i there is a Y_i^* mapping into $\mathbb{R}^{d'}$ such that $X_i = AY_i^*$ and $Y_i^* \sim p^*_{d'}$. Hence we get

$$\begin{aligned}
&\int \exp\Big(- \frac{\|z - AA^\top z\|^2 + \|AA^\top z - x\|^2}{2h^2}\Big) \,\mathrm{d}\mathbb{P}^*(x) \\
&\qquad = \int \exp\Big(- \frac{\|z - AA^\top z\|^2 + \|AA^\top z - Ay\|^2}{2h^2}\Big) p^*_{d'}(y) \,\mathrm{d}y \\
&\qquad = \int \exp\Big(- \frac{\|z - AA^\top z\|^2 + \|A^\top z - y\|^2}{2h^2}\Big) p^*_{d'}(y) \,\mathrm{d}y,
\end{aligned}$$

where the last equality holds due to the orthonormal columns of A. For (4.6) we proceed analogously. Now we can decompose every $z \in \mathbb{R}^d$ uniquely into

$$z = Au + Bv, \quad \text{for some} \quad u \in \mathbb{R}^{d'}, v \in \mathbb{R}^{d-d'},$$

where $B \in \mathbb{R}^{d\times(d-d')}$ is a matrix that maps $\mathbb{R}^{d-d'}$ into $V^\perp$. As $A^\top B = 0$ and therefore

$$A^\top z = A^\top(Au + Bv) = A^\top Au = u \quad \text{and similar} \quad (I_d - AA^\top)z = v.$$

We get for (4.5) and (4.6)

$$\begin{aligned}
&\sup_{f\in\mathrm{Lip}(1)} \int \frac{f(z)}{(2\pi h^2)^{d/2}} \Big(\int \exp\Big(- \frac{\|z - AA^\top z\|^2 + \|A^\top z - y\|^2}{2h^2}\Big) p^*_{d'}(y) \,\mathrm{d}y \\
&\qquad\qquad - \frac{1}{n}\sum_{i=1}^{n} \exp\Big(- \frac{\|z - AA^\top z\|^2 + \|A^\top z - Y_i^*\|^2}{2h^2}\Big)\Big) \,\mathrm{d}z \\
= &\sup_{f\in\mathrm{Lip}(1)} \int\int \frac{f(Au + Bv)}{(2\pi h^2)^{(d-d')/2}} \exp\Big(- \frac{\|v\|^2}{2h^2}\Big) \,\mathrm{d}v \Big(\int \frac{1}{(2\pi h^2)^{d'/2}} \exp\Big(- \frac{\|u - y\|^2}{2h^2}\Big) p^*_{d'}(y) \,\mathrm{d}y \\
&\qquad\qquad - \frac{1}{n}\sum_{i=1}^{n} \frac{1}{(2\pi h^2)^{d'/2}} \exp\Big(- \frac{\|u - Y_i^*\|^2}{2h^2}\Big)\Big) \,\mathrm{d}u.
\end{aligned}$$

For fixed $v \in \mathbb{R}^{d-d'}$ and $u, w \in \mathbb{R}^{d'}$, we have that

$$\begin{aligned}
&\Big| \int \frac{f(Au + Bv)}{(2\pi h^2)^{(d-d')/2}} \exp\Big(- \frac{\|v\|^2}{2h^2}\Big) \,\mathrm{d}v - \int \frac{f(Aw + Bv)}{(2\pi h^2)^{(d-d')/2}} \exp\Big(- \frac{\|v\|^2}{2h^2}\Big) \,\mathrm{d}v \Big| \\
&\qquad \le \int \frac{|f(Au + Bv) - f(Aw + Bv)|}{(2\pi h^2)^{(d-d')/2}} \exp\Big(- \frac{\|v\|^2}{2h^2}\Big) \,\mathrm{d}v \\
&\qquad \le \int \frac{|A(u - w)|}{(2\pi h^2)^{(d-d')/2}} \exp\Big(- \frac{\|v\|^2}{2h^2}\Big) \,\mathrm{d}v \\
&\qquad = |u - w|,
\end{aligned}$$

where we used the Lipschitz continuity of f in the second inequality, the orthogonality of the

columns of A and the density of $\mathcal{N}(0, I_{d-d'})$ in the last equality. We conclude that the integral is Lipschitz in u with Lipschitz constant 1. Hence

$$\begin{aligned}\sup_{f\in\mathrm{Lip}(1)} \int\int \frac{f(Au+Bv)}{(2\pi h^2)^{(d-d')/2}} \exp\Big(-\frac{\|v\|^2}{2h^2}\Big)\,\mathrm{d}v\Big(\int \frac{1}{(2\pi h^2)^{d'/2}} \exp\Big(-\frac{\|u-y\|^2}{2h^2}\Big)p^*_{d'}(y)\,\mathrm{d}y \\ - \frac{1}{n}\sum_{i=1}^n \frac{1}{(2\pi h^2)^{d'/2}} \exp\Big(-\frac{\|u-Y_i^*\|^2}{2h^2}\Big)\Big)\,\mathrm{d}u \\ \le \sup_{f\in\mathrm{Lip}(1)} \int f(u)\,\mathrm{d}v\Big(\int \frac{1}{(2\pi h^2)^{d'/2}} \exp\Big(-\frac{\|u-y\|^2}{2h^2}\Big)p^*_{d'}(y)\,\mathrm{d}y \\ - \frac{1}{n}\sum_{i=1}^n \frac{1}{(2\pi h^2)^{d'/2}} \exp\Big(-\frac{\|u-Y_i^*\|^2}{2h^2}\Big)\Big)\,\mathrm{d}u.\end{aligned}$$

Now we can proceed along the proof of the second term of Theorem 4.3 to obtain the desired bound. □

Chapter 5

Generative Flow Matching

The second generative model we are going to investigate is Flow Matching as introduced by Lipman et al. (2023). Again, to ease readability, we quickly recall the definitions from the introduction and introduce the setting that will be studied in this chapter.
Assume we observe an i.i.d. sample $X_1^*, \dots, X_n^*$ from an unknown distribution $\mathbb{P}^*$ on $\mathbb{R}^d$. Further assume that $\mathbb{P}^*$ has a density p^* with respect to the Lebesgue measure and finite first moment. For a time dependent vector field $v\colon [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$ we consider the flow $\psi\colon [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$ given as the solution to the ODE

$$\frac{d}{dt}\psi_t(x) = v_t(\psi_t(x)), \quad \psi_0(x) = x. \tag{5.1}$$

For a fixed latent distribution with Lebesgue density p_0, the vector field v generates a probability density path $p\colon [0,1] \times \mathbb{R}^d \to \mathbb{R}_{>0}$ with $\int p_t(x)\mathrm{d}x = 1$ for all t via the push-forward distributions

$$p_t = [\psi_t]_{\#} p_0, \quad \text{i.e. } \psi_t(Z) \sim p_t \text{ for } Z \sim \mathbb{U},$$

where $\mathbb{U}$ is a chosen latent distribution in $\mathbb{R}^d$ that admits a density with respect to the Lebesgue measure and has finite first moment. The ODE (5.1) corresponds to the Lagrangian description (in terms of particle trajectories) of the conservation of mass formula (Villani, 2008, p. 14). The change of variables formula links it to the Eulerian description: A necessary and sufficient condition for v_t to generate p_t is

$$\frac{d}{dt}p_t + \operatorname{div}(p_t v_t) = 0, \tag{5.2}$$

see Villani (2008, p. 14).
Approximating a given vector field v that generates a certain density density path p using a parameterized function $\tilde{v}$ leads to the Flow Matching objective

$$\mathbb{E}_{\substack{t\sim\mathcal{U}[0,1] \\ X_t\sim p_t}}\left[|v_t(X_t) - \tilde{v}_t(X_t)|^2\right] \tag{5.3}$$

as in Lipman et al. (2023). To interpolate between $\mathbb{U}$ and an approximation of $\mathbb{P}^*$, they have considered a probability path of the form

$$p_t(x, p^*) = \int p_t(x|y)p^*(y)\ \mathrm{d}y, \tag{5.4}$$

where $p_t(\cdot|y)\colon \mathbb{R}^d \to \mathbb{R}$ is a *conditional probability path* generated by some vector field $v_t(\cdot|y)\colon \mathbb{R}^d \to \mathbb{R}^d$ for $y \in \mathbb{R}^d$. As we shall see in Lemma 5.4, one example of such a pair in case $\mathbb{U} = \mathcal{N}(0, I_d)$ is

$$v_t(x|y) \coloneqq \frac{y - (1-\sigma_{\min})x}{1-(1-\sigma_{\min})t} \quad \text{and} \quad p_t(x|y) \propto \exp\Big(-\frac{|x-ty|^2}{2(1-(1-\sigma_{\min})t)^2}\Big).$$

The vector field generating (5.4) is then given by

$$v_t(x, p^*) = \int v_t(x|y)\frac{p_t(x|y)p^*(y)}{p_t(x)}dy. \tag{5.5}$$

In this setting Lipman et al. (2023) show that minimizing the Flow Matching objective (5.3) with respect to the parameters of $\tilde{v}$ is equivalent minimizing the conditional Flow Matching objective

$$\Psi(\tilde{v}) \coloneqq \mathbb{E}_{\substack{t\sim\mathcal{U}[0,1] \\ Y\sim p^* \\ X_t\sim p_t(\cdot|Y)}} \big[\, |\tilde{v}_t(X_t) - v_t\left(X_t|Y\right)|^2 \,\big]. \tag{5.6}$$

Note that we changed the variable naming slightly compared to Chapter 1 to ease notation in this chapter and to align with the literature. In the following, we are going to omit the dependency of p_t and v_t on p^*.

Flow Matching algorithms have been successfully used in many different applications that benefit from efficient sampling, such as text-to-speech (Guo et al., 2024) and text-to-image (Yang et al., 2025; Esser et al., 2024) settings, the production of novel molecular or protein structures (Dunn & Koes, 2024; Bose et al., 2024) or the construction of surrogate models in high energy physics (Bieringer et al., 2024). It has also been adjusted theoretically to different settings. Atanackovic et al. (2025) adapted Flow Matching to the case of interacting particles, Gat et al. (2024) explored the discrete setting. Chen & Lipman (2024) generalized the Euclidean setting to the Riemannian setting, allowing for more general geometries. Kerrigan et al. (2024) extended Flow Matching to function spaces. However, the statistical properties have only recently been studied by Fukumizu et al. (2025) and Gao et al. (2024b) in the Wasserstein-2 distance. They do not use the exact setting of Lipman et al. (2023), but rather introduce stopping times that depend on the number of samples, enabling the transfer of methods known from the statistical analysis of diffusion models.

Related work Lipman et al. (2023) use a fixed latent distribution. Similar approaches for flows between two possibly unknown distributions $\mathbb{P}$ and $\mathbb{Q}$ are studied by Tong et al. (2024), Liu et al. (2023) and Albergo & Vanden-Eijnden (2023). Tong et al. (2024) also generalize the mentioned methods. Gao et al. (2024b) prove a suboptimal rate of convergence in the Wasserstein-2 distance. Benton et al. (2024) analyzed Flow Matching excluding the approximation error by imposing assumptions on the covariance that lead to global Lipschitz bounds on the vector field. They also focus on different constructions than Lipman et al. (2023). Gong et al. (2025) study the properties of ReLU networks to approximate a vector field corresponding to higher order trajectories.

As Lipman et al. (2023) point out, Flow Matching is closely related to diffusion models. We will provide a short overview of this connection in Section 5.1.2. For an overview of generative diffusion models, see Cao et al. (2024). Indeed, even in cases that are not constructed to be consistent with diffusion models, the approximation of a score function has similar properties to the approximation of the Flow Matching vector field. Fukumizu et al. (2025) build up on this connection. However, their proof has a critical flaw concerning the bounds of the integral in the exponent on page 14[1]. Furthermore, they rely on Oko et al. (2023, Theorem C.4), which is also not completely correct, as pointed out by Yakovlev & Puchkin (2025). This second flaw appears to be fixable (Stéphanovitch et al., 2025). The statistical properties of score matching are an area of ongoing research, see for example Chen et al. (2023a), Chen et al. (2023b), Chen et al. (2023c), Oko et al. (2023), Tang & Yang (2024), Azangulov et al. (2024), Zhang et al. (2024), Yakovlev & Puchkin (2025). Marzouk et al. (2024) study the statistical properties of continuous normalizing flows (CNFs) trained by likelihood maximization.

Own contribution In this chapter, we first demonstrate that Flow Matching in the setting of Lipman et al. (2023) is closely related to the classical kernel density estimation (KDE). This connection allows us to analyze Flow Matching from a new perspective. First, we show that the motivation of Flow Matching also holds for its empirical counterparts. For sufficiently large network classes, the resulting generative algorithm coincides exactly with a kernel density estimator, where the kernel is given by the density of the latent distribution.
Then we show convergence rates that build up on the rates obtained in Section 4.1. Unlike Fukumizu et al. (2025), we do exploit on the similarities to diffusion models. Separating the error of the kernel density estimator allows us to use empirical risk minimization without the need for Bernstein-type bounds. This avoids the problems pointed out by Yakovlev & Puchkin (2025) in the analysis of diffusion models. Overall, the analysis of Flow Matching is more delicate, since Girsanov's theorem does not apply and thus the strategy of Chen et al. (2023c) cannot be used. Therefore, our bound depends exponentially on the Lipschitz constant of the vector field. This is one of the reasons for the difficulties in the proof of Fukumizu et al. (2025).
This analysis allows us to show convergence rates in Wasserstein-1 distance in the case where finite neural networks are used for the vector field. While these rates are minimax optimal up to a logarithmic constant, they suffer from the curse of dimensionality. In case the support of $\mathbb{P}^*$ is concentrated on a linear subspace, we improve our results in a way that the convergence rate depends only on the intrinsic dimension. This provides a first justification for the excellent performance of Flow Matching for high-dimensional data sets.
In a completely different second ansatz, we focus on smaller networks, that correspond more to real world scenarios. This approach requires a detailed study of the Lipschitz constant of the underlying population vector field (5.5). We provide upper and lower bounds on this Lipschitz constant and develop conditions under which the unknown distribution admits a Lipschitz controlled vector field. We demonstrate that our assumptions can be met even without assuming

[1]The separation of the outer integral does not imply the same separation of the integral in the exponent. The correct factor in the integral is $e^{2\int_{T_0}^{t_j} L_u du}$, where L_u is the Lipschitz constant of the approximated vector field, which depends on T_0 and is thus not bounded by a universal constant.

log-concavety. Then, using the Bernstein-type inequality presented in Theorem 2.11, we obtain a rate of convergence that allows for significantly smaller networks than the result obtained before. As a consequence, the rate obtained is not minimax optimal, but still improves existing rates for similar settings.

5.1. Overview of related methods

In the following, we give a quick overview of generative models related to Flow Matching. First, we present predecessors of Flow Matching. Afterwards, we draw a connection from Flow Matching to diffusions, which is restricted to properties that will be revisited later and is therefore by far not exhaustive.

5.1.1. From Normalizing Flows to Flow Matching

A natural starting point for a generative model is the change of variables theorem. If $Z \sim p_Z$, then $\psi(Z)$ admits the density

$$p_{\psi(Z)}(x) = p_Z(\psi^{-1}(x))|\det D_x\psi^{-1}(x)|. \tag{5.7}$$

Applying the logarithm on both sides leads to

$$\log(p_{\psi(Z)}(x)) = \log(p_Z(\psi^{-1}(x))) + \log(|\det D_x\psi^{-1}(x)|).$$

The use of concatenations of diffeomorphisms for the function ψ gave rise to normalizing flows. The first approaches were defined by Tabak & Vanden-Eijnden (2010) and Tabak & Turner (2013), but popularized by Rezende & Mohamed (2015) and Dinh et al. (2015). There have been numerous approaches to construct the function ψ in this setting, for an overview see Kobyzev et al. (2020). Although the models were constructed to limit computational costs of the calculation of the determinant of the Jacobian, this remained the bottleneck of normalizing flows.
Expanding the discrete setting of concatenations of diffeomorphism to a continuous setting gave rise to CNFs (Chen et al., 2018). Instead of (5.7), they showed and used the following continuous adaptation named *Instantaneous Change of Variables*.

Theorem 5.1. *(Chen et al., 2018, Theorem 1) Let $v : \mathbb{R}^d \times [0,T] \to \mathbb{R}^d$ be uniformly Lipschitz in the first component and continuous in the second component. Further, let $\psi_t : \mathbb{R}^d \to \mathbb{R}^d$ be the flow resulting from v, i.e.*

$$\frac{\partial \psi_t(x)}{\partial t} = v\left(\psi_t(x), t\right), \quad \psi_0(x) = x$$

Let Z be a random variable such that for each $t \in [0,T]$, the distribution $\mathbb{P}^{\psi_t(Z)}$ is absolutely continuous w.r.t. the Lebesgue measure with density p_t. Then for almost every $t \in [0,T]$ and for all x with $p_t\left(\psi_t(x)\right) > 0$,

$$\frac{d}{dt}\log p_t\left(\psi_t(x)\right) = -\operatorname{tr}\left(D_x v\left(\psi_t(x), t\right)\right).$$

Theorem 5.1 replaces the calculation of the determinant with a trace operation, which is computationally cheaper and linear. Chen et al. (2018) then use a maximum likelihood method to optimize over a parameterized class of functions v. However, the training requires simulations of the ODE, which results in high computational cost, see Grathwohl et al. (2019).
The Flow Matching approach offers an alternative objective circumventing the need for simulations during training. Next to Lipman et al. (2023), Albergo & Vanden-Eijnden (2023) and Liu et al. (2023) developed similiar models with the same advantage. Additionally, there are several approaches that extend these models. For an overview, we refer to Tong et al. (2024).

5.1.2. Diffusion models

The regression objective of Flow Matching is similar to the score matching objective (Hyvärinen, 2005; Vincent, 2011), which is used to train diffusion models based on SDEs, introduced by Song et al. (2021). In the following, we want to make this comparison more precise. To this end, we start by recalling the standard definitions and constructions of Song et al. (2021). For exhaustive definitions and assumptions necessary for existence and uniqueness of solutions to SDEs and reverse processes, we refer to Øksendal (2003). Then we illustrate the connection of diffusions based on Ornstein-Uhlenbeck processes to Flow Matching.
Song et al. (2021) introduced SDE-based diffusions via the very general SDE

$$\mathrm{d}X_t = f(X_t, t)\,\mathrm{d}t + g(t)\,\mathrm{d}B_t, \tag{5.8}$$

where B is a d-dimensional Brownian motion, $f\colon\ \mathbb{R}^d \times [0,T] \to \mathbb{R}^d$ is the drift coefficient and $g\colon [0,T] \to \mathbb{R}$ is the diffusion coefficient for $T > 0$. For conditions that guarantee existence and uniqueness of solutions of (5.8), we refer to Øksendal (2003, Theorem 5.2.1). The diffusion is then initialized using the unknown distribution, e.g. $X_0 \sim \mathbb{P}^*$. The density of X_t is denoted by p_t.
Under certain assumptions on the drift, the diffusion and p_t, see Anderson (1982), the reverse process of (5.8), i.e. the process satisfying $\overleftarrow{X}_t \sim X_{T-t}$, is given by the SDE

$$\mathrm{d}\overleftarrow{X}_t = [f(\overleftarrow{X}_t, T-t) - g(T-t)^2 \nabla_x \log p_{T-t}(\overleftarrow{X}_t)]\,\mathrm{d}t + g(T-t)\,\mathrm{d}\overleftarrow{B}_t,$$

where $\overleftarrow{B}$ is another d-dimensional Brownian motion and the diffusion is initialized at $\overleftarrow{X}_0 \sim \mathbb{P}^{X_T}$. The function $\nabla_x \log p_t$ is then approximated for all $t \in [0,T]$ by some parameterized function s_θ. The denoising score matching objective is obtained by using the equality

$$\mathbb{E}_{X_t}[|\nabla_x \log p_t(X_t) - s_\theta(X_t,t)|^2] = \mathbb{E}_{X_0}[\mathbb{E}_{X_t|X_0}[|\nabla_x \log p_t(X_t|X_0) - s_\theta(X_t,t)|^2]] + C,$$

where C is a constant independent of s_θ, which is due to Vincent (2011). In the above equation $p_t(\cdot|x_0)$ is the conditional density of X_t given $X_0 = x_0$. The idea of approximating $\nabla_x \log p_t$ dates back to Hyvärinen (2005) and is naturally connected to the Hyvärinen score presented in Section 2.3.
Song et al. (2021) already present two different instances, the variance exploding and the variance

preserving diffusion. For the variance exploding diffusion, they set $f \equiv 0$ and use a function $\sigma^2(t)$ with $\sigma^2(0) = 0$ to build the diffusion coefficient g. Thus, they recover the SDE

$$\mathrm{d}X_t = \sqrt{\frac{\mathrm{d}\sigma^2(t)}{\mathrm{d}t}}\,\mathrm{d}B_t.$$

In this setting

$$X_t|X_0 \sim \mathcal{N}(X_0, \sigma^2(t)I_d).$$

This explains the name variance exploding: if σ is chosen such that it grows in t, then $p_t(\cdot|X_0)$ adds noise around the fixed mean X_0. This model has been studied from a statistical perspective by Zhang et al. (2024) and Dou et al. (2024).
The variance preserving diffusion behaves differently. For a function $\beta\colon \mathbb{R} \to \mathbb{R}_{>0}$, consider the SDE

$$\mathrm{d}X_t = -\frac{1}{2}\beta(t)X_t\,\mathrm{d}t + \sqrt{\beta(t)}\,\mathrm{d}B_t.$$

This is a linear SDE with solution

$$X_t = \exp\Big(-\frac{1}{2}\int_0^t \beta(s)\,\mathrm{d}s\Big)\Big(X_0 + \int_0^t \exp\Big(\frac{1}{2}\int_0^s \beta(z)\,\mathrm{d}z\Big)\sqrt{\beta(s)}\,\mathrm{d}B_s\Big).$$

Define

$$\alpha_t \coloneqq \exp\Big(-\frac{1}{2}\int_0^t \beta(s)\,\mathrm{d}s\Big),$$

then the Ito isometry, see e.g. Øksendal (2003, Lemma 3.1.5) and standard integration leads to

$$X_t|X_0 \sim \mathcal{N}(\alpha_t X_0, (1-\alpha_t^2)I_d).$$

A special instance of the variance preserving diffusion corresponds to the choice $\beta(t) = 2$. Then

$$X_t|X_0 \sim \mathcal{N}(\exp(-t)X_0, (1-\exp(-2t))I_d). \tag{5.9}$$

and for $t \to \infty$ we recover the standard normal distribution. The resulting process is a special instance of an Ornstein-Uhlenbeck process. From (5.9), the name variance preserving is apparent: for growing t, the mean converges towards 0 and the variance converges towards 1. This model or slight adaptations of this model, such as the multiplication with the factor 2 or the multiplication of the diffusion coefficient with some factor $\sigma > 0$ have also been studied extensively in the literature, see for example Chen et al. (2023a), Chen et al. (2023c), Oko et al. (2023), Stéphanovitch et al. (2025), Arsenyan et al. (2025).

The asymptotic behavior of (5.9) for $t \to \infty$ is the same as the behavior of Flow Matching for $t \to 0$, if $\mathcal{N}(0, I_d)$ is chosen as the latent distribution which we will assume in this subsection. Thus, we consider (5.9) for our comparison. In contrast to diffusions, Flow Matching works in a finite time horizon. Additionally, the time runs in reverse direction, i.e. for $t = 0$ we have that $X_0 \sim \mathbb{P}^*$. To avoid confusion, we will use $X^* \sim \mathbb{P}^*$ without a time index for the remainder of

this subsection. Switching the time regime of the diffusion leads to

$$X_t|X^* \sim e^{-(1-t)}X^* + \sqrt{1-e^{-2(1-t)}}Z,$$

for $t \in (-\infty, 1)$, where $Z \sim \mathcal{N}(0, I_d)$ and $X^* \sim \mathbb{P}^*$ are independent. Set $\sigma_t := \sqrt{1-e^{-2(1-t)}}$. For $t < 1$ the density of the marginal is then given by

$$\begin{aligned}\tilde{p}_t(x) &= \int \frac{1}{(2\pi\sigma_t^2)^{d/2}} \exp\Big(-\frac{|x-y|^2}{2\sigma_t^2}\Big) e^{d(1-t)} p^*\Big(\frac{y}{e^{-(1-t)}}\Big)\,\mathrm{d}y \\ &= \int \frac{1}{(2\pi\sigma_t^2)^{d/2}} \exp\Big(-\frac{|x-e^{-(1-t)}z|^2}{2\sigma_t^2}\Big) p^*(z)\,\mathrm{d}z.\end{aligned}$$

Theoretical results of diffusion models need early stopping at a time $\underline{t} \in (0,1)$ of the backward process to prevent the score from blowing up. In Flow Matching as introduced in Chapter 1, the analogue of this is the choice of a small variance $\sigma_{\min}$. We note that for the mean shift, there is a slight difference: while early stopping leads to a small factor $e^{-(1-\underline{t})}$ for $\underline{t}$ close to 1, the mean shift in Flow Matching guarantees that $p_1(x|y)$ is the density of a Gaussian with mean y. Additionally, the backward process is started from $\mathcal{N}(0, I_d)$ at some time $-T$ ($T+1$ in the original time regime) for a large T. This additional setting does not occur in the Flow Matching setting, as the vector field is constructed such that it transfers mass from the latent distribution directly. Accounting for these differences, we can still compare the shift of variance across the different models. To do so, we squeeze the diffusion time interval $[-T, 1]$ into the Flow Matching time interval $[0, 1]$. Figure 5.1 illustrates the difference in variance functions for different choices of early stopping and different $T > 0$. Figure 5.2 illustrates the difference in the factors of the mean shift.

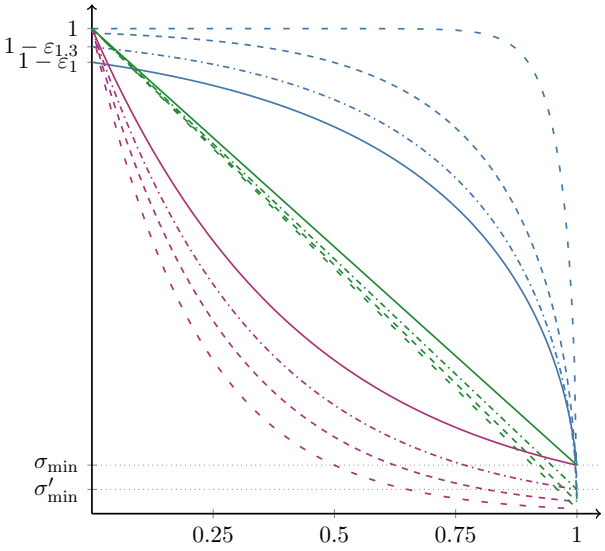


Figure 5.1.: Variance shifts of diffusion models (blue), the linear choice in Lipman et al. (2023) (green) and the choice in Section 5.5.2 (red) for different values of $\sigma_{\min}$/early stopping times: solid line 0.1, dashdotted 0.05, dashed 0.025, loosely dashed 0.01. ε_T is the difference resulting from running the diffusion only to $-T$ instead of $-\infty$. Only $\sigma_{\min} = 0.1$ and $\sigma'_{\min} = 0.05$ are labeled on the axis.

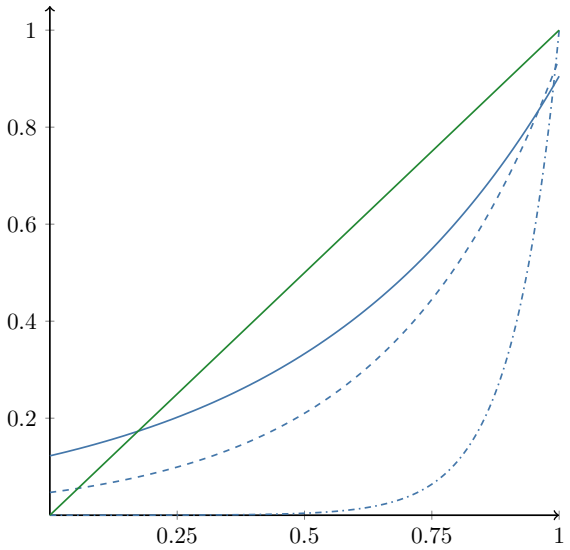


Figure 5.2.: Mean shift factors of diffusion models (blue) and the linear choice in Lipman et al. (2023) (green). We consider $T = 1$ (solid), $T = 2$ (dashed line), $T = 10$ (dashdotted line). Early stopping leads to a bias in case of diffusions. Note that we only consider the factors of the mean shift.

The upper time T is typically chosen as a monotonously increasing function of n, the stopping time as a monotonously decreasing function of n. In the squeezed reverse-time interval $[0, 1]$, this leads to a steeper and later decline of the variance. We are going to show that there are classes of unknown distributions that allow for a broad range of variance functions while preserving controlled Lipschitz continuity of the vector field from (5.5).

5.2. Connection to kernel density estimation

As p^* is unknown, in practice, the expectation in $Y \sim p^*$ in (5.6) is replaced by the empirical counterpart based on i.i.d. observations $X_1^*, \dots, X_n^*$ from the unknown distribution $\mathbb{P}^*$. This leads to the empirical counterparts of (5.4) and (5.5) given by

$$p_t^n(x) = \frac{1}{n} \sum_{i=1}^{n} p_t(x|X_i^*) \qquad \text{and} \qquad v_t^n(x) = \sum_{i=1}^{n} v_t(x|X_i^*) \frac{p_t(x|X_i^*)}{\sum_{j=1}^{n} p_t(x|X_j^*)}. \tag{5.10}$$

With this modification we recover the sufficient condition for v_t^n to generate p_t^n analogously to Lipman et al. (2023, Theorem 1).

Lemma 5.2. *If $v_t(\cdot|X_i^*)$ generates $p_t(\cdot|X_i^*)$ for all $i = 1, \dots, n$, then v_t^n generates p_t^n.*

The motivation for the conditional Flow Matching objective (5.6) is the equivalence to the unconditioned Flow Matching objective with respect to the optimizing arguments (Lipman et al., 2023, Theorem 2). Using the empirical counterparts, this still holds true.

Theorem 5.3. *Let $p_t(\cdot|y) \colon \mathbb{R}^d \to \mathbb{R}$ be a probability path generated by a vector field $v_t(\cdot|y) \colon \mathbb{R}^d \to \mathbb{R}^d$ for $y \in \mathbb{R}^d$. Using p_t^n and v_t^n from* (5.10)*, and a class of parameterized vector fields $\mathcal{M}$, which*

is constructed such that the minimal arguments exist, we have that

$$\begin{aligned}&\operatorname*{argmin}_{\hat{v}\in\mathcal{M}} \int_0^1 \mathbb{E}_{X_t\sim p_t^n}\left[|\hat{v}_t(X_t) - v_t^n(X_t)|^2\right] \mathrm{d}t \\ &\qquad = \operatorname*{argmin}_{\hat{v}\in\mathcal{M}} \int_0^1 \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\tilde{X}_t\sim p_t(\cdot|X_i^*)}\left[|\hat{v}_t(\tilde{X}_t) - v_t(\tilde{X}_t|X_i^*)|^2\right] \mathrm{d}t.\end{aligned}$$

Hence the *empirical conditional Flow Matching* objective

$$\tilde{\Psi}(\tilde{v}) \coloneqq \int_0^1 \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\tilde{X}_t\sim p_t(\cdot|X_i^*)}\left[|\tilde{v}_t(\tilde{X}_t) - v_t(\tilde{X}_t|X_i^*)|^2\right] \mathrm{d}t \tag{5.11}$$

is justified theoretically. Compared to Gao et al. (2024b), we use $v_t(\cdot|X_i^*)$ directly as proposed by Lipman et al. (2023) and stick to the entire time interval. Note that for simplicity we do not sample t.

From minimizing (5.11) in $\tilde{v}$ over a class $\mathcal{M}$, which is constructed such that the minimal argument exists and all corresponding ODEs admit a solution, we obtain an optimal argument $\hat{v}$. Solving the ODE (5.1) using $\hat{v}$, we obtain a flow $\hat{\psi}_t$, i.e. $\hat{\psi}$ is given by

$$\frac{d}{dt}\hat{\psi}_t(x) = \hat{v}_t(\hat{\psi}_t(x)), \quad \hat{\psi}_0(x) = x, \qquad \text{for} \quad \hat{v} \in \operatorname*{argmin}_{\tilde{v}\in\mathcal{M}} \tilde{\Psi}(\tilde{v}). \tag{5.12}$$

We use this flow to push forward the known, latent distribution $\mathbb{U}$ to time $t = 1$. In accordance with the goal of generative modeling, the distribution of this pushforward should mimic $\mathbb{P}^*$.

In order to apply Flow Matching, we have to construct a class of conditional probability paths. Let $Z \sim \mathbb{U}$ and let K denote the density of $\mathbb{U}$. Consider the *variance function* $\sigma\colon [0,1]\times\mathbb{R}^d \to \mathbb{R}_{>0}$ and the *mean shift* $\mu\colon [0,1]\times\mathbb{R}^d \to \mathbb{R}^d$. Set for $t \in [0,1]$ and a given $X_i^*, i \in \{1,..,n\}$,

$$\psi_t(Z|X_i^*) \coloneqq \sigma_t(X_i^*)Z + \mu_t(X_i^*). \tag{5.13}$$

The density of $\psi_t(Z|X_i^*)$ is by the transformation formula

$$p_t(x|X_i^*) = \frac{1}{\sigma_t^d(X_i^*)} K\Big(\frac{x - \mu_t(X_i^*)}{\sigma_t(X_i^*)}\Big).$$

We call $p_t(\cdot|X_i^*)$ the *conditional kernel probability path.* Setting

$$\frac{d}{dt}\psi_t(x|X_i^*) = v_t(\psi_t(x|X_i^*)|X_i^*), \tag{5.14}$$

we recover the same result like Lipman et al. (2023, Theorem 3).

Lemma 5.4. *Let $p_t(x|X_i^*)$ be a conditional kernel probability path, and $\psi_t(\cdot|X_i^*)$ its corresponding flow as in* (5.13)*. Then, the unique vector field that defines $\psi_t(\cdot|X_i^*)$ via* (5.14) *has the form*

$$v_t(x|X_i^*) = \frac{\frac{\partial\sigma_t}{\partial t}(X_i^*)}{\sigma_t(X_i^*)}\big(x - \mu_t(X_i^*)\big) + \frac{\partial\mu_t}{\partial t}(X_i^*).$$

Set $\sigma_{\min} > 0$. To flow from $p_0^n(x) = \frac{1}{n}\sum_{i=1}^n K(x) = K(x)$ to $p_1^n(x) = \frac{1}{n\sigma_{\min}^d}\sum_{i=1}^n K(\frac{x-X_i^*}{\sigma_{\min}})$, we can choose any differentiable functions σ_t and μ_t such that for every $x \in \mathcal{X}$

$$\mu_0(x) = 0, \quad \mu_1(x) = x \quad \text{and} \quad \sigma_0(x) = 1, \quad \sigma_1(x) = \sigma_{\min}.$$

At time $t = 1$ the distribution $\mathbb{P}^{\psi_1^n(Z)}$ then coincides with the kernel density estimator

$$p_1^n(x) = \frac{1}{n\sigma_{\min}^d}\sum_{i=1}^n K\Big(\frac{x - X_i^*}{\sigma_{\min}}\Big), \tag{5.15}$$

where the kernel is given by the latent distribution $\mathbb{U}$. Choosing $\mathbb{U} = \mathcal{N}_d(0,1)$, i.e., we consider the d-dimensional Gaussian kernel $K(x) = (2\pi)^{-d/2}\exp(-|x|^2/2)$, yields the proposed flow from Lipman et al. (2023, Section 4). Moreover, considering general kernels is in line with methods by Tong et al. (2024), Liu et al. (2023) and Albergo & Vanden-Eijnden (2023), which transform an unknown distribution to another. Interestingly, a similiar connection in case of diffusions has been studied by Li et al. (2024).

5.3. Wasserstein distance in Flow Matching

The aim of this chapter is to evaluate how well Flow Matching performs depending on the number of observations n. To this end, we have to control the distance between $\mathbb{P}^*$ and $\mathbb{P}^{\hat{\psi}_1(Z)}$ with the flow $\hat{\psi}$ from (5.12). For the evaluation metric, we again use the Wasserstein-1 metric.
As already motivated in the introduction, Flow Matching admits two natural reference models: the first is the model corresponding to the KDE, with p_t^n and v_t^n from (5.10). Using v_t^n as a vector field for the ODE (5.1), we obtain the flow ψ_t^n. Then we can bound

$$\mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{\hat{\psi}_1(Z)}) \le \mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{\psi_1^n(Z)}) + \mathsf{W}_1(\mathbb{P}^{\psi_1^n(Z)}, \mathbb{P}^{\hat{\psi}_1(Z)}). \tag{5.16}$$

A second reference model is the population model with p_t and v_t from (5.4) and (5.5). Using v_t as a vector field for the ODE (5.1), we obtain the flow ψ_t. In this case we can bound

$$\mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{\hat{\psi}_1(Z)}) \le \mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{\psi_1(Z)}) + \mathsf{W}_1(\mathbb{P}^{\psi_1(Z)}, \mathbb{P}^{\hat{\psi}_1(Z)}). \tag{5.17}$$

The following theorem provides a first comparison of the performance of $\hat{\psi}_t$ and a flow $\tilde{\psi}_t$ obtained as a solution of the ODE (5.1) using a vector field $\tilde{v}$ such that the solution exists.

Theorem 5.5. *Let $\tilde{\psi}$ be the solution of the ODE (5.1) using a Lipschitz continuous vector field $\tilde{v}$. Assume that all functions in $\mathcal{M}$ are Lipschitz continuous for fixed t with Lipschitz constant Γ_t. Then for any $\hat{v}\colon [0,1]\times\mathbb{R}^d \to \mathbb{R}^d$, with $\hat{v} \in \mathcal{M}$ and the corresponding solution $\hat{\psi}$ of the ODE (5.1),*

$$\begin{aligned}\mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{\hat{\psi}_1(Z)}) &\le \mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{\tilde{\psi}_1(Z)}) + \sqrt{2}e e^{\int_0^1 \Gamma_t \,\mathrm{d}t}\Big(\int_0^1 \int |\tilde{v}_t(x) - \hat{v}_t(x)|^2 \tilde{p}_t(x)\,\mathrm{d}x\,\mathrm{d}t\Big)^{\frac{1}{2}} \\ &\le \mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{\tilde{\psi}_1(Z)}) + \sqrt{2}e e^{\int_0^1 \Gamma_t \,\mathrm{d}t}\|\tilde{v} - \hat{v}\|_\infty.\end{aligned} \tag{5.18}$$

The first term in (5.18) depends on the choice of the conditional densities $p_t(\cdot|X_i^*)$ and on the number of observations, but is independent of the class $\mathcal{M}$. Within the framework of Section 2.6, we used $\mathbb{P}^{\tilde{\psi}_1(Z)}$ as the reference measure. Using the empirical reference (5.16), $\mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{\psi_1^n(Z)})$ is the estimation error of a kernel density estimator studied in Chapter 4. Note that the empirical reference is not the empirical distribution, but a smoothed version of it. Using the population reference as in (5.17), $\mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{\psi_1(Z)})$ is the error of the convolution of the unknown distribution with the latent distribution.

The second term depends on the set $\mathcal{M}$ and its ability to approximate $\tilde{v}$. In case of the empirical reference, we can use the fact that $\hat{v}$ minimizes (5.11) directly. In contrast to Section 2.6, the Flow Matching model regularizes intrinsically and the relation between the evaluation distance and the optimization criterion is nonlinear. For a sufficiently rich class $\mathcal{M}$ this second term will be negligible in this setting. Using the population reference, we first need to apply a concentration inequality to reach an empirical risk minimization setting, which leads to the classical trade-off in the size of the class $\mathcal{M}$. As a consequence, we need to treat the Lipschitz constant more carefully in this case.

The proof of Theorem 5.5 uses Grönwall's Lemma, which leads to the factor $e^{\int_0^1 \Gamma_t \, \mathrm{d}t}$ in the second term in (5.18). This is standard in the analysis of ODEs, in context of flow-based generative models see Fukumizu et al. (2025) and Albergo & Vanden-Eijnden (2023). Compared to the SDE setting of diffusion models, there is no such result like Girsanov's theorem, which is typically used to study these models following Chen et al. (2023c) and leads to the avoidance of the exponential dependence on the Lipschitz constant. This greatly complicates the analysis of Flow Matching. Given this dependence, a Lipschitz regularization in the Flow Matching objective seems to be theoretically beneficial. This could be a possible avenue for future research.

Remark 5.6. *A bound for arbitrary Lipschitz $\tilde{v}_t$ cannot be better than $\mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{\psi_1^n(Z)})$. If v_t^n is contained in $\mathcal{M}$, then the Picard-Lindelöf theorem yields $\hat{\psi}_t = \psi_t^n$ and hence $\mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{\hat{\psi}_1}) = \mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{\psi_1^n})$.*

In Section 5.4, we are going to study the setting using the empirical reference based on the results of Section 4.1. Subsequently, in Section 5.5 we investigate the setting using the population reference.

5.4. Rate of convergence in the over-parameterized setting

In this section, we focus on the case where $\mathbb{P}^*$ is a distribution on a compact set $\mathcal{X} \subset [-1, 1]^d$.

If the empirical vector field v^n is contained in $\mathcal{M}$, then the second error term in Theorem 5.5 vanishes and the Flow Matching model inherits, for suitable choices of the latent distribution, the optimal rate of convergence from Theorem 4.3. In practice, (5.11) is optimized over a class of neural networks to find a good approximation of v^n. We thus have to take this approximation error into account.

To analyze the ability of a network to approximate v^n, it is necessary to know the properties of v^n for all $t \in [0, 1]$. Hence we are going to specify σ_t, μ_t and a latent distribution for the subsequent analysis in this subsection.

Assumption 5.7.

1. *We consider the following choices of* $\sigma\colon [0,1]\times\mathbb{R}^d \to \mathbb{R}_{>0}$ *and* $\mu\colon [0,1]\times\mathbb{R}^d \to \mathbb{R}^d$ *:*

$$\sigma_t(X_i^*) = 1-(1-\sigma_{\min})t \quad \textit{and} \quad \mu_t(X_i^*) = tX_i^*.$$

2. *We choose* $\mathbb{U} = \mathcal{N}(0, I_d)$ *for the latent distribution.*

3. $d \geq 2$.

The choice of σ_t and μ_t in Assumption 5.7 coincides with the setting of Lipman et al. (2023) in Example II. In Section 5.1.2 we already saw other specific shapes of mean shifts and variance functions. We are going to study v_t for general σ_t in Section 5.5.
Now we want to combine the approximation properties of ReLU networks with the results of Chapter 4 to evaluate the performance of the Flow Matching mechanism from (5.12) where $\mathcal{M} = \mathsf{NN}(L, M, \Gamma)$ is a set of ReLU networks $v_{\mathsf{NN}}\colon [0,1]\times\mathbb{R}^d \to \mathbb{R}^d, (t,x)\mapsto v_t(x)$ with a fixed maximal number of layers L, at most M nonzero weights and Lipschitz constant of v_t of at most Γ for any t. We obtain the following rate:

Theorem 5.8. *Grant Assumption 5.7 and assume* $p^* \in B^{\alpha}_{1,\infty}(M, [-1,1]^d), \alpha \in (0,1]$. *Set* $\sigma_{\min} = n^{-\frac{1}{2\alpha+d}}$. *Then there are sequences* $L_n, M_n, \Gamma_n \in \mathbb{N}$ *such that for* n *big enough*

$$\mathbb{E}\big[\mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{\hat{\psi}_1})\big] \lesssim n^{-\frac{1+\alpha}{2\alpha+d}} \log^2(n),$$

where $\hat{\psi}$ *is given by* (5.12) *with* $\mathcal{M} = \mathsf{NN}(L_n, M_n, \Gamma_n)$.

Remark 5.9.

1. *The Lipschitz constraint in Theorem 5.8 is typically not enforced in practice. However, we only require a bound on the Lipschitz constant of order* $\Gamma_n = \frac{2d+1}{\sigma_{\min}^3} + \frac{1}{2}$ *which is a mild restriction for* $\sigma_{\min} \to 0$ *as* $n \to \infty$.

2. *Compared to the literature on diffusion models and the first results on Flow Matching, we do not use early stopping times, but rather set* $\sigma_{\min} > 0$ *according to* n. *In terms of variance, both procedures are interchangeable: calculating* σ_{t_*} *in Fukumizu et al. (2025, Theorem 9) leads to exactly the same variance. We note that this separation is unaffected by the flaw earlier in their proof. However, early stopping leads to a bias induced by the mean function* μ, *which is not* 1 *at a time* $t < 1$. *This could be fixed using an adapted mean function, which is* 1 *at the time of the early stopping, but this leads to much more complicated structures of* μ.

Theorem 5.8 shows that Flow Matching achieves minimax optimal rates (up to logarithmic factors) for certain classes of unknown densities, giving some justification of their empirical success.
The rate in Theorem 5.8 depends on the dimension d and therefore exhibits the classical curse of dimensionality. In the following, we analyze the case, which enabled to circumvent the curse

of dimensionality in Section 4.2. Hence, we assume the setting of Theorem 4.5, where $\mathbb{P}^*$ lives on a linear subspace of dimension d'. For a detailed definition, we refer to the previous chapter. Analogous to Theorem 5.8, we obtain a rate of convergence in case ReLU networks are used for the vector field.

Theorem 5.10. *Grant the setting of Theorem 4.5, Assumption 5.7 and assume that for every $y \in \operatorname{supp}(p^*_{d'})$ we have that $Ay \in [-1,1]^d$, where A is the matrix from the setting of Theorem 4.5. Set $\sigma_{\min} = n^{-\frac{1}{d'}}$. Then there are sequences $L_n, M_n, \Gamma_n \in \mathbb{N}$ such that for n big enough*

$$\mathbb{E}\big[\mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{\hat{\psi}_1})\big] \lesssim n^{-\frac{1}{d'}} \log^2(n),$$

where $\hat{\psi}$ is given by (5.12) with $\mathcal{M} = \mathsf{NN}(L_n, M_n, \Gamma_n)$.

Theorem 4.5 and Theorem 5.10 demonstrate that Flow Matching can benefit from a lower intrinsic dimension. Note that the method automatically adapts to the linear subspace, only the dimension d' is required for the choice of $\sigma_{\min}$. However, the smaller error bound comes at the cost of even larger networks to ensure that the approximation error of the network remains negligible compared to the faster rate of convergence in the dominating first term in decomposition (5.18). Additionally, the rate does not exploit regularity of $p^*_{d'}$. Therefore, the above results can only serve as first step to an analysis in a dimension reduction setting. A natural next step would be a generalization from linear subspaces to d'-dimensional sub-manifolds.

5.5. Rate of convergence via Lipschitz guarantees

While the rate in Theorem 5.8 is minimax optimal up to logarithmic factors, the network size is extremely large. Although this is part of the novelty of the result using the empirical reference, it is unrealistic in practical implementations. In this section, we are going to study the population reference. The use of a concentration inequality will lead to a trade-off in the network size. While this results in smaller networks, which align more closely with practice, it also prohibits the compensation for the exponential bound on the Lipschitz constant. Thus, we need to study the Lipschitz constant more carefully.
We are going to focus on one specific probability path and assume the following:

Assumption 5.11. *We assume that*

$$p_t(x|y) \propto \exp\Big(-\frac{|x-\mu_t(y)|^2}{2\sigma_t^2}\Big),$$

where the mean shift $\mu\colon [0,1]\times\mathbb{R}^d \to \mathbb{R}^d$ and the variance function $\sigma\colon [0,1] \to \mathbb{R}_{>0}$ are smooth, (coordinate-wise) monotone functions such that

$$\mu_0(y) = 0, \quad \mu_1(y) = y, \qquad \sigma_0 = 1, \quad \sigma_1 = \sigma_{\min},$$

for a $\sigma_{\min} \in (0,1)$.

Assumption 5.11 means in particular that the latent distribution is $\mathcal{N}(0, I_d)$. The monotonicity assumption guarantees that there is no unnecessary movement of mass. Assumption 5.11 coincides with the construction in Lipman et al. (2023, Section 4). Very similar generative models based on probability flows have been studied by Gao et al. (2024b,a). Additionally, the use of $\mathcal{N}(0, I_d)$ for the latent distribution is pervasive in generative modeling, e.g. in diffusions building up on Sohl-Dickstein et al. (2015), which lead to the SDE based diffusions introduced in Section 5.1.2.
For the smoothness of the unknown distribution, we stick to the setting of Section 5.4.

Assumption 5.12. *We assume that*

1. $d \geq 2$.
2. $p^* \in B^{\alpha}_{1,\infty}(\mathbb{R}^d), \quad \alpha \in (0, 1]$.

5.5.1. Lipschitz constant of the vector field

Under Assumption 5.11 Lipman et al. (2023), show that the corresponding vector field that generates $p_t(\cdot|y)$ for every $y \in \mathbb{R}^d$ is given by

$$v_t(x|y) = \frac{\sigma_t'}{\sigma_t}(x - \mu_t(y)) + \mu_t'(y), \tag{5.19}$$

where σ_t' and μ_t' are the derivatives in time. This is a special instance of Lemma 5.4 with shortened notation. We conclude from this result that the definition of the variance function σ_t is critical for bounding the Lipschitz constant, while only very special choices of the mean shift μ_t have an influence. Therefore, we will focus on polynomial choices of μ_t, which generalize the choice of Lipman et al. (2023).

Assumption 5.13. *We consider the following choices of* $\mu\colon [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$ *for* $\gamma > 1$*:*

$$\mu_t(y) = t^\gamma y.$$

A natural first idea is to aim for a global Lipschitz constant, this corresponds to the assumption that v has a bounded Lipschitz constant. The next result shows that such a constant cannot exist in this setting. We can upper and lower bound the Lipschitz constant of v:

Theorem 5.14. *Let*

$$B^t_{i,j} \coloneqq \Big\| \frac{\sigma_t'}{\sigma_t} \mathbb{1}_{i=j} + \big(\gamma t^{\gamma-1} - \frac{\sigma_t' t^\gamma}{\sigma_t}\big) \frac{t^\gamma}{\sigma_t^2} \operatorname{Cov}(Y^{\cdot,t})_{ij} \Big\|_\infty,$$

where $Y^{x,t}$ *is a random variable with density* $q \propto p_t(x|\cdot)p^*(\cdot)$ *for* $x \in \mathbb{R}^d$*. Assume* $B_{i,j}$ *exists for all* $i, j \in \{1, .., d\}$ *and define* $B \coloneqq (B_{ij})_{i,j=1,\dots,d}$*.*
For the Lipschitz constant in space Γ_t *of* v_t *on* $\mathbb{R}^d$*, we have that*

$$\max_{ij} B_{i,j} \leq \Gamma_t \leq d \max_{ij} B_{i,j}.$$

The upper and lower bound in Theorem 5.14 depend on the choice of σ_t. We want to choose σ_t such that $\left|\frac{\sigma_t'}{\sigma_t}\right|$ does not grow too fast for small σ_t or large $|\sigma_t'|$. The next result shows that it is not possible to choose a function σ_t such that the absolute value of the quotient is smaller than $\log(\sigma_{\min}^{-1})$ for all $t \in [0,1]$ or that the integral over $\left|\frac{\sigma_t'}{\sigma_t}\right|$ remains small.

Lemma 5.15.

1. *There is a $t^* \in [0,1]$ such that*
$$\frac{\sigma_{t^*}'}{\sigma_{t^*}} = \log(\sigma_{\min}),$$
for any choice of σ_t satisfying Assumption 5.11.

2. *For any choice of σ_t satisfying Assumption 5.11,*
$$\int_0^1 \left|\frac{\sigma_t'}{\sigma_t}\right| \mathrm{d}t = \log(\sigma_{\min}^{-1}).$$

We note that Lemma 5.15 does not depend on the specific probability path, it holds for all quotients of smooth choices of σ_t. Furthermore, similar terms arise in all cases, where the chosen probability path implies vector fields of the form (5.19). The first part of Lemma 5.15 reveals that a global Lipschitz bound on v_t independent of $\sigma_{\min}$ and the covariance term is not feasible in this setting. To compensate for $\frac{\sigma_t'}{\sigma_t} = \log(\sigma_{\min})$ in the case of $i = j$, the second term must have the same absolute value. However, this prohibits a smaller bound when $i \neq j$. The second part of Lemma 5.15 implies that, even the case of $\mathrm{Cov}(Y^{x,t})_{ij} \lesssim \sigma_t^2$ for all i, j will lead to a logarithmic dependency on $\sigma_{\min}^{-1}$, which ultimately influences the rate.
To control the Lipschitz constant, we thus need to assume that p^* is such that the covariance term in Theorem 5.14 decays in a controlled order for $\sigma_t \to \sigma_{\min}$ for $i = j$ and the same decay of the off-diagonal elements is fast enough.

Assumption 5.16. *Assume there is a $t^* \in \left[\frac{1}{2^{1/\gamma}}, 1\right)$ independent of $\sigma_{\min}$ such that for all $x \in \mathbb{R}^d$*

$$\begin{aligned}
&\text{(I)} && \mathrm{Cov}(Y^{x,t})_{ij} \lesssim \left(\frac{\sigma_t}{t^\gamma}\right)^3, && i \neq j, && t > t^*,\\
&\text{(II)} && \mathrm{Var}(Y_i^{x,t}) = \left(\frac{\sigma_t}{t^\gamma}\right)^2 \left(1 + O\left(\left(\frac{\sigma_t}{t^\gamma}\right)^{\frac{1}{\kappa}}\right)\right), && \text{for all } i, && t > t^*,\\
&\text{(III)} && \mathrm{Cov}(Y^{x,t})_{ij} \leq C, && \text{for all } i, j, && t \leq t^*,
\end{aligned}$$

where $Y^{x,t}$ is a random variable with density $q \propto p_t(x|\cdot)p^*(\cdot)$ and $\kappa \in \mathbb{R}_{\geq 1}$ and C are fixed constants.

Under these assumptions, any choice of σ_t satisfying Assumption 5.11 will lead to a bounded Lipschitz constant.

Theorem 5.17. *Grant Assumption 5.11, Assumption 5.13 and Assumption 5.16 with fixed parameters γ, κ and t^*. Then there is a constant C independent of $\sigma_{\min}$ such that*

$$\int_0^1 \Gamma_t \, \mathrm{d}t \leq C.$$

In order to allow for comparison with other works, we are going to show that the very abstract Assumption 5.16 is satisfied if the unknown distribution is of the form

$$p^*(x) \propto \exp\big(-\frac{|x|^2}{2} - a(x)\big), \quad \|a\|_{C^2} = L < \infty. \tag{5.20}$$

This setting was used in Stéphanovitch (2024) to construct Lipschitz continuous pushforward maps in diffusion models. It specifically allows for unbounded distribution that are not log-concave (i.e the logarithmic density is concave) and thus the Brascamp-Lieb inequality, Theorem 2.24, cannot be applied directly. The log-concavity assumption has been popular in the analysis of diffusion models and Flow Matching, see for example Gao & Zhu (2025); Bruno et al. (2025); Gao et al. (2025, 2024b). First, we note that by Lemma 5.33, which can be found in Section 5.6.4, the second assumption in Assumption 5.12 is satisfied for densities of the form (5.20) with $\alpha = 1$. For simplicity, we choose $\gamma = 1$, but the result extends easily to other γ.

Theorem 5.18. *Set $\gamma = 1$. Assume that $\mathbb{P}^*$ is of the form (5.20). Then Assumption 5.16 is fulfilled with*

$$C = e^{2L}, \quad \kappa = 1, \quad \text{and} \quad t^* \quad \text{such that} \quad \sigma_{t^*} = c(L,d),$$

where $c(L,d) \in (0,1)$ is a constant that depends only on L and d.

The uniform covariance bound is proven using the fact that we can control the effect of bounded perturbations of Gaussians on the variance, which follows from the Holley-Strooke perturbation principle. The bound on the diagonal entries employs proof techniques from the Cramer-Rao lower bound. The bound on the off-diagonal elements is developed from the Brascamp-Lieb type covariance estimate of Menz (2014).
The proof of Theorem 5.18 reveals that for $t > t^*$ large enough and $i \neq j$, the covariance decay is of order σ_t^4 and $\kappa = 1$. This suggests that the class of distributions for which Theorem 5.17 applies is significantly larger.

5.5.2. Rate of convergence

In order to evaluate the performance of Flow Matching, we are going to derive a rate of convergence using neural networks for the set $\mathcal{M}$. Since $\tilde{v}$ is chosen such that it minimizes (5.11) instead of (5.3), we cannot use the minimization property in Theorem 5.5 directly. Within the framework of Section 2.6, this corresponds to $\varepsilon_c^n \neq 0$. The next theorem provides a more detailed decomposition of the error in the form of a classical oracle inequality. In order to merge the results later, we choose μ_t such that the results of Section 5.5.1 apply. For the variance function σ_t, the analysis in Section 5.5.1 reveals that we are not restricted to the linear case. Thus, we can choose a variance function that is suited for the application of a Bernstein-type inequality. Since Bernstein-type inequalities rely on absolute value bounds, we are going to choose σ_t such that the absolute value of (5.19) is as small as possible. Our choice follows from Lemma 5.15 and the solution of the ODE

$$\frac{\sigma_t'}{\sigma_t} = \log(\sigma_{\min}), \quad \sigma_0 = 1.$$

Assumption 5.19. *Assume that*

$$\mu_t(y) = ty \quad \textit{and} \quad \sigma_t = (\sigma_{\min})^t.$$

First, we choose $\mathcal{M}$ as a set of continuous measurable functions $\tilde{v}\colon [0,1] \times \mathbb{R}^d \to \mathbb{R}$ such that for the j-th component function $\tilde{v}^j$ of $\tilde{v}$

$$|\tilde{v}^j|_\infty \leq e^{2L} \log(n)^3 + (1 + e^{2L}) \log(n)^2 + \log(n) + 1. \tag{5.21}$$

We are going to justify this bound in the proof of the following theorem.

Theorem 5.20. *Let p^* be of the form* (5.20) *and grant Assumption 5.11 and Assumption 5.19. Assume that $\log(\sigma_{\min}) \asymp \log(n)$. Then we have for every $a \in (0,1], \tau \in \mathbb{R}_{>0}, b > 1$ and n large enough with probability of $1 - \frac{1}{n}$*

$$\begin{aligned}\mathbb{E}_{\substack{t\sim\mathcal{U}[0,1]\\ X_t\sim p_t}}\left[|v_t(X_t) - \hat{v}_t(X_t)|^2\right] \lesssim \inf_{\tilde{v}\in\mathcal{M}} \int\int |\tilde{v}_t(x) - v_t(x)|^2 p_t(x)\,\mathrm{d}x\,\mathrm{d}t + n^{-\frac{1}{2}} \\ + \frac{a^{-1}\log(n)^7}{n} \log\left(2\mathcal{N}(\tau, g(\mathcal{M}), \|\cdot\|_\infty)\right) + (2+a)\tau + a\log(n)^6.\end{aligned} \tag{5.22}$$

The proof of Theorem 5.20 uses the Bernstein-type concentration inequality of Chen et al. (2023b), which is commonly used in the analysis of diffusion models, see for example Yakovlev & Puchkin (2025). Theorem 5.20 shows that a careful choice of the variance shift can allow for to a logarithmic dependency of the oracle bounds on $\sigma_{\min}^{-1}$ instead of a linear dependency.
Using networks of the form (2.22) clipped at (5.21) for the set $\mathcal{M}$, the next result combines the findings of Section 5.5.1, Theorem 4.3 and Theorem 5.20 with the approximation theory of Theorem 2.6 that allows for simultaneous approximation of a function and its derivative. To facilitate the transfer of results, we refrain from inserting $\alpha = 1$, which is the applicable smoothness in this setting as shown in Lemma 5.33.

Theorem 5.21. *Let p^* be of the form* (5.20) *and grant Assumption 5.11 and Assumption 5.19. Then for n large enough with probability of $1 - \frac{1}{n}$ for fixed $\eta > 0$*

$$\mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{\hat{\psi}_1(Z)}) \lesssim \mathrm{polylog}(n) n^{-\frac{1+\alpha}{d+4\alpha+5+\eta}},$$

where $\hat{\psi}$ is the solution of an ODE whose vector field is in $\mathcal{M}$ given by a ReLU neural network, with no more than $c \cdot \log(n)$ layers, $c \cdot n^{c'(d,\alpha,\eta)} \cdot \log^2(n)$ nonzero weights, where c and $c'(d,\alpha,\eta)$ is a constant independent of n.

The proof of Theorem 5.21 exploits the fact that the vector field v from (5.5) is by construction in C^∞. In order to bound the higher order derivatives of v, we leverage the fact that densities of the form (5.20) satisfy a logarithmic Sobolev inequality with a controlled constant. The rate in Theorem 5.21 benefits from smoothness in the unknown distribution. Since smoothing is an intrinsic property of Flow Matching, it is desirable that the rate also reflects this property. Even though the rate is not optimal, the gap to the optimal rate of $n^{-\frac{1+\alpha}{2\alpha+d}}$ is small for large d.

Compared to Theorem 5.8, our result applies to networks with logarithmically growing depth and polynomially growing numbers of non-zero weights. On the other hand, the rate deteriorates slightly. However, the two results cannot be compared directly. Theorem 5.8 assumes that the support of p^* is compact, whereas Theorem 5.21 assumes full support on $\mathbb{R}^d$. The evaluation metric differs from that in Gao et al. (2024b), who use the Wasserstein-2 metric. According to Villani (2008, Remark 6.5), their rate $n^{-\frac{1}{d+5}}$ is also valid for the Wasserstein-1 metric. For $\alpha \neq 0$ and large d, the rate obtained in Theorem 5.21 is faster. However, this is just a rough comparison, as the unknown distributions studied are different. Furthermore, they use a linear variance shift and an early stopping approach, which hinders direct comparison even more.

5.6. Proofs

5.6.1. Proofs of Section 5.2

Proof of Lemma 5.2. A necessary and sufficient condition for v_t to generate p_t is given in (5.2). We thus verify

$$\begin{aligned}\frac{d}{dt}p_t^n(x) &= \frac{1}{n}\sum_{i=1}^n \frac{d}{dt}p_t(x|X_i^*) = -\frac{1}{n}\sum_{i=1}^n \operatorname{div}(p_t(x|X_i^*)v_t(x|X_i^*)) \\ &= -\operatorname{div}\Big(\frac{1}{n}\sum_{i=1}^n p_t(x|X_i^*)\sum_{i=1}^n \frac{p_t(x|X_i^*)v_t(x|X_i^*)}{\sum_{i=1}^n p_t(x|X_i^*)}\Big) = -\operatorname{div}(p_t^n(x)v_t^n(x)).\end{aligned}$$ □

Proof of Theorem 5.3. For fixed $t \in [0,1]$ we have

$$\begin{aligned}|\hat{v}_t(x) - v_t^n(x)|^2 &= |\hat{v}_t(x)|^2 - 2\langle \hat{v}_t(x), v_t^n(x)\rangle + |v_t^n(x)|^2, \\ |\hat{v}_t(x) - v_t(x|X_i^*)|^2 &= |\hat{v}_t(x)|^2 - 2\langle \hat{v}_t(x), v_t(x|X_i^*)\rangle + |v_t(x|X_i^*)|^2.\end{aligned}$$

The last term does not influence the minimal argument in $\hat{v}$. For the first two we have

$$\begin{aligned}\mathbb{E}_{X_t\sim p_t^n}[|\hat{v}_t(X_t)|^2] &= \int |\hat{v}_t(x)|^2 p_t^n(x)\,\mathrm{d}x \\ &= \frac{1}{n}\sum_{i=1}^n \int |\hat{v}_t(x)|^2 p_t(x|X_i^*)\,\mathrm{d}x \\ &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\tilde{X}_t\sim p_t(\cdot|X_i^*)}[|\hat{v}_t(\tilde{X}_t)|^2],\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}_{X_t\sim p_t^n}[\langle \hat{v}_t(X_t), v_t^n(X_t)\rangle] &= \int \langle \hat{v}_t(x), v_t^n(x)\rangle p_t^n(x)\,\mathrm{d}x \\ &= \int \Big\langle \hat{v}_t(x), \sum_{i=1}^n v_t(x|X_i^*)\frac{p_t(x|X_i^*)}{\sum_{i=1}^n p_t(x|X_i^*)}\Big\rangle \frac{1}{n}\sum_{i=1}^n p_t(x|X_i^*)\,\mathrm{d}x \\ &= \frac{1}{n}\sum_{i=1}^n \int \langle \hat{v}_t(x), v_t(x|X_i^*)\rangle p_t(x|X_i^*)\,\mathrm{d}x\end{aligned}$$

$$= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\tilde{X}_t \sim p_t(\cdot|X_i^*)}[\langle \hat{v}_t(\tilde{X}_t), v_t(\tilde{X}_t|X_i^*)\rangle]. \qquad \square$$

5.6.2. Proofs of Section 5.3

Proof of Lemma 5.4. As $\sigma_t(X_i^*) > 0$ for all t, X_i^*, we have that

$$\psi_t^{-1}(x|X_i^*) = \frac{x - \mu_t(X_i^*)}{\sigma_t(X_i^*)}.$$

Hence

$$\frac{\partial \psi_t}{\partial t}(z|X_i^*)\Big|_{z=\psi_t^{-1}(x|X_i^*)} = v_t(x|X_i^*),$$

and

$$\frac{\partial \psi_t}{\partial t}(x|X_i^*) = \frac{\partial \sigma_t}{\partial t}(X_i^*)x + \frac{\partial \mu_t}{\partial t}(X_i^*).$$

Thus, we get

$$v_t\left(x|X_i^*\right) = \frac{\frac{\partial \sigma_t}{\partial t}\left(X_i^*\right)}{\sigma_t\left(X_i^*\right)}\big(x - \mu_t\left(X_i^*\right)\big) + \frac{\partial \mu_t}{\partial t}\left(X_i^*\right). \qquad \square$$

Proof of Theorem 5.5. Since we assumed that $\mathbb{U}$ has finite first moment and $\tilde{v}$ and $\hat{v}$ are Lipschitz continuous, $\mathbb{P}^{\tilde{\psi}_1(Z)}$ and $\mathbb{P}^{\hat{\psi}_1(Z)}$ have finite first moment. This implies finiteness of $\mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{\hat{\psi}_1(Z)}), \mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{\tilde{\psi}_1(Z)})$ and $\mathsf{W}_1(\mathbb{P}^{\tilde{\psi}_1(Z)}, \mathbb{P}^{\hat{\psi}_1(Z)})$. As W_1 satisfies the triangle inequality,

$$\mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{\hat{\psi}_1(Z)}) \leq \mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{\tilde{\psi}_1(Z)}) + \mathsf{W}_1(\mathbb{P}^{\tilde{\psi}_1(Z)}, \mathbb{P}^{\hat{\psi}_1(Z)}).$$

Using Villani (2008, Remark 6.5) we get

$$\begin{aligned}
\mathsf{W}_1(\mathbb{P}^{\tilde{\psi}_1(Z)}, \mathbb{P}^{\hat{\psi}_1(Z)}) &\leq \mathsf{W}_2(\mathbb{P}^{\tilde{\psi}_1(Z)}, \mathbb{P}^{\hat{\psi}_1(Z)}) \\
&\leq \sqrt{\mathbb{E}[|\tilde{\psi}_1(Z) - \hat{\psi}_1(Z)|^2]} \\
&= \Big(\int |\tilde{\psi}_1(x) - \hat{\psi}_1(x)|^2 p_0(x)\,\mathrm{d}x\Big)^{1/2}.
\end{aligned}$$

Like in Albergo & Vanden-Eijnden (2023, Proposition 3) set

$$Q_t = \int |\tilde{\psi}_t(x) - \hat{\psi}_t(x)|^2 p_0(x)\,\mathrm{d}x.$$

Since $p^*, \tilde{\psi},\ \hat{\psi}, \tilde{v}$ and $\hat{v}$ are integrable with respect to the Lebesgue measure, we can use the dominated convergence theorem and interchange integration and differentiation to obtain

$$\begin{aligned}
\frac{d}{dt}Q_t &= 2\int (\tilde{\psi}_t(x) - \hat{\psi}_t(x))^\top (\tilde{v}_t(\tilde{\psi}_t(x)) - \hat{v}_t(\hat{\psi}_t(x)))p_0(x)\,\mathrm{d}x \\
&= 2\int (\tilde{\psi}_t(x) - \hat{\psi}_t(x))^\top (\tilde{v}_t(\tilde{\psi}_t(x)) - \hat{v}_t(\tilde{\psi}_t(x)))p_0(x)\,\mathrm{d}x \\
&\quad + 2\int (\tilde{\psi}_t(x) - \hat{\psi}_t(x))^\top (\hat{v}_t(\tilde{\psi}_t(x)) - \hat{v}_t(\hat{\psi}_t(x)))p_0(x)\,\mathrm{d}x.
\end{aligned}$$

As $0 \leq |\tilde{\psi}_t(x) - \hat{\psi}_t(x) - (\tilde{v}_t(\tilde{\psi}_t(x)) - \hat{v}_t(\tilde{\psi}_t(x)))|^2$, we get

$$2(\tilde{\psi}_t(x) - \hat{\psi}_t(x))^\top (\tilde{v}_t(\tilde{\psi}_t(x)) - \hat{v}_t(\tilde{\psi}_t(x))) \leq |\tilde{\psi}_t(x) - \hat{\psi}_t(x)|^2 + |\tilde{v}_t(\tilde{\psi}_t(x)) - \hat{v}_t(\tilde{\psi}_t(x))|^2,$$

and since for fixed t the vector field $\hat{v}_t$ is Γ_t Lipschitz continuous

$$2(\tilde{\psi}_t(x) - \hat{\psi}_t(x))^\top (\hat{v}_t(\tilde{\psi}_t(x)) - \hat{v}_t(\hat{\psi}_t(x))) \leq 2\Gamma_t |\tilde{\psi}_t(x) - \hat{\psi}_t(x)|^2.$$

Therefore

$$\frac{d}{dt} Q_t \leq (1 + 2\Gamma_t) Q_t + 2 \int |\tilde{v}_t(\tilde{\psi}_t(x)) - \hat{v}_t(\tilde{\psi}_t(x))|^2 p_0(x) \, \mathrm{d}x.$$

Now we use the formulation of Grönwall's lemma in Walter (1970) and use $Q_0 = 0$ (this holds since $\psi_0(x) = \hat{\psi}_0(x) = x$). This leads to

$$Q_1 \leq e^{\int_0^1 1 + 2\Gamma_t \, \mathrm{d}t} 2 \int_0^1 \int |\tilde{v}_t(\tilde{\psi}_t(x)) - \hat{v}_t(\tilde{\psi}_t(x))|^2 p_0(x) \, \mathrm{d}x \, \mathrm{d}t. \tag{5.23}$$

By Theorem 5.3 the vector field $\hat{v}$ is chosen such that it minimizes $\int_0^1 \mathbb{E}_{X_t \sim p_t} \left[|\hat{v}_t(X_t) - \tilde{v}_t(X_t)|^2 \right] \mathrm{d}t$ and $X_t = \psi_t(Z)$, thus we get for every $\hat{v} \in \mathcal{M}$

$$\begin{aligned} Q_1 &\leq 2e^{\int_0^1 1 + 2\Gamma_t \, \mathrm{d}t} \int_0^1 \int |\tilde{v}_t(\tilde{\psi}_t(x)) - \hat{v}_t(\tilde{\psi}_t(x))|^2 p_0(x) \, \mathrm{d}x \, \mathrm{d}t \\ &\leq 2e^{\int_0^1 1 + 2\Gamma_t \, \mathrm{d}t} \int_0^1 \|\tilde{v}_t - \hat{v}_t\|_\infty^2 p_0(x) \, \mathrm{d}t, \\ &\leq 2e^{\int_0^1 1 + 2\Gamma_t \, \mathrm{d}t} \|\tilde{v} - \hat{v}\|_\infty^2. \end{aligned}$$

Taking the square root leads to the result. □

5.6.3. Proofs of Section 5.4

Preparations for the proof of Theorem 5.8 The proof of Theorem 5.8 requires some additional results. In order to control the error $e^{\int_0^1 \Gamma_t \, \mathrm{d}t} \|v^n - \hat{v}\|_\infty$ (where we omitted constants) we need to approximate the function v^n and control the Lipschitz constant of the apprximation $\hat{v}_t$ for fixed t. However, we cannot expect to approximate a non-tivial function on $\mathbb{R}^d$ with respect to the supremum norm. Additionally, v_t^n as constructed by Lipman et al. (2023) is only locally Lipschitz in x for fixed t and the bound of the Lipschitz constant grows for $\sigma_{\min} \to 0$ (see Lemma 5.23). Hence we cannot expect to construct a network $\hat{v}$ that approximates v^n and its Lipschitz constant on $\mathbb{R}^d$ with a small absolute error for every $x \in \mathbb{R}$.
In the given setting, a rapid decay of the kernel function results in less significant approximation errors outside of the support of p^*. We first show that in case of rapidly decreasing kernels the error $\mathsf{W}_1(\mathbb{P}^{\psi(Z)}, \mathbb{P}^{\hat{\psi}(Z)})$ can be decomposed into two approximation errors, one depending on the approximation of v on a compact domain and one corresponding to the area outside with little probability mass.

Lemma 5.22. *Assume that* $\operatorname{supp}(p^*) \subset (-1, 1)^d$ *and* $\sigma_{\min} \leq 1$. *Additionally, assume that* $K(x) = \prod_{i=1}^d \varphi(x_i)$, *where* φ *is a one-dimensional symmetric density that decays faster than* $x \mapsto e^{-\lambda x}$

for $|x| > 2$ and some $\lambda > 0$. Then for every $a > 1$ there is a network $\tilde{v}$, which is cutoff outside $(-a, a)^d$ such that

$$\mathsf{W}_1(\mathbb{P}^{\psi_1^n(Z)}, \mathbb{P}^{\hat{\psi}_1(Z)}) \leq \sqrt{2}e e^{\int_0^1 \Gamma_t \,\mathrm{d}t} \Big(\int_0^1 \int_{(-a,a)^d} |v_t^n(\psi_t^n(x)) - \tilde{v}_t(\psi_t^n(x))|^2 p_0(x) \,\mathrm{d}x \,\mathrm{d}t + c\frac{a^{2d+2}e^{-da}}{\sigma_{\min}^2} \Big)^{1/2}.$$

Next we want to choose a in Lemma 5.22 and a neural network of finite size (depending on n) such that the bound of $\mathsf{W}_1(\mathbb{P}^{\psi^n(Z)}, \mathbb{P}^{\hat{\psi}(Z)})$ decays at a rate of $n^{-\frac{1+\alpha}{2\alpha+d}}$. Hence we need to approximate

$$v_t^n(x) = \sum_{i=1}^n v_t(x|X_i^*) \frac{p_t(x|X_i^*)}{\sum_{i=1}^n p_t(x|X_i^*)}$$

and simultaneously control the Lipschitz constant of the approximation $\hat{v}$ on $(-a, a)^d$. This is indeed sufficient, since using the same reasoning as in the proof of Lemma 5.22, it holds that if $x \in (-a, a)^d$, then $v_t^n(x) \in (-a, a)^d$ for every t. As $\hat{v}$ approximates v^n in the supremum norm, the Lipschitz constant of $\hat{v}$ will be at least of the order of the Lipschitz constant of v^n. The same applies to the supremum norm. Hence we need bounds for both, the Lipschitz constant and the supremum norm of v^n.

Lemma 5.23. *Assume that* $\operatorname{supp}(p^*) \subset [-1, 1]^d$. *For every $x \in (-a, a)$ we have that*

$$|v_t^n(x)| \leq \frac{\sqrt{d}(1+a)}{1-(1-\sigma_{\min})t} \leq \frac{\sqrt{d}(1+a)}{\sigma_{\min}}.$$

Further for the Gaussian kernel and $\operatorname{supp}(p^*) \subset [-1, 1]^d$, *for any $x \in \mathbb{R}^d$ and every $t \in [0, 1]$*

$$\operatorname{Lip}(v_t^n) \leq \frac{1}{\sigma_t} + \frac{2d}{\sigma_t^3}.$$

Looking at the proof of Lemma 5.23 and using standard analysis, we can bound the second partial derivatives of v_t^n by $\frac{C}{\sigma_t^5}$ for some $C > 0$. Now we are ready to prove Theorem 5.8.

Proof of Theorem 5.8. From Lemma 5.22, we know that

$$\mathsf{W}_1(\mathbb{P}^{\psi_1^n(Z)}, \mathbb{P}^{\hat{\psi}_1(Z)}) \leq \sqrt{2}e e^{\int_0^1 \tilde{L}_t \,\mathrm{d}t} \inf_{\tilde{v} \in \mathcal{M}} \Big(\int_0^1 \int_{(-a,a)^d} |v_t^n(\psi_t^n(x)) - \tilde{v}_t(\psi_t^n(x))|^2 p_0(x) \,\mathrm{d}x \,\mathrm{d}t + c\frac{a^{2d+2}e^{-da}}{\sigma_{\min}^2} \Big)^{1/2},$$

where we used the choice of (5.12) before dividing the area in (5.24) in the proof of Lemma 5.22 and $\tilde{L}$ is the Lipschitz constant of $\tilde{v}$ at time t. We neglect an arbitrary small additional error on the right hand side to account for the use of the infimum, as the error bound corresponding to the kernel density estimator in (5.18) is nonzero anyway. Note that due to the restriction of the network class,

$$e^{\int_0^1 \tilde{L}_t \,\mathrm{d}t} \leq e \cdot e^{\int_0^1 \frac{2d+1}{\sigma_t^3} \,\mathrm{d}t}.$$

Integration yields for n big enough

$$\int_0^1 \frac{1}{\sigma_t^3}\,\mathrm{d}t = \int_0^1 \frac{1}{(1-(1-\sigma_{\min})t)^3}\,\mathrm{d}t \leq \frac{1}{2(1-\sigma_{\min})\sigma_{\min}^2} - \frac{1}{2(1-\sigma_{\min})} \leq \frac{1}{\sigma_{\min}^2}.$$

To bound the tail, we need to assure a is chosen such that

$$e^{\frac{2d+1}{\sigma_{\min}^2}} \frac{a^{d+1}e^{-\frac{da}{2}}}{\sigma_{\min}} \leq n^{-\frac{1+\alpha}{2\alpha+d}}.$$

As in the setting $d \geq 1$, we can use that for $a \geq 15$

$$a^{d+1}e^{-da/2} \leq e^{-\frac{da}{8}}.$$

For $\sigma_{\min} \asymp n^{-\frac{1}{2\alpha+d}}$, the above inequality is satisfied if

$$a \geq \max\Big(\frac{8}{d}\Big(\frac{2+\alpha}{2\alpha+d}\log(n) + (2d+1)n^{\frac{2}{2\alpha+d}}\Big), 15\Big).$$

To cover $(-a,a)^d$ with cubes of size $(0,1)^d$, we need $\lceil 2a \rceil^d$ little cubes.
Now we want to bound the approximation term. We can bound

$$\Big(\int_0^1 \int_{(-a,a)^d} |v_t^n(\psi_t^n(x)) - \tilde{v}_t(\psi_t^n(x))|^2 p_0(x)\,\mathrm{d}x\,\mathrm{d}t\Big)^{1/2} \leq \|v_t^n - \tilde{v}_t\|_{\infty,(-a,a)^d}.$$

Note that $v_t^n \in C^\infty$ in case of the Gaussian kernel. Hence we can apply Theorem 2.6. The bound of the second partial derivatives as well as the supremum norm bound of v^n influence the coefficients in Theorem 2.6, therefore we construct the network such that $n^{-\frac{5}{2\alpha+d}}v^n$ is approximated and choose the last weight matrix such that it scales the result up. This addition of one layer does not influence the order of the layers and non zero weights of the network. However, the approximation error gets scaled up too, hence we need to ensure a smaller error. We need to choose ε in Theorem 2.6 such that

$$e^{(2d+1)n^{\frac{2}{2\alpha+d}}}\sqrt{\varepsilon} \leq n^{-\frac{6+\alpha}{2\alpha+d}} \quad \Longleftrightarrow \quad \varepsilon \leq n^{-\frac{12+2\alpha}{2\alpha+d}} e^{-(4d+2)n^{\frac{2}{2\alpha+d}}}.$$

The addition of one cutting layer to ensure the assumptions of Lemma 5.22 does not influence the order of the layers and non zero weights of the network either. Hence choosing

$$\begin{aligned} L &\gtrsim \Big(\max\Big(\log(n) + n^{\frac{2}{2\alpha+d}}, 15\Big)\Big)^d \cdot \Big(\log^2(n) + n^{\frac{4}{2\alpha+d}}\Big), \\ M &\gtrsim \Big(\max\Big(\log(n) + n^{\frac{2}{2\alpha+d}}, 15\Big)\Big)^d \cdot \Big(n^{\frac{12+2\alpha}{2\alpha+d}} e^{(4d+2)n^{\frac{2}{2\alpha+d}}}\Big)^d \cdot \Big(\log^2(n) + n^{\frac{4}{2\alpha+d}}\Big), \end{aligned}$$

yields the desired rate. □

Proof of Theorem 5.10. We proceed completely analogous to Theorem 5.8 replacing d with d'. However, since the vector field we want to approximate is still d-dimensional, we still need $\lceil 2a \rceil^d$

little cubes. Hence we choose

$$L \gtrsim \Big(\max\Big(\log(n) + n^{\frac{2}{d'}}, 15\Big)\Big)^d \cdot \Big(\log^2(n) + n^{\frac{4}{d'}}\Big),$$
$$M \gtrsim \Big(\max\Big(\log(n) + n^{\frac{2}{d'}}, 15\Big)\Big)^d \cdot \Big(n^{\frac{12}{d'}} e^{(4d'+2)n^{\frac{2}{d'}}}\Big)^d \cdot \Big(\log^2(n) + n^{\frac{4}{d'}}\Big),$$

to obtain the desired rate. □

Additional proofs of Section 5.6.3

Proof of Lemma 5.22. First, recall the proof from Theorem 5.20. For any $a > 0$ we can bound the term from (5.23)

$$\begin{aligned} Q_1 &\leq \sqrt{e} e^{2\int_0^1 \Gamma_t \,\mathrm{d}t} \int_0^1 \int |v_t^n(\psi_t^n(x)) - \tilde{v}_t(\psi_t^n(x))|^2 p_0(x) \,\mathrm{d}x \,\mathrm{d}t \\ &= \sqrt{e} e^{2\int_0^1 \Gamma_t \,\mathrm{d}t} \Big(\int_0^1 \int_{(-a,a)^d} |v_t^n(\psi_t^n(x)) - \tilde{v}_t(\psi_t^n(x))|^2 p_0(x) \,\mathrm{d}x \,\mathrm{d}t \\ &\quad + \int_0^1 \int_{\mathbb{R}^d \setminus (-a,a)^d} |v_t^n(\psi_t^n(x)) - \tilde{v}_t(\psi_t^n(x))|^2 p_0(x) \,\mathrm{d}x \,\mathrm{d}t \Big). \end{aligned} \tag{5.24}$$

The second term can be bounded by

$$\begin{aligned} &\int_0^1 \int_{\mathbb{R}^d \setminus (-a,a)^d} |v_t^n(\psi_t^n(x)) - \tilde{v}_t(\psi_t^n(x))|^2 p_0(x) \,\mathrm{d}x \,\mathrm{d}t \\ &\qquad \leq 2 \int_0^1 \int_{\mathbb{R}^d \setminus (-a,a)^d} \Big(|v_t^n(\psi_t^n(x))|^2 + |\tilde{v}_t(\psi_t^n(x))|^2 \Big) p_0(x) \,\mathrm{d}x \,\mathrm{d}t. \end{aligned} \tag{5.25}$$

Now

$$\begin{aligned} |v_t^n(\psi_t^n(x))|^2 &\leq \max_{i \in \{1,\dots,n\}} |v_t(\psi_t^n(x) | X_i^*)|^2 \\ &\leq \frac{|X_i^* - (1-\sigma_{\min})\psi_t^n(x)|^2}{(1-(1-\sigma_{\min})t)^2} \\ &\leq 2\frac{|X_i^*|^2 + (1-\sigma_{\min})^2 |\psi_t^n(x)|^2}{(1-(1-\sigma_{\min})t)^2}. \end{aligned}$$

As $\mathbb{P}^*$ has compact support within $(-a,a)^d$,

$$|X_i^*|^2 \leq d \cdot a^2.$$

To bound $\psi_t^n(x)$ for $x \in \mathbb{R}^d \setminus (-a,a)^d$, we use that

$$|\psi_t^n(x)| \leq |x|. \tag{5.26}$$

To see why this is true, consider two cases. First, let t be small such that $\psi_t^n(x) \notin (-a,a)^d$. Then for every $X_i^*, i \in \{1,\dots,n\}$ we have by construction that $v_t^n(\psi_t^n(x)|X_i^*)$ is in the smallest convex cone that includes $\psi_t^n(x)$ and $(-1,1)^d$, where $\psi_t^n(x)$ is the vertex of the convex cone. Any linear

combination of $v_t^n(\psi_t^n(x)|X_i^*), i = 1, \dots, n$, must be in this convex cone as well. Therefore $v_t^n(x)$ must be in the convex cone. Since

$$\frac{d}{dt}\psi_t^n(x) = v_t^n(\psi_t(x)), \quad \psi_0^n(x) = x,$$

this means that for small t, ψ_t pushes the initial condition x in the direction of $(-a, a)^d$ and therefore (5.26) is true.
Second, let t' be the smallest t such that $\psi_{t'}^n(x) \in (-a, a)^d$. Then for all $t > t'$ we have that $\psi_t^n(x) \in (-a, a)^d$. This holds since within $(-a, a)^d \setminus (-1, 1)^d$ using the same argument as in the first case, $v_t^n(x)$ is a linear combination of $v_t^n(\psi_t^n(x)|X_i^*)$ and thus ψ_t pushes x in the direction of $(-1, 1)^d$. If t is such that $\psi_t(x) \in (-1, 1)^d$, there cannot be a $t' > t$ such that $\psi_{t'}(x) \notin (-a, a)^d$, since by continuity of ψ_t, there must be a t'' with $t < t'' < t'$ such that $\psi_{t''}(x) \in (-a, a)^d$. In this case the corresponding vector field is oriented towards $(-1, 1)^d$, which leads to ψ_t pushing x back to $(-1, 1)^d$. Hence, in any case for $x \notin (-a, a)^d$, we have that (5.26) is true.
Therefore we get for (5.25) for $\sigma_{\min} \leq 1$

$$\begin{aligned}
&\int_0^1 \int_{\mathbb{R}^d \setminus (-a,a)^d} |v_t^n(\psi_t^n(x)) - \tilde{v}_t(\psi_t^n(x))|^2 p_0(x) \, \mathrm{d}x \, \mathrm{d}t \\
&\qquad \leq 8 \int_0^1 \int_{\mathbb{R}^d \setminus (-a,a)^d} \frac{da^2 + (1 - \sigma_{\min})^2 |x|^2}{(1 - (1 - \sigma_{\min})t)^2} p_0(x) \, \mathrm{d}x \, \mathrm{d}t \\
&\qquad \leq \frac{8da^2}{\sigma_{\min}^2} \int_0^1 \int_{\mathbb{R}^d \setminus (-a,a)^d} p_0(x) \, \mathrm{d}x \, \mathrm{d}t + \frac{(1 - \sigma_{\min})^2}{\sigma_{\min}^2} \int_0^1 \int_{\mathbb{R}^d \setminus (-a,a)^d} |x|^2 p_0(x) \, \mathrm{d}x \, \mathrm{d}t.
\end{aligned}$$

For the first term we get

$$\frac{8da^2}{\sigma_{\min}^2} \int_{\mathbb{R}^d \setminus (-a,a)^d} p_0(x) \, \mathrm{d}x = \frac{8da^2}{\sigma_{\min}^2} \Big(\int_{\mathbb{R} \setminus (-a,a)} \varphi(x) \, \mathrm{d}x \Big)^d,$$

where φ is the PDF of the one-dimensional kernel distribution. Since $a > 1$, we get for the second term

$$\begin{aligned}
\frac{(1 - \sigma_{\min})^2}{\sigma_{\min}^2} \int_{\mathbb{R}^d \setminus (-a,a)^d} |x|^2 p_0(x) \, \mathrm{d}x &= \frac{(1 - \sigma_{\min})^2}{\sigma_{\min}^2} \sum_{i=1}^d \int_{\mathbb{R}^d \setminus (-a,a)^d} x_i^2 p_0(x) \, \mathrm{d}x_i \\
&\leq \frac{(1 - \sigma_{\min})^2}{\sigma_{\min}^2} d \Big(\int_{\mathbb{R} \setminus (-a,a)} x^2 \varphi(x) \, \mathrm{d}x \Big)^d.
\end{aligned}$$

By assumption φ decays faster than e^{-x}. Using the upper incomplete Γ-function as defined in Gabcke (1979, Satz 4.4.3), we obtain for $a > 2$

$$\begin{aligned}
\int_{\mathbb{R} \setminus (-a,a)} \varphi(x) \, \mathrm{d}x &\leq \int_{\mathbb{R} \setminus (-a,a)} x^2 \varphi(x) \, \mathrm{d}x \\
&\leq \int_{\mathbb{R} \setminus (-a,a)} x^2 e^{-x} \, \mathrm{d}x \\
&= 2 \int_a^\infty x^2 e^{-x} \, \mathrm{d}x
\end{aligned}$$

$$
\begin{aligned}
&= 2\Gamma(3,a)\\
&\leq 6e^{-a}a^2,
\end{aligned}
$$

where we accept this double usage of Γ for both the Lipschitz constant and the Gamma function, due to the universal notation of the Gamma function. Hence

$$\int_0^1 \int_{\mathbb{R}^d\setminus(-a,a)^d} |v_t^n(\psi_t^n(x)) - \tilde{v}_t(\psi_t^n(x))|^2 p_0(x)\,\mathrm{d}x\,\mathrm{d}t \leq c\frac{a^{2d+2}e^{-da}}{\sigma_{\min}^2}. \qquad \square$$

Proof of Lemma 5.23. For the supremum norm bound, observe that

$$
\begin{aligned}
|v_t^n(x)| &= \Big|\sum_{i=1}^n v_t(x|X_i^*)\frac{p_t(x|X_i^*)}{\sum_{i=1}^n p_t(x|X_i^*)}\Big|\\
&\leq \sum_{i=1}^n |v_t(x|X_i^*)|\frac{p_t(x|X_i^*)}{\sum_{i=1}^n p_t(x|X_i^*)}\\
&\leq \max_{i\in\{1,\dots,n\}} |v_t(x|X_i^*)|\\
&= \max_{i\in\{1,\dots,n\}} \frac{|(\sigma_t - t)X_i^* + x|}{1-(1-\sigma_{\min})t}\\
&\leq \max_{i\in\{1,\dots,n\}} \frac{|X_i^*| + |x|}{1-(1-\sigma_{\min})t}\\
&\leq \frac{\sqrt{d}(1+a)}{1-(1-\sigma_{\min})t}.
\end{aligned}
$$

For the Lipschitz bound, note that

$$
\begin{aligned}
\nabla_x v_t^n(x) &= \nabla_x \sum_{i=1}^n \Big(\frac{\frac{\partial\sigma_t}{\partial t}}{\sigma_t}x - \mu_t(X_i) + \frac{\partial\mu_t}{\partial t}(X_i)\Big)\frac{p_t(x|X_i)}{\sum_{j=1}^n p_t(x|X_j)} \qquad (5.27)\\
&= \nabla_x \frac{-(1-\sigma_{\min})}{\sigma_t}x - \nabla_x \frac{-(1-\sigma_{\min})t}{\sigma_t}\frac{\sum_{i=1}^n X_i p_t(x|X_i)}{\sum_{j=1}^n p_t(x|X_j)} + \nabla_x \frac{\sum_{i=1}^n X_i p_t(x|X_i)}{\sum_{j=1}^n p_t(x|X_j)}.
\end{aligned}
$$

Now we get for the partial derivative of the ℓ-st coordinate function with respect to x_k, $\ell, k \in \{1,\dots,d\}$

$$
\begin{aligned}
&\frac{\partial}{\partial x_k}\frac{\sum_{i=1}^n X_{i,\ell} p_t(x|X_i)}{\sum_{j=1}^n p_t(x|X_j)}\\
&= \frac{\big(\sum_{i=1}^n \big(-\frac{x_k - tX_{ik}}{\sigma_t^2}\big)X_{i,\ell}p_t(x|X_i)\big)\big(\sum_{j=1}^n p_t(x|X_j)\big)}{\big(\sum_{j=1}^n p_t(x|X_j)\big)^2}\\
&\quad - \frac{\big(\sum_{i=1}^n X_{i,\ell}p_t(x|X_i)\big)\big(\sum_{j=1}^n \big(-\frac{x_k - tX_{jk}}{\sigma_t^2}\big)p_t(x|X_j)\big)}{\big(\sum_{j=1}^n p_t(x|X_j)\big)^2}\\
&= \frac{t}{\sigma_t^2}\left(\frac{\sum_{i=1}^n X_{ik}X_{i\ell}p_t(x|X_i)}{\sum_{j=1}^n p_t(x|X_i)} - \frac{\big(\sum_{i=1}^n X_{ik}p_t(x|X_i)\big)\big(\sum_{i=1}^n X_{i\ell}p_t(x|X_i)\big)}{\big(\sum_{j=1}^n p_t(x|X_i)\big)^2}\right).
\end{aligned}
$$

Since $\operatorname{supp}(p^*) \subset [-1,1]^d$, we can bound

$$\frac{\partial}{\partial x_k} \frac{\sum_{i=1}^n X_{i,\ell} p_t(x|X_i)}{\sum_{j=1}^n p_t(x|X_j)} \leq \frac{2t}{\sigma_t^2}.$$

Using $t \in [0,1]$, $\sigma_{\min} \leq 1$, we get for (5.27)

$$\nabla_x v_t^n(x) \leq \frac{1}{\sigma_t} I_d + \frac{2}{\sigma_t^3} J_d,$$

where I_d denotes the $d \times d$ identity matrix, J_d denotes the $d \times d$ matrix consisting of ones and $\leq$ denotes entry wise inequality. Using the mean value theorem, we obtain for $x, y \in \mathbb{R}^d$

$$|v_t^n(x) - v_t^n(y)| \leq \left\| \frac{1}{\sigma_t} I_d + \frac{2}{\sigma_t^3} J_d \right\|_2 |x - y|.$$

As

$$\begin{aligned} \left\| \frac{1}{\sigma_t} I_d + \frac{2}{\sigma_t^3} J_d \right\|_2 &\leq \left\| \frac{1}{\sigma_t} I_d \right\|_2 + \left\| \frac{2}{\sigma_t^3} J_d \right\|_2 \\ &= \frac{1}{\sigma_t} + \frac{2d}{\sigma_t^3}, \end{aligned}$$

we get the desired bound. □

5.6.4. Proofs of Section 5.5

First, we calculate the constant leading to the equivalence of the Flow Matching objectives.

Lemma 5.24. *Let $p_t(x) > 0$ for all $x \in \mathbb{R}^d$. In the above setting, it holds that for every measurable function $\tilde{v}$ and v_t from* (5.5)

$$\begin{aligned} \mathbb{E}_{\substack{t\sim\mathcal{U}[0,1] \\ X_t\sim p_t}} \left[|v_t(X_t) - \tilde{v}_t(X_t)|^2 \right] = \mathbb{E}_{\substack{t\sim\mathcal{U}[0,1] \\ Y\sim p^* \\ X_t\sim p_t(\cdot|Y)}} \left[|\tilde{v}_t(X_t) - v_t(X_t|Y)|^2 \right] \\ - \mathbb{E}_{\substack{t\sim\mathcal{U}[0,1] \\ Y\sim p^* \\ X_t\sim p_t(\cdot|Y)}} \left[|v_t(X_t) - v_t(X_t|Y)|^2 \right]. \end{aligned}$$

Proof of Lemma 5.24. From Lipman et al. (2023), we know that there is a constant $C \in \mathbb{R}$ independent of $\tilde{v}$ such that

$$\mathbb{E}_{\substack{t\sim\mathcal{U}[0,1] \\ X_t\sim p_t}} \left[|v_t(X_t) - \tilde{v}_t(X_t)|^2 \right] = \mathbb{E}_{\substack{t\sim\mathcal{U}[0,1] \\ Y\sim p^* \\ X_t\sim p_t(\cdot|Y)}} \left[|\tilde{v}_t(X_t) - v_t(X_t|Y)|^2 \right] + C.$$

Setting $\tilde{v}_t = v_t$, we obtain

$$0 = \mathbb{E}_{\substack{t\sim\mathcal{U}[0,1] \\ Y\sim p^* \\ X_t\sim p_t(\cdot|Y)}} \left[|v_t(X_t) - v_t(X_t|Y)|^2 \right] + C \quad \Longleftrightarrow \quad C = -\mathbb{E}_{\substack{t\sim\mathcal{U}[0,1] \\ Y\sim p^* \\ X_t\sim p_t(\cdot|Y)}} \left[|v_t(X_t) - v_t(X_t|Y)|^2 \right].$$

This concludes the proof. □

Proofs of Section 5.5.1

Proof of Theorem 5.14. First we show that we can calculate the Jacobian of v explicitly. The proof and all subsequent proofs of auxiliary results are deferred to Section 5.6.4.

Lemma 5.25. *Fix $t \in [0,1]$. The Jacobian with respect to x of v_t is given by*

$$D_x v_t(x) = \frac{\sigma_t'}{\sigma_t} I_d + \big(\gamma t^{\gamma-1} - \frac{\sigma_t' t^\gamma}{\sigma_t}\big)\frac{t^\gamma}{\sigma_t^2}\operatorname{Cov}(Y^{x,t}). \tag{5.28}$$

The matrix $(B_{ij})_{i,j=1,\dots,d}$ consists of the component wise supremum norm of $D_x v_t$.

By assumption $\operatorname{Cov}(Y^{x,t})_{ji}$ is bounded for all x. For the upper bound we use that by the mean value theorem, there exists an $\xi \in \mathbb{R}^d$ such that for the i-th coordinate function v_t^j

$$|v_t^j(x) - v_t^j(y)| = \langle \nabla v_t^j(\xi), x-y\rangle \le |v_t^j(\xi)||x-y| \le \sqrt{d}|x-y| \max_{i\in\{1,\dots,d\}} \Big\|\frac{\partial}{\partial x_i} v_t^j\Big\|_\infty.$$

Then

$$|v_t(x) - v_t(y)| = \left|\begin{pmatrix} v_t^1(x) - v_t^1(y) \\ \vdots \\ v_t^d(x) - v_t^d(y)\end{pmatrix}\right| \le d|x-y| \max_{j\in\{1,\dots d\}} \max_{i\in\{1,\dots,d\}} \big\|\frac{\partial}{\partial x_i} v_t^j\big\|_\infty.$$

Therefore

$$|v_t(x) - v_t(y)| \le d|x-y| \max_{ij} \Big\|\frac{\sigma_t'}{\sigma_t}\mathbb{1}_{i=j} + \Big(\gamma t^{\gamma-1} - \frac{\sigma_t' t^\gamma}{\sigma_t}\Big)\frac{t^\gamma}{\sigma_t^2}\operatorname{Cov}(Y^{\cdot,t})_{ij}\Big\|_\infty = d|x-y| \max_{ij} B_{i,j}.$$

Hence v_t is Lipschitz continuous. For the lower bound we can use that by a Taylor expansion for $h, a \in \mathbb{R}^d$

$$v_t(a+h) = v_t(a) + D_x v_t(a)h + r(h), \tag{5.29}$$

with

$$\lim_{|h|\to 0} \frac{|r(h)|}{|h|} = 0.$$

Let the smallest Lipschitz constant of v_t be Γ_t. Using (5.29) we can conclude

$$\begin{aligned} |D_x v_t(a)h| = |v_t(a+h) - v_t(a) - r(h)| &\le |v_t(a+h) - v_t(a)| + |r(h)| \\ &\le \Gamma_t|h| + |r(h)|. \end{aligned}$$

Now

$$\|D_x v_t(a)\| = \limsup_{|h|\to 0} \frac{|D_x v_t(a)h|}{|h|} \le \limsup_{|h|\to 0} \Gamma_t + \frac{|r(h)|}{|h|} = \Gamma_t.$$

Let v_t^i be the i-th component function of v_t. Then for every $i \in \{1,\dots,d\}$ and every $j \in \{1,\dots,d\}$

$$\sup_{a\in\mathbb{R}^d} \|D_x v_t(a)\| = \sup_{a\in\mathbb{R}^d} \sup_{|w|=1} |D_x v_t(a)w| \ge \sup_{a\in\mathbb{R}^d} \sup_{|w|=1} |e_i^\top D_x v_t(a)w| = \sup_{a\in\mathbb{R}^d} \sup_{|w|=1} |\langle \nabla v_t^i(a), w\rangle|.$$

As the dual norm of the Euclidean norm is the Euclidean norm,

$$\sup_{a\in\mathbb{R}^d}\sup_{|w|=1}|\langle\nabla v_t^i(a),w\rangle| = \sup_{a\in\mathbb{R}^d}|\nabla v_t^i(a)| \geq \sup_{a\in\mathbb{R}^d}|D_x v_t(a)_{ij}|.$$

Since i,j were arbitrary and the entries of the Jacobian are of the form (5.28), we obtain the result. □

Proof of Lemma 5.15.

1. We begin by setting
$$\frac{\sigma_t'}{\sigma_t} = h_t,$$
where h_t is a continuous function on $[0,1]$. By separations of variables, all of the solutions of this ODE are of the form
$$\sigma_t = ce^{H_t},$$
where H_t is an anti-derivative of h_t and $c\in\mathbb{R}$. We use $\sigma_0 = 1$ as initial condition, which leads to
$$1 = ce^{H_0} \quad\Longleftrightarrow\quad c = \frac{1}{e^{H_0}}.$$
To assure $\sigma_1 = \sigma_{\min}$, we need to choose H such that
$$\sigma_{\min} = \frac{e^{H_1}}{e^{H_0}} \quad\Longleftrightarrow\quad H_1 - H_0 = \log(\sigma_{\min}).$$
By the mean-value-theorem there is a $t^*\in[0,1]$ such that $H_{t^*}' = h_{t^*} = \log(\sigma_{\min})$.

2. By change of variables we have that
$$\int_0^1\left|\frac{\sigma_t'}{\sigma_t}\right|\mathrm{d}t = -\int_0^1\frac{\sigma_t'}{\sigma_t}\,\mathrm{d}t = \int_{\sigma_{\min}}^1\frac{1}{u}\,\mathrm{d}u = \log(\sigma_{\min}^{-1}).$$ □

Proof of Theorem 5.17. For small t, we can use the following simple bound, which is independent of the decay of $\mathrm{Cov}(Y^{x,t})$ in t:

Lemma 5.26. *Let t^* be such that $\sigma_{t^*} = \frac{1}{\vartheta}$ for $\vartheta\in\mathbb{R}_{\geq 1}$. Grant Assumption 5.16 (III). Then*

$$\int_0^{t^*}\Gamma_t\,\mathrm{d}t \lesssim \vartheta^2(1+\log(\vartheta)).$$

For large t, we need to assume that $\mathrm{Cov}(Y^{x,t})_{ij}$ decays fast enough for $i\neq j$ and $t\to 1$ to bound the integral over all B_{ij} for $i\neq j$. Under Assumption 5.16 (I), we know that for all $i\neq j$ and all $x\in\mathbb{R}^d$

$$\begin{aligned}\int_{t^*}^1\left|\Big(\gamma t^{\gamma-1} - \frac{\sigma_t' t^\gamma}{\sigma_t}\Big)\frac{t^\gamma}{\sigma_t^2}\,\mathrm{Cov}(Y^{x,t})_{ij}\right|\mathrm{d}t &\leq (t^*)^{-2\gamma}\int_{t^*}^1\left|(\gamma t^{\gamma-1}\sigma_t - \sigma_t' t^\gamma)\right|\mathrm{d}t\\ &\leq (t^*)^{-2\gamma}\gamma\int_0^1\sigma_t - (t^*)^{-2\gamma}\int_0^1\sigma_t'\,\mathrm{d}t = (t^*)^{-2\gamma}(\gamma+1-\sigma_{\min}).\end{aligned}$$

For the bound of

$$\int_{t^*}^1 \Big|\frac{\sigma_t'}{\sigma_t} + \Big(\gamma t^{\gamma-1} - \frac{\sigma_t' t^\gamma}{\sigma_t}\Big)\frac{t^\gamma}{\sigma_t^2}\operatorname{Var}(Y_i^{x,t})\Big|\,\mathrm{d}t$$

we need to use Assumption 5.16 (II). Inserting the expression for the variance $\operatorname{Var}(Y_i^{x,t})$ we obtain

$$\begin{aligned}
&\int_{t^*}^1 \Big|\frac{\sigma_t'}{\sigma_t} + \Big(\gamma t^{\gamma-1} - \frac{\sigma_t' t^\gamma}{\sigma_t}\Big)\frac{t^\gamma}{\sigma_t^2}\operatorname{Var}(Y_i^{x,t})\Big|\,\mathrm{d}t \\
&\quad= \int_{t^*}^1 \Big|\frac{\sigma_t'}{\sigma_t} + \Big(\gamma t^{\gamma-1} - \frac{\sigma_t' t^\gamma}{\sigma_t}\Big)\frac{t^\gamma}{\sigma_t^2}\Big(\frac{\sigma_t}{t^\gamma}\Big)^2\Big(1 + O\Big(\Big(\frac{\sigma_t}{t^\gamma}\Big)^{\frac{1}{\kappa}}\Big)\Big)\Big|\,\mathrm{d}t \\
&\quad\le \int_{t^*}^1 \Big|\frac{\sigma_t'}{\sigma_t} + \Big(\gamma t^{\gamma-1} - \frac{\sigma_t' t^\gamma}{\sigma_t}\Big)\frac{t^\gamma}{\sigma_t^2}\Big(\frac{\sigma_t}{t^\gamma}\Big)^2\Big|\,\mathrm{d}t + \int_{t^*}^1 \Big|\Big(\gamma t^{\gamma-1} - \frac{\sigma_t' t^\gamma}{\sigma_t}\Big)\frac{t^\gamma}{\sigma_t^2}\Big(\frac{\sigma_t}{t^\gamma}\Big)^2 O\Big(\Big(\frac{\sigma_t}{t^\gamma}\Big)^{\frac{1}{\kappa}}\Big)\Big|\,\mathrm{d}t.
\end{aligned}$$

The first term simplifies to

$$\int_{t^*}^1 \Big|\frac{\sigma_t'}{\sigma_t} + \Big(\gamma t^{\gamma-1} - \frac{\sigma_t' t^\gamma}{\sigma_t}\Big)\frac{t^\gamma}{\sigma_t^2}\Big(\frac{\sigma_t}{t^\gamma}\Big)^2\Big|\,\mathrm{d}t = \int_{t^*}^1 |\gamma t^{-1}|\,\mathrm{d}t = -\gamma\log(t^*).$$

For the second term we obtain

$$\begin{aligned}
\int_{t^*}^1 \Big|\Big(\gamma t^{\gamma-1} - \frac{\sigma_t' t^\gamma}{\sigma_t}\Big)\frac{t^\gamma}{\sigma_t^2}\Big(\frac{\sigma_t}{t^\gamma}\Big)^2 O\Big(\Big(\frac{\sigma_t}{t^\gamma}\Big)^{\frac{1}{\kappa}}\Big)\Big|\,\mathrm{d}t &\le \int_{t^*}^1 \Big|\Big(\gamma t^{\gamma-1} - \frac{\sigma_t' t^\gamma}{\sigma_t}\Big)\Big|\frac{1}{t^\gamma} O\Big(\Big(\frac{\sigma_t}{t^\gamma}\Big)^{\frac{1}{\kappa}}\Big)\Big|\,\mathrm{d}t \\
&\lesssim \int_{t^*}^1 \gamma t^{-1-\frac{\kappa}{\gamma}}\sigma_t^{\frac{1}{\kappa}}\,\mathrm{d}t + \int_{t^*}^1 \frac{\sigma_t'}{\sigma_t}\frac{\sigma_t^{\frac{1}{\kappa}}}{t^{\frac{\gamma}{\kappa}}}\,\mathrm{d}t \\
&\le \gamma\int_{t^*}^1 t^{-1-\frac{\kappa}{\gamma}}\,\mathrm{d}t + (t^*)^{-\frac{\gamma}{\kappa}}\int_{t^*}^1 \frac{\sigma_t'}{\sigma_t}\sigma_t^{\frac{1}{\kappa}}\,\mathrm{d}t \\
&= \frac{\gamma^2}{\kappa}((t^*)^{-\frac{\kappa}{\gamma}} - 1) + (t^*)^{-\frac{\gamma}{\kappa}}\kappa(\sigma_{\min}^{\frac{1}{\kappa}} - \sigma_{t^*}^{\frac{1}{\kappa}}).
\end{aligned}$$

Using that $\sigma_{\min}^{\frac{1}{\kappa}} - \sigma_{t^*}^{\frac{1}{\kappa}} \le 1$ yields the following bound on the integral over the Lipschitz constant

$$\int_0^1 \Gamma_t\,\mathrm{d}t \lesssim \vartheta^2(1 + \log(\vartheta)) + (t^*)^{-2\gamma}(\gamma+1) + \frac{\gamma^2}{\kappa}((t^*)^{-\frac{\kappa}{\gamma}} - 1) + (t^*)^{-\frac{\gamma}{\kappa}}\kappa. \qquad \square$$

Proof of Theorem 5.18.

Property (III):

For the uniform bound, we use that the distribution with density

$$q \propto \exp\Big(-\frac{|x-ty|^2}{2\sigma_t^2} - \frac{|y|^2}{2}\Big) = \exp\Big(-\frac{1}{2}\Big(1 + \frac{t^2}{\sigma_t^2}\Big)\Big(|y|^2 - \Big\langle \frac{t}{\sigma_t^2}\frac{x}{1+\frac{t^2}{\sigma_t^2}}, y\Big\rangle\Big)\Big) \tag{5.30}$$

is a Gaussian distribution with variance $\big(1 + \frac{t^2}{\sigma_t^2}\big)^{-1} I_d$. By Theorem 2.13, the density defined above satisfies the Poincaré inequality with constant $\big(1 + \frac{t^2}{\sigma_t^2}\big)^{-1}$. Using the Holley-Strooke perturbation principle Holley & Stroock (1987), in the form of Ledoux (2001, Lemma 1.2), we can bound the Poincaré constant ρ of the perturbed Gaussian via

$$\rho \le e^{4L}\Big(1 + \frac{t^2}{\sigma_t^2}\Big)^{-1}.$$

Thus

$$\mathrm{Var}(Y_i^{x,t}) \le \Big(e^{-4L}\Big(1+\frac{t^2}{\sigma_t^2}\Big)\Big)^{-1} \le \frac{e^{4L}}{1+\frac{t^2}{\sigma_t^2}} \le e^{4L},$$

with L from (5.20). Using

$$\mathrm{Cov}(Y_i^{x,t}, Y_j^{x,t}) \le \sqrt{\mathrm{Var}(Y_i^{x,t})\,\mathrm{Var}(Y_j^{x,t})},$$

we conclude that for all $t \in [0,1]$

$$\mathrm{Cov}(Y_i^{x,t}, Y_j^{x,t}) \le e^{4L}.$$

Hence we can set $C = e^{4L}$.

Property (II):

For the variance, we use that

$$\mathrm{Var}(Y_i^{x,t}) = \mathbb{E}[(Y_i^{x,t})^2] - \mathbb{E}[Y_i^{x,t}]^2.$$

Let φ denote the density of $\mathcal{N}(0, I_d)$. Further, let $t > 0$. Then

$$\mathbb{E}[Y_i^{x,t}] = \frac{\int y_i \varphi\big(\frac{x-ty}{\sigma_t}\big) p^*(y)\,\mathrm{d}y}{\int \varphi\big(\frac{x-ty}{\sigma_t}\big) p^*(y)\,\mathrm{d}y} = \frac{\int \big(\frac{x_i - z_i\sigma_t}{t}\big)\varphi(z) p^*\big(\frac{x-z\sigma_t}{t}\big)\,\mathrm{d}y}{\int \varphi(z) p^*\big(\frac{x-z\sigma_t}{t}\big)\,\mathrm{d}y} = \frac{x_i}{t} - \frac{\sigma_t}{t}\frac{\int z_i \varphi(z) p^*\big(\frac{x-z\sigma_t}{t}\big)\,\mathrm{d}y}{\int \varphi(z) p^*\big(\frac{x-z\sigma_t}{t}\big)\,\mathrm{d}y}.$$

Similarly

$$\begin{aligned}\mathbb{E}[\big(Y_i^{x,t}\big)^2] &= \frac{\int \big(\frac{x_i - z_i\sigma_t}{t}\big)^2 \varphi(z) p^*\big(\frac{x-z\sigma_t}{t}\big)\,\mathrm{d}y}{\int \varphi(z) p^*\big(\frac{x-z\sigma_t}{t}\big)\,\mathrm{d}y} \\ &= \big(\frac{x_i}{t}\big)^2 - \frac{2x_i\sigma_t}{t^2}\frac{\int z_i \varphi(z) p^*\big(\frac{x-z\sigma_t}{t}\big)\,\mathrm{d}y}{\int \varphi(z) p^*\big(\frac{x-z\sigma_t}{t}\big)\,\mathrm{d}y} + \big(\frac{\sigma_t}{t}\big)^2 \frac{\int z_i^2 \varphi(z) p^*\big(\frac{x-z\sigma_t}{t}\big)\,\mathrm{d}y}{\int \varphi(z) p^*\big(\frac{x-z\sigma_t}{t}\big)\,\mathrm{d}y}.\end{aligned}$$

Now we define

$$A(z_i) \coloneqq \frac{\int z_i \varphi(z) p^*\big(\frac{x-z\sigma_t}{t}\big)\,\mathrm{d}z}{\int \varphi(z) p^*\big(\frac{x-z\sigma_t}{t}\big)\,\mathrm{d}z}, \quad A(z_i^2) \coloneqq \frac{\int z_i^2 \varphi(z) p^*\big(\frac{x-z\sigma_t}{t}\big)\,\mathrm{d}z}{\int \varphi(z) p^*\big(\frac{x-z\sigma_t}{t}\big)\,\mathrm{d}z}.$$

We obtain for the variances

$$\begin{aligned}\mathbb{E}[(Y_i^{x,t})^2] - \mathbb{E}[Y_i^{x,t}]^2 &= \Big(\frac{x_i}{t}\Big)^2 - \frac{2x_i\sigma_t}{t^2}A(z_i) + \Big(\frac{\sigma_t}{t}\Big)^2 A(z_i^2) - \Big(\frac{x_i}{t} - \frac{\sigma_t}{t}A(z_i)\Big)^2 \\ &= \Big(\frac{\sigma_t}{t}\Big)^2 \Big(A(z_i^2) - A(z_i)^2\Big). \end{aligned} \tag{5.31}$$

Hence we need to bound the component variances of a random variable Z with density

$$\begin{aligned}p_Z(z) &= \frac{\varphi(z) p^*\big(\frac{x-z\sigma_t}{t}\big)}{\int \varphi(z) p^*\big(\frac{x-z\sigma_t}{t}\big)\,\mathrm{d}z} \\ &= \frac{\exp\big(-\frac{|z|^2}{2} - \frac{|x-z\sigma_t|^2}{2t^2} - a\big(\frac{x-z\sigma_t}{t}\big)\big)}{\int \exp\big(-\frac{|z|^2}{2} - \frac{|x-z\sigma_t|^2}{2t^2} - a\big(\frac{x-z\sigma_t}{t}\big)\big)\,\mathrm{d}z}\end{aligned}$$

$$= \frac{\exp\big(-\frac{1}{2}\big(1+\frac{\sigma_t^2}{t^2}\big)\big(|z|^2 - 2\langle z, x\frac{\sigma_t}{t^2(1+(\frac{\sigma_t}{t})^2)}\rangle\big) - a(\frac{x-z\sigma_t}{t})\big)}{\int \exp\big(-\frac{1}{2}\big(1+\frac{\sigma_t^2}{t^2}\big)\big(|z|^2 - 2\langle z, x\frac{\sigma_t}{t^2(1+(\frac{\sigma_t}{t})^2)}\rangle\big) - a(\frac{x-z\sigma_t}{t})\big)\,\mathrm{d}z}. \tag{5.32}$$

First we bound the influence of the perturbation function a on the expected value. To do so, we note that

$$\begin{aligned}\mathbb{E}_{p_Z}[\partial_{z_i}\log(p(Z))] &= \int \partial_{z_i} p(z)\,\mathrm{d}z \\ &= \int ... \int_{z_i\in\mathbb{R}} \partial_{z_i} p(z)\,\mathrm{d}z_i\,\mathrm{d}(z_1,...,z_{i-1},z_{z+1},...,z_d) \\ &= 0.\end{aligned}$$

This implies for (5.32) that

$$\begin{aligned} & 0 = -\mathbb{E}\Big[\Big(1+\frac{\sigma_t^2}{t^2}\Big)\Big(Z_i - x_i\frac{\sigma_t}{t^2(1+(\frac{\sigma_t}{t})^2)}\Big) - \frac{\sigma_t}{t}\partial_{z_i}a\Big(\frac{x-Z\sigma_t}{t}\Big)\Big] \\ \iff\quad & \Big(1+\frac{\sigma_t^2}{t^2}\Big)\Big(\mathbb{E}[Z_i] - x_i\frac{\sigma_t}{t^2(1+(\frac{\sigma_t}{t})^2)}\Big) = \mathbb{E}\Big[\frac{\sigma_t}{t}\partial_{z_i}a\Big(\frac{x-Z\sigma_t}{t}\Big)\Big] \\ \iff\quad & \Big|\mathbb{E}[Z_i] - x_i\frac{\sigma_t}{t^2(1+(\frac{\sigma_t}{t})^2)}\Big| = \Big|\frac{\frac{\sigma_t}{t}\mathbb{E}\Big[\partial_{z_i}a(\frac{x-Z\sigma_t}{t})\Big]}{\big(1+\frac{\sigma_t^2}{t^2}\big)}\Big|. \end{aligned}$$

Hence

$$\Big|\mathbb{E}[Z_i] - x_i\frac{\sigma_t}{t^2(1+(\frac{\sigma_t}{t})^2)}\Big| \le \frac{\sigma_t}{t}\frac{L}{\big(1+\frac{\sigma_t^2}{t^2}\big)}, \tag{5.33}$$

where L is from (5.20). From (5.32) we can see that this bounds the influence of the perturbation function a on the expected value.
Now we bound the variance. To do so, we first note that

$$\mathbb{E}_{p_Z}[(\partial_{z_i}\log(p(Z)))^2] = \int \frac{(\partial_{z_i}p_Z(z))^2}{p_Z(z)}\,\mathrm{d}z. \tag{5.34}$$

As

$$\partial_{z_i}\log(p_Z(z)) = \frac{\partial_{z_i}p_Z(z)}{p_Z(z)}, \quad \partial^2_{z_i}\log(p_Z(z)) = \frac{\partial^2_{z_i}p_Z(z)}{p_Z(z)} - \frac{(\partial_{z_i}p_Z(z))^2}{(p_Z(z))^2},$$

we have for (5.34)

$$\mathbb{E}_{p_Z}[(\partial_{z_i}\log(p(Z)))^2] = \mathbb{E}_{p_Z}\Big[\frac{\partial^2_{z_i}p_Z(z)}{p_Z(z)}\Big] - \mathbb{E}_{p_Z}\big[\partial^2_{z_i}\log(p_Z(Z))\big].$$

Since $\partial_{z_i}p_Z(z) \to 0$ for $|z| \to \infty$, we conclude

$$\mathbb{E}_{p_Z}[(\partial_{z_i}\log(p(Z)))^2] = -\mathbb{E}_{p_Z}[\partial^2_{z_i}\log(p_Z(Z))]. \tag{5.35}$$

For the right hand side, we obtain

$$-\mathbb{E}_{p_Z}[\partial^2_{z_i}\log(p_Z(Z))] = \Big(1+\Big(\frac{\sigma_t}{t}\Big)^2\Big) + \Big(\frac{\sigma_t}{t}\Big)^2\mathbb{E}_{p_Z}\Big[\partial^2_{z_i}a\Big(\frac{x-\sigma_t z}{t}\Big)\Big].$$

The left side of (5.35) can be rewritten as

$$\begin{aligned}
&\mathbb{E}_{p_Z}[(\partial_{z_i}\log(p(Z)))^2] = \mathbb{E}_{p_Z}\Big[\Big(\Big(1+\frac{\sigma_t^2}{t^2}\Big)\Big(Z_i - x_i\frac{\sigma_t}{t^2(1+(\frac{\sigma_t}{t})^2)}\Big) + \frac{\sigma_t}{t}\partial_{z_i}a\Big(\frac{x-\sigma_t Z}{t}\Big)\Big)^2\Big] \\
&= \mathbb{E}_{p_Z}\Big[\Big(\Big(1+\frac{\sigma_t^2}{t^2}\Big)\Big(Z_i - x_i\frac{\sigma_t}{t^2(1+(\frac{\sigma_t}{t})^2)}\Big)\Big)^2\Big] \\
&\quad + 2\mathbb{E}_{p_Z}\Big[\Big(1+\frac{\sigma_t^2}{t^2}\Big)\Big(Z_i - x_i\frac{\sigma_t}{t^2(1+(\frac{\sigma_t}{t})^2)}\Big)\frac{\sigma_t}{t}\partial_{z_i}a\Big(\frac{x-\sigma_t Z}{t}\Big)\Big] \\
&\quad + \Big(\frac{\sigma_t}{t}\Big)^2\mathbb{E}_{p_Z}\Big[\Big(\partial_{z_i}a\Big(\frac{x-\sigma_t Z}{t}\Big)\Big)^2\Big].
\end{aligned}$$

Then

$$\begin{aligned}
&\mathbb{E}_{p_Z}\Big[\Big(\Big(1+\frac{\sigma_t^2}{t^2}\Big)\Big(Z_i - x_i\frac{\sigma_t}{t^2(1+(\frac{\sigma_t}{t})^2)}\Big)\Big)^2\Big] \\
&= \mathbb{E}_{p_Z}\Big[\Big(\Big(1+\frac{\sigma_t^2}{t^2}\Big)\Big(Z_i - \mathbb{E}[Z_i] + \mathbb{E}[Z_i] - x_i\frac{\sigma_t}{t^2(1+(\frac{\sigma_t}{t})^2)}\Big)\Big)^2\Big] \\
&= \Big(1+\frac{\sigma_t^2}{t^2}\Big)^2\mathbb{E}_{p_Z}\big[(Z_i-\mathbb{E}[Z_i])^2\big] + 2\Big(1+\frac{\sigma_t^2}{t^2}\Big)^2\Big(\mathbb{E}[Z_i] - x_i\frac{\sigma_t}{t^2(1+(\frac{\sigma_t}{t})^2)}\Big)\mathbb{E}_{p_Z}\big[(Z_i-\mathbb{E}[Z_i])\big] \\
&\quad + \Big(1+\frac{\sigma_t^2}{t^2}\Big)^2\Big(\mathbb{E}[Z_i] - x_i\frac{\sigma_t}{t^2(1+(\frac{\sigma_t}{t})^2)}\Big)^2 \\
&= \Big(1+\frac{\sigma_t^2}{t^2}\Big)^2\Big(\mathbb{E}_{p_Z}\Big[(Z_i-\mathbb{E}[Z_i])^2\Big] + \Big(\mathbb{E}[Z_i] - x_i\frac{\sigma_t}{t^2(1+(\frac{\sigma_t}{t})^2)}\Big)^2\Big).
\end{aligned}$$

Additionally

$$\begin{aligned}
&\mathbb{E}_{p_Z}\Big[\Big(1+\frac{\sigma_t^2}{t^2}\Big)(Z_i - x_i\frac{\sigma_t}{t^2(1+(\frac{\sigma_t}{t})^2)})\frac{\sigma_t}{t}\partial_{z_i}a\Big(\frac{x-\sigma_t Z}{t}\Big)\Big] \\
&= \mathbb{E}_{p_Z}\Big[\Big(1+\frac{\sigma_t^2}{t^2}\Big)\Big(Z_i - \mathbb{E}[Z_i] + \mathbb{E}[Z_i] - x_i\frac{\sigma_t}{t^2(1+(\frac{\sigma_t}{t})^2)}\Big)\frac{\sigma_t}{t}\partial_{z_i}a\Big(\frac{x-\sigma_t Z}{t}\Big)\Big] \\
&= \Big(1+\frac{\sigma_t^2}{t^2}\Big)\frac{\sigma_t}{t}\mathbb{E}_{p_Z}\Big[(Z_i-\mathbb{E}[Z_i])\partial_{z_i}a\Big(\frac{x-\sigma_t Z}{t}\Big)\Big] \\
&\quad + \Big(1+\frac{\sigma_t^2}{t^2}\Big)\frac{\sigma_t}{t}\Big(\mathbb{E}[Z_i] - x_i\frac{\sigma_t}{t^2(1+(\frac{\sigma_t}{t})^2)}\Big)\mathbb{E}_{p_Z}\Big[\partial_{z_i}a\Big(\frac{x-\sigma_t Z}{t}\Big)\Big].
\end{aligned}$$

Hence

$$\begin{aligned}
&\Big(1+\frac{\sigma_t^2}{t^2}\Big)^2\mathbb{E}_{p_Z}\Big[\Big(Z_i-\mathbb{E}[Z_i]\Big)^2\Big] \\
&\qquad = \Big(1+\Big(\frac{\sigma_t}{t}\Big)^2\Big) + \Big(\frac{\sigma_t}{t}\Big)^2\mathbb{E}_{p_Z}\Big[\partial_{z_i}^2 a\Big(\frac{x-\sigma_t z}{t}\Big)\Big] \\
&\qquad\quad - \Big(\frac{\sigma_t}{t}\Big)^2\mathbb{E}_{p_Z}\Big[\Big(\partial_{z_i}a\Big(\frac{x-\sigma_t Z}{t}\Big)\Big)^2\Big] \\
&\qquad\quad - \Big(1+\frac{\sigma_t^2}{t^2}\Big)^2\Big(\mathbb{E}[Z_i] - x_i\frac{\sigma_t}{t^2\big(1+\big(\frac{\sigma_t}{t}\big)^2\big)}\Big)^2 \\
&\qquad\quad - 2\Big(1+\frac{\sigma_t^2}{t^2}\Big)\frac{\sigma_t}{t}\mathbb{E}_{p_Z}\Big[(Z_i-\mathbb{E}[Z_i])\partial_{z_i}a\Big(\frac{x-\sigma_t Z}{t}\Big)\Big]
\end{aligned}$$

$$- 2\Big(1+\frac{\sigma_t^2}{t^2}\Big)\frac{\sigma_t}{t}\Big(\mathbb{E}[Z_i] - x_i\frac{\sigma_t}{t^2(1+(\frac{\sigma_t}{t})^2)}\Big)\mathbb{E}_{p_Z}\Big[\partial_{z_i}a\Big(\frac{x-\sigma_t Z}{t}\Big)\Big].$$

Combining this with (5.33) and the fact that a has bounded derivatives, we obtain the following upper and lower bounds:

$$\Big(1+\frac{\sigma_t^2}{t^2}\Big)^2\mathbb{E}_{p_Z}\Big[\Big(Z_i-\mathbb{E}[Z_i]\Big)^2\Big] \leq \Big(1+\Big(\frac{\sigma_t}{t}\Big)^2\Big) + \Big(\frac{\sigma_t}{t}\Big)^2 L + 2\Big(1+\frac{\sigma_t^2}{t^2}\Big)\frac{\sigma_t}{t}L\sqrt{\mathbb{E}_{p_Z}\big[(Z_i-\mathbb{E}[Z_i])^2\big]} + 2\Big(1+\frac{\sigma_t^2}{t^2}\Big)\frac{\sigma_t^2}{t^2}\frac{L^2}{\Big(1+\frac{\sigma_t^2}{t^2}\Big)},$$

and

$$\begin{aligned}\Big(1+\frac{\sigma_t^2}{t^2}\Big)^2\mathbb{E}_{p_Z}\Big[\Big(Z_i-\mathbb{E}[Z_i]\Big)^2\Big] \geq{}& \Big(1+\Big(\frac{\sigma_t}{t}\Big)^2\Big) - \Big(\frac{\sigma_t}{t}\Big)^2 L - \Big(\frac{\sigma_t}{t}\Big)^2 L^2 \\ &- \Big(1+\frac{\sigma_t^2}{t^2}\Big)^2\Big(\frac{\sigma_t}{t}\frac{L}{\Big(1+\frac{\sigma_t^2}{t^2}\Big)}\Big)^2 \\ &- 2\Big(1+\frac{\sigma_t^2}{t^2}\Big)\frac{\sigma_t}{t}L\sqrt{\mathbb{E}_{p_Z}\big[(Z_i-\mathbb{E}[Z_i])^2\big]} \\ &- 2\Big(1+\frac{\sigma_t^2}{t^2}\Big)\frac{\sigma_t^2}{t^2}\frac{L^2}{\big(1+\frac{\sigma_t^2}{t^2}\big)}.\end{aligned}$$

With the same reasoning via the Poincaré constant of a Gaussian and the Holley-Stroock perturbation principle, we conclude that

$$\mathbb{E}_{p_Z}\big[(Z_i-\mathbb{E}[Z_i])^2\big] \leq e^{4L}.$$

As $\big(1+\frac{\sigma_t^2}{t^2}\big)^2 > 0$ and

$$\frac{1}{1+\frac{\sigma_t^2}{t^2}} = 1 - \frac{\frac{\sigma_t^2}{t^2}}{1+\frac{\sigma_t^2}{t^2}},$$

dividing the first term in the upper and lower bound loosens the bound further. Thus we obtain

$$\mathbb{E}_{p_Z}\big[\big(Z_i-\mathbb{E}[Z_i]\big)^2\big] = 1 + O\Big(\frac{\sigma_t}{t}\Big).$$

Multiplying with $\frac{\sigma_t^2}{t}$ yields the result for the variances via (5.31), we arrive at

$$\mathrm{Var}(Y_i^{x,t}) = \big(\frac{\sigma_t}{t}\big)^2\big(1+O\Big(\frac{\sigma_t}{t}\Big)\big).$$

Property (I):

For the covariances, we need to find a stricter upper bound. We are going to use Menz (2014, Theorem 2.3) applied to the component functions $f_i\colon \mathbb{R}^d \to \mathbb{R},\ f_i(y) = y_i$. First we verify that the assumptions are fulfilled in our setting. By construction, the distribution of $Y_i^{x,t}$ is a bounded perturbation of a Gaussian. Hence Assumption 2.2 in Menz (2014) is satisfied.

For the Poincaré constant of the i-th conditional measure, e.g. the Poincaré constant of the

distribution with the density

$$p(y_i|y_1,...,y_{i-1},y_{i+1},...,y_d) \propto \exp\Big(-\frac{(x_i-ty_i)^2}{2\sigma_t^2}-\frac{y_i^2}{2}-a(y_i|y_1,...,y_{i-1},y_{i+1},...,y_d)\Big),$$

where $a(\cdot|y_1,...,y_{i-1},y_{i+1},...,y_d) \coloneqq a(y_1,...,y_{i-1},\cdot,y_{i+1},...,y_d)$ and $i \in \{1,...,d\}$, we can use the same arguments as in the proof of property (III) to obtain

$$\tilde{\rho}_i^t \leq e^{4L}\Big(1+\frac{t^2}{\sigma_t^2}\Big)^{-1}.$$

Note that $a(\cdot|y_1,...,y_{i-1},y_{i+1},...,y_d)$ is still bounded. Furthermore, we know that the off-diagonal entries of the Hessian of the log-density are bounded by L. To profit from easier notation later, we define the matrix

$$A_t = (A_{t,ij})_{i,j=1,...,d}, \quad \text{where} \quad A_{t,ij} \coloneqq \begin{cases} \rho_i^t - e^{-4L}\frac{t^2}{\sigma_t^2}, & i=j, \\ -L, & i \neq j. \end{cases}$$

Here $\rho_i^t = (\tilde{\rho}_i^t)^{-1}$. This inversion provides accordance with the notation in Menz (2014), who defines the Poincaré constant as the inverse of the constant in Definition 2.12. We need to find a bound on t^* such that the matrix $A_t + e^{-4L}\frac{t^2}{\sigma_t^2}I_d$ is positive definite. As A_t is symmetric by construction, we know that the eigenvalues of $A_t + e^{-4L}\frac{t^2}{\sigma_t^2}I_d$ are real numbers. By Gerschgorin's theorem (Gerschgorin, 1931, Satz 2), we know that the eigenvalues of $A_t + e^{-4L}\frac{t^2}{\sigma_t^2}I_d$ are in the following union of intervals:

$$D = \bigcup_{i=1}^{d}\Big[\rho_i^t - L(d-1), \rho_i^t + L(d-1)\Big]. \tag{5.36}$$

Inserting the lower bound of ρ_i^t, we choose t^* such that for all $t < t^*$

$$e^{-4L}\Big(1+\frac{t^2}{\sigma_t^2}\Big) > L(d-1) \quad \Longleftarrow \quad \frac{\sigma_t}{t} < \frac{1}{\sqrt{e^{4L}L(d-1)-1}}. \tag{5.37}$$

For this choice of t^*, we conclude that the matrix $A_{t,ij} + e^{-4L}\frac{t^2}{\sigma_t^2}I_d$ is positive definite. Thus all of the assumptions in Menz (2014, Theorem 2.3) are satisfied.
We conclude that

$$|\operatorname{Cov}(Y^{x,t})_{ij}| \leq \Big(\Big(e^{-4L}\frac{t^2}{\sigma_t^2}I_d + A_t\Big)^{-1}\Big)_{ij}.$$

Now

$$\Big(e^{-4L}\frac{t^2}{\sigma_t^2}I_d + A_t\Big)^{-1} = e^{4L}\frac{\sigma_t^2}{t^2}\Big(I_d + e^{4L}\frac{\sigma_t^2}{t^2}A_t\Big)^{-1}. \tag{5.38}$$

As A_t is a symmetric matrix, the spectral norm is the absolute value of the largest eigenvalue. Using Gerschgorin's theorem (Gerschgorin, 1931, Satz II) again, we know that the eigenvalues

of A_t are in the following union of intervals:

$$D = \bigcup_{i=1}^{d} \Big[\rho_i^t - e^{-4L}\frac{t^2}{\sigma_t^2} - L(d-1), \rho_i^t - e^{-4L}\frac{t^2}{\sigma_t^2} + L(d-1)\Big]. \tag{5.39}$$

As $\rho_i^t - e^{-4L}\frac{t^2}{\sigma_t^2} = e^{-4L}$ we conclude that the largest eigenvalue is bounded by

$$\lambda_{\max}(A_t) \leq e^{-4L} + L(d+1).$$

We therefore deduce that

$$\Big\|e^{4L}\frac{\sigma_t^2}{t^2}A_t\Big\| = e^{4L}\frac{\sigma_t^2}{t^2}\Big\|A_t\Big\| \leq \frac{\sigma_t^2}{t^2}(1 + e^{4L}L(d-1)).$$

If we choose t^* such that for $t > t^*$

$$\frac{\sigma_t^2}{t^2}(1 + e^{4L}L(d-1)) \leq \frac{\sigma_t}{t} < 1 \quad \Longleftrightarrow \quad \frac{\sigma_t}{t} \leq \frac{1}{1 + e^{4L}L(d-1)},$$

we can use a Neumann series. This gives

$$\Big(I_d + e^{4L}\frac{\sigma_t^2}{t^2}A_t\Big)^{-1} = \sum_{k=0}^{\infty}(-1)^k e^{4Lk}\frac{\sigma_t^{2k}}{t^{2k}}A_t^k = I_d - e^{4L}\frac{\sigma_t^2}{t^2}A_t + \sum_{k=2}^{\infty}(-1)^k e^{4Lk}\frac{\sigma_t^{2k}}{t^{2k}}A_t^k.$$

Now for the ij-th element, we obtain

$$\Big(I_d + e^{4L}\frac{\sigma_t^2}{t^2}A_t\Big)^{-1}_{ij} \leq \Big(\mathbb{1}_{i=j} + e^{4L}\frac{\sigma_t^2}{t^2}A_{t,ij} + \Big\|\sum_{k=2}^{\infty}(-1)^k e^{4Lk}\frac{\sigma_t^{2k}}{t^{2k}}A_t^k\Big\|\Big).$$

Then we can bound

$$\Big\|\sum_{k=2}^{\infty}(-1)^k e^{4Lk}\frac{\sigma_t^{2k}}{t^{2k}}A_t^k\Big\| \leq \sum_{k=2}^{\infty}\Big\|e^{4Lk}\frac{\sigma_t^{2k}}{t^{2k}}A_t^k\Big\| \leq \sum_{k=2}^{\infty}\Big\|e^{4L}\frac{\sigma_t^2}{t^2}A_t\Big\|^k \leq \sum_{k=2}^{\infty}\big(\frac{\sigma_t}{t}\big)^k$$

Using the convergence of the geometric series, we get that

$$\sum_{k=2}^{\infty}\big(\frac{\sigma_t}{t}\big)^k = \frac{\frac{\sigma_t^2}{t^2}}{1 - \frac{\sigma_t}{t}} \leq \frac{\sigma_t^2}{t^2}.$$

Inserting everything into (5.38), we obtain

$$\Big(I_d + \frac{\sigma_t^2}{t^2}A_t\Big)^{-1}_{ij} \lesssim \frac{\sigma_t^2}{t^2}\Big(\mathbb{1}_{i=j} + \frac{\sigma_t^2}{t^2}\Big).$$

If we choose $t^* \geq \frac{1}{2}$, we know that $\frac{\sigma_t^2}{t^2} \leq 2\sigma_t^2$. This gives the smaller bound on the covariances.

□

Proofs of Section 5.5.2

Proof of Theorem 5.20. To validate the bound on $|v_t^j|_\infty$, we insert $\log(\sigma_{\min}) \asymp \log(n)$ a bit later. We start by defining $A \coloneqq [-\log(n), \log(n)]^d$ and

$$g\colon \mathcal{M} \times \mathbb{R}^d \to \mathbb{R}^d, \quad g(v, y) \coloneqq \int_0^1 \int |v_t(x) - v_t(x|y)|^2 p_t(x|y)\, \mathrm{d}x\, \mathrm{d}t. \tag{5.40}$$

Then we use Lemma 5.24,

$$\mathbb{E}_{\substack{t\sim\mathcal{U}[0,1]\\ X_t\sim p_t}}\left[|v_t(X_t) - \hat{v}_t(X_t)|^2\right] = \mathbb{E}_{\substack{t\sim\mathcal{U}[0,1]\\ Y\sim p^*\\ X_t\sim p_t(\cdot|Y)}}\left[\,|\hat{v}_t(X_t) - v_t\,(X_t|Y)|^2\,\right] - C, \tag{5.41}$$

where

$$C \coloneqq \mathbb{E}_{\substack{t\sim\mathcal{U}[0,1]\\ Y\sim p^*\\ X_t\sim p_t(\cdot|Y)}}\left[|v_t(X_t) - v_t(X_t|Y)|^2\right]. \tag{5.42}$$

We can split the integral

$$\begin{aligned}\mathbb{E}_{\substack{t\sim\mathcal{U}[0,1]\\ Y\sim p^*\\ X_t\sim p_t(\cdot|Y)}}\left[\,|\hat{v}_t(X_t) - v_t\,(X_t|Y)|^2\,\right] &= \int_A \int_0^1 \int |\hat{v}_t(x) - v_t(x|y)|^2 p_t(x|y)\, \mathrm{d}x\, \mathrm{d}t\, p^*(y)\, \mathrm{d}y \\ &\quad + \int_{\mathbb{R}^d\setminus A} \int_0^1 \int |\hat{v}_t(x) - v_t(x|y)|^2 p_t(x|y)\, \mathrm{d}x\, \mathrm{d}t\, p^*(y)\, \mathrm{d}y.\end{aligned}$$

Note that

$$\int_A \int_0^1 \int |\hat{v}_t(x) - v_t(x|y)|^2 p_t(x|y)\, \mathrm{d}x\, \mathrm{d}t\, p^*(x)\, \mathrm{d}y = \mathbb{E}_{Y\sim\mathbb{P}^*}[g(\hat{v}, Y)\mathbb{1}_{Y\in A}].$$

We have for every $x \in \mathbb{R}^d$ and $j \in \{1, ..., d\}$ and the t-th component function v_t^j of v_t

$$\begin{aligned}|v_t^j(x)| &= \Big| \int v_t^j(x|y) \frac{p_t(x|y)}{\int p_t(x|z)p^*(z)\, \mathrm{d}z} p^*(y)\, \mathrm{d}y \Big| \\ &= \Big| \int \Big(\frac{\sigma_t'}{\sigma_t} x_j + \Big(1 - \frac{\sigma_t'}{\sigma_t} t\Big) y_j \Big) \frac{p_t(x|y)}{\int p_t(x|z)p^*(z)\, \mathrm{d}z} p^*(y)\, \mathrm{d}y \Big| \\ &\le \Big| \frac{\sigma_t'}{\sigma_t} \Big| |x_j| + \Big| \Big(1 - \frac{\sigma_t'}{\sigma_t} t\Big) \Big| \frac{\int |y_j| \exp\big(-\frac{|x-ty|^2}{2\sigma_t^2} - \frac{|y|^2}{2} - a(y)\big)\, \mathrm{d}y}{\int \exp\big(-\frac{|x-ty|^2}{2\sigma_t^2} - \frac{|y|^2}{2} - a(y)\big)\, \mathrm{d}y} \\ &\le \Big| \frac{\sigma_t'}{\sigma_t} \Big| |x_j| + e^{2L} \Big| \Big(1 - \frac{\sigma_t'}{\sigma_t} t\Big) \Big| \mathbb{E}_{Z\sim\mathcal{N}(\frac{t}{t^2+\sigma_t^2}x, (1+\frac{t^2}{\sigma_t^2})^{-1} I_d)}[|Z_j|] \\ &\le \Big| \frac{\sigma_t'}{\sigma_t} \Big| |x_j| + e^{2L} \Big| \Big(1 - \frac{\sigma_t'}{\sigma_t} t\Big) \Big| \Big(\frac{t}{t^2+\sigma_t^2} |x_j| + \Big(1 + \frac{t^2}{\sigma_t^2}\Big)^{-\frac{1}{2}} \Big) \\ &= \Big(\Big| \frac{\sigma_t'}{\sigma_t} \Big| + e^{2L} \Big| \Big(1 - \frac{\sigma_t'}{\sigma_t} t\Big) \Big| \frac{t}{t^2+\sigma_t^2} \Big) |x_j| + \Big| \Big(1 - \frac{\sigma_t'}{\sigma_t} t\Big) \Big| \Big(1 + \frac{t^2}{\sigma_t^2}\Big)^{-\frac{1}{2}}.\end{aligned}$$

As

$$\max_{t\in[0,1]} \Big| \frac{\sigma_t'}{\sigma_t} \Big| = \log(\sigma_{\min}^{-1}), \quad \Big(1 + \frac{t^2}{\sigma_t^2}\Big)^{-\frac{1}{2}} \le 1$$

and due to Lemma 5.32

$$\frac{t}{t^2+\sigma_t^2} \leq \max(\log(\sigma_{\min}^{-1}), e^2)$$

we obtain for the maximum over $t \in [0,1]$ for n big enough

$$|v_t^j(x)| \lesssim \log(n)^2|x_j| + \log(n).$$

The constants lead to the bound in (5.21). An analogous calculation shows that, for $t \in [0,1]$, the norm of v_t is bounded by

$$|v_t(x)| \lesssim \log(n)^2|x| + \log(n). \tag{5.43}$$

Therefore, we obtain for every $v \in \mathcal{M}$ and $y \in A$ using the construction of $v_t(\cdot|\cdot)$ and fact that the functions in $\mathcal{M}$ are cut at (5.21) $[-\log(n), \log(n)]^d$ for all t,

$$\begin{aligned} g(v,y) &= \int_0^1 \int |v_t(x) - v_t(x|y)|^2 p_t(x|y) \,\mathrm{d}x \,\mathrm{d}t \\ &\lesssim \log(n)^6 + \int_0^1 \int |v(x|y)|^2 p_t(x|y) \,\mathrm{d}x \,\mathrm{d}t. \end{aligned}$$

Since

$$\begin{aligned} \int |v(x|y)|^2 p_t(x|y) \,\mathrm{d}x &= \int \Big|\frac{\sigma_t'}{\sigma_t}x + \Big(1 - \frac{\sigma_t'}{\sigma_t}t\Big)y\Big|^2 p_t(x|y) \,\mathrm{d}x \\ &= \Big|\frac{\sigma_t'}{\sigma_t}\Big|^2 \int |x|^2 p_t(x|y) \,\mathrm{d}x + \Big(1 - \frac{\sigma_t'}{\sigma_t}t\Big)^2 |y|^2, \end{aligned}$$

and as $p_t(\cdot|y)$ is the density of a Gaussian with mean y and variance $\sigma_t^2 I_d$,

$$\int |x|^2 p_t(x|y) \,\mathrm{d}x = t^2|y|^2 + \sigma_t^2 d, \tag{5.44}$$

we obtain for $g(v,y)$

$$g(v,y) \lesssim \log(n)^6 + \int_0^1 \Big|\frac{\sigma_t'}{\sigma_t}\Big|^2 (t^2|y|^2 + \sigma_t^2 d) \,\mathrm{d}t + |y|^2 \int_0^1 (1 - \frac{\sigma_t'}{\sigma_t}t)^2 \,\mathrm{d}t \lesssim \log(n)^6.$$

We denote the constant in the bound of $g(v,y)$ by D. Now we can use Theorem 2.11 and conclude that for every $a \in (0,1], \delta_1 \in (0, \frac{1}{3})$ and $\tau \in \mathbb{R}_{>0}$

$$\begin{aligned} \mathbb{P}\Big(\sup_{\bar{v} \in \mathcal{M}} \mathbb{E}_{Y \sim \mathbb{P}^*}[g(\bar{v}, Y) \mathbb{1}_{Y \in A}] - \frac{1+a}{n} \sum_{i=1}^n g(\bar{v}, X_i) \mathbb{1}_{X_i \in A} \\ > \frac{(1+6/a)D\log(n)^6}{3n} \log\Big(\frac{\mathcal{N}(\tau, g(\mathcal{M}), \|\cdot\|_\infty)}{\delta_1}\Big) + (1+a)\tau\Big) \leq \delta_1. \end{aligned}$$

We keep the term $\frac{a}{n}\sum_{i=1}^n g(\bar{v}, X_i)\mathbb{1}_{X_i\in A}$ separate. For $\frac{1}{n}\sum_{i=1}^n g(\bar{v}, X_i)\mathbb{1}_{X_i\in A}$ we conclude that with a probability of at least $1-\delta_1$

$$\begin{aligned}
&\int_A \int\int |\hat{v}_t(x) - v_t(x|y)|^2 p_t(x|y)\,\mathrm{d}x\,\mathrm{d}t\,p^*(x)\,\mathrm{d}y \\
&\qquad \leq \frac{1+a}{n}\sum_{i=1}^n g(\hat{v}, X_i)\mathbb{1}_{X\in A} + \frac{(1+6/a)D\log(n)^6}{3n}\log\Big(\frac{\mathcal{N}(\tau, g(\mathcal{M}), \|\cdot\|_\infty)}{\delta_1}\Big) + (1+a)\tau.
\end{aligned}$$

Due to the choice of $\hat{v}$ as the empirical risk minimizer, we know that for every $\tilde{v}\in\mathcal{M}$

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n g(\hat{v}, X_i) &= \frac{1}{n}\sum_{i=1}^n \Big(\int\int |\hat{v}_t(x) - v_t(x|X_i)|^2 p_t(x|X_i)\,\mathrm{d}x\,\mathrm{d}t\Big)\mathbb{1}_{X_i\in A} \\
&= \frac{1}{n}\sum_{i=1}^n \int\int |\hat{v}_t(x) - v_t(x|X_i)|^2 p_t(x|X_i)\,\mathrm{d}x\,\mathrm{d}t \\
&\quad - \frac{1}{n}\sum_{i=1}^n \Big(\int\int |\hat{v}_t(x) - v_t(x|X_i)|^2 p_t(x|X_i)\,\mathrm{d}x\,\mathrm{d}t\Big)\mathbb{1}_{X_i\notin A} \\
&\leq \frac{1}{n}\sum_{i=1}^n \int\int |\tilde{v}_t(x) - v_t(x|X_i)|^2 p_t(x|X_i)\,\mathrm{d}x\,\mathrm{d}t \\
&\quad - \frac{1}{n}\sum_{i=1}^n \Big(\int\int |\hat{v}_t(x) - v_t(x|X_i)|^2 p_t(x|X_i)\,\mathrm{d}x\,\mathrm{d}t\Big)\mathbb{1}_{X_i\notin A}.
\end{aligned}$$

As

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \int\int |\tilde{v}_t(x) - v_t(x|X_i)|^2 p_t(x|X_i)\,\mathrm{d}x\,\mathrm{d}t \\
&\qquad = \frac{1}{n}\sum_{i=1}^n \Big(\int\int |\tilde{v}_t(x) - v_t(x|X_i)|^2 p_t(x|X_i)\,\mathrm{d}x\,\mathrm{d}t\Big)\mathbb{1}_{X_i\in A} \\
&\qquad\quad + \frac{1}{n}\sum_{i=1}^n \Big(\int\int |\tilde{v}_t(x) - v_t(x|X_i)|^2 p_t(x|X_i)\,\mathrm{d}x\,\mathrm{d}t\Big)\mathbb{1}_{X_i\notin A},
\end{aligned}$$

we can use the other case of Theorem 2.11 to conclude that for every $a\in(0,1], \delta_2\in(0,\frac{1}{3})$ and $\tau\in\mathbb{R}_{>0}$

$$\begin{aligned}
&\mathbb{P}\Big(\sup_{\bar{v}\in\mathcal{M}} \frac{1}{n}\sum_{i=1}^n g(\bar{v}, X_i)\mathbb{1}_{X_i\in A} - (1+a)\mathbb{E}_{Y\sim\mathbb{P}^*}[g(\bar{v}, Y)\mathbb{1}_{Y\in A}] \\
&\qquad > \frac{(1+3/a)D\log(n)^6}{3n}\log\Big(\frac{\mathcal{N}(\tau, g(\mathcal{M}), \|\cdot\|_\infty)}{\delta_2}\Big) + (1+a)\tau\Big) \leq \delta_2.
\end{aligned}$$

We conclude that with a probability of $1-\delta_2$

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \Big(\int\int |\tilde{v}_t(x) - v_t(x|X_i)|^2 p_t(x|X_i)\,\mathrm{d}x\,\mathrm{d}t\Big)\mathbb{1}_{X_i\in A} \\
&\qquad \leq (1+a)\mathbb{E}_{Y\sim\mathbb{P}^*}[g(\tilde{v}, Y)\mathbb{1}_{Y\in A}] + \frac{(1+3/a)D\log(n)^6}{3n}\log\Big(\frac{\mathcal{N}(\tau, g(\mathcal{M}), \|\cdot\|_\infty)}{\delta_2}\Big)
\end{aligned}$$

$$+ (1+a)\tau.$$

Like before, we separate $a\mathbb{E}_{Y\sim\mathbb{P}^*}[g(\tilde{v},Y)\mathbb{1}_{Y\in A}]$. Using Lemma 5.24 again, we obtain

$$\mathbb{E}_{Y\sim\mathbb{P}^*}[g(\tilde{v},Y)\mathbb{1}_{Y\in A}] = \mathbb{E}_{\substack{t\sim\mathcal{U}[0,1]\\ Y\sim p^*\\ X_t\sim p_t(\cdot|Y)}}\left[\,|\tilde{v}_t(X_t) - v_t\left(X_t|Y\right)|^2\,\right] - \int_{\mathbb{R}^d\setminus A}\int\int |\tilde{v}_t(x) - v_t(x|y)|^2 p_t(x|y)\,\mathrm{d}x\,\mathrm{d}t\,p^*(x)\,\mathrm{d}y.$$

Further

$$\mathbb{E}_{\substack{t\sim\mathcal{U}[0,1]\\ Y\sim p^*\\ X_t\sim p_t(\cdot|Y)}}\left[\,|\tilde{v}_t(X_t) - v_t\left(X_t|Y\right)|^2\,\right] = \mathbb{E}_{\substack{t\sim\mathcal{U}[0,1]\\ X_t\sim p_t}}\left[|v_t(X_t) - \tilde{v}_t(X_t)|^2\right] + C.$$

Note that C cancels with the same constant in (5.41). The terms arising from the restriction of the Bernstein-bound to the set A can be bounded with high probability by the following result.

Lemma 5.27. *Fix $C\in\mathbb{R}$. Let $\log(\sigma_{\min}^{-1})\asymp\log(n)$. With a probability of $1-\frac{1}{3n}$ for $n\geq e^{3(d+7+\frac{1}{2})}$*

$$|\mathbb{E}[g(\hat{v},Y)\mathbb{1}_{Y\notin A}] - \mathbb{E}[g(\tilde{v},Y)\mathbb{1}_{Y\notin A}] + \sum_{i=1}^{n} g(\tilde{v},X_i)\mathbb{1}_{X_i\notin A} - g(\hat{v},X_i)\mathbb{1}_{X_i\notin A}| \leq n^{-\frac{1}{2}}.$$

Collecting all other terms, inserting $\log(\sigma_{\min})\asymp\log(n)$, setting $\delta_1=\delta_2=\frac{1}{3n}$ and using the union bound leads to the result. The infimum can be used due to the presence of other nonzero terms. □

Preparation for the proof of Theorem 5.21. The proof of Theorem 5.21 requires some additional results to bound the covering number and apply the approximation result. We further need to look carefully into the approximation result of Theorem 2.6 and the proof of Gühring et al. (2020).

As we cannot approximate a function on $\mathbb{R}^d$ with a finite neural network with fixed precision, we are going to look at functions that approximate v on the set $[-\log(n),\log(n)]^d\times[0,1]$ and show that the error on the complement is small. Hence we want to determine a network necessary to obtain

$$\int\int_{[-\log(n),\log(n)]^d} |\tilde{v}_t(x) - v_t(x)|^2 p_t(x)\,\mathrm{d}x\,\mathrm{d}t \leq \varepsilon,$$

for a given approximation error $\varepsilon>0$. We map any point outside of $[-\log(n),\log(n)]^d$ back to this hypercube using the one layer ReLU net associated with the following clipping function

$$\mathrm{clip}_{\mathrm{input}}(x_i,-\log(n),\log(n)) \coloneqq x_i - \phi(x_i-\log(n)) + \phi(-x_i-\log(n)),$$

where ϕ is the ReLU activation function. Thus, the Lipschitz constant of a network that approximates v on the set $[-\log(n),\log(n)]^d\times[0,1]$ is bounded by the Lipschitz constant of the network on $[-\log(n),\log(n)]^d\times[0,1]$. A similar one layer clipping function can be used to clip the component functions at (5.21). The order of the number of layers and the number of nonzero weights will not change with both adaptations. For the error on the complement, we

use the following result:

Lemma 5.28. *For every $\tilde{v}$ in $\mathcal{M}$ and $\log(\sigma_{\min}^{-1}) \asymp \log(n)$, if $n \geq e^{4(\frac{1}{2}+6+d)}$ then*

$$\int_0^1 \int_{\mathbb{R}^d \setminus [-\log(n),\log(n)]^d} |v_t(x) - \tilde{v}_t(x)|^2 p_t(x) \, \mathrm{d}x \, \mathrm{d}t \leq n^{-\frac{1}{2}}.$$

Bound of the covering number:

Lemma 5.29. *We have for $\bar{v}^1, \bar{v}^2 \in \mathcal{M}$ and every $y \in [-\log(n), \log(n)]^d$, that*

$$|g(\bar{v}^1, y) - g(\bar{v}^2, y)| \lesssim \|\bar{v}^1 - \bar{v}^2\|_\infty \log(n)^4.$$

Hence we can bound

$$\mathcal{N}(\tau, g(\mathcal{M}), \|\cdot\|_\infty) \leq \mathcal{N}(\frac{\tau}{\log(n)^4}, \mathcal{M}, \|\cdot\|_\infty).$$

From Suzuki (2019, Lemma 3) we know that the covering number of the set ReLU networks of a cube $[0,1]^d$, denoted by $\mathsf{NN}_{\mathrm{ReLU}}$, is

$$\log(\mathcal{N}(\tau, \mathsf{NN}_{\mathrm{ReLU}}, \|\cdot\|_{\infty,[0,1]^d})) \leq 2SL \log\left(\tau^{-1} L(B \vee 1)(W+1)\right),$$

where L is the number of layers, S is the number of nonzero weights, B is the maximal absolute value of a single weight and W is the maximal width. A careful inspection of the proof of the approximation result will later reveal the specific choices.
To consider functions on $[-\log(n), \log(n)]^d \times [0,1]$, we proceed analogously to Yakovlev & Puchkin (2025, p. 46), multiplying the first d coordinates of the weight matrix of the first lemma with $\log(n)$ and dividing the input vector with the same value. This has of course an impact on the bound of the weights, which will scale up by the factor $\log(n)$. Additionally, the weights of the clipping layers are of order $\log(n)^3$. We thus obtain

$$\log(\mathcal{N}(\tau, g(\mathcal{M}), \|\cdot\|_\infty)) \lesssim 2SL \log\left(\tau^{-1} \log(n)^4 L(B \vee 1) \log(n)^3 (W+1)\right).$$

Approximation result:
Since the covering number depends on the maximum bound of the weights, we need to track this in the proof of Theorem 2.6 by Gühring et al. (2020). Gühring et al. (2020, Lemma C.3 (v)) reveals that the biggest single weight is set to N, which is later chosen as $\left\lceil \left(\frac{\varepsilon}{2CL}\right)^{-\frac{1}{s-1}} \right\rceil$, where L is the bound on the partial derivatives up to the highest order considered. To apply Theorem 2.6 without tracking the bound on the partial derivatives in their proof, we approximate v divided by this bound and scale the approximated function up in the end. This can be achieved adding a layer in front and a layer after the network. The maximal weight used is of the size of the bound.
As v is in C^∞, we can apply Theorem 2.6 for arbitrary large s. However, the use of a larger s will lead to a larger bound on the absolute value of the partial derivatives up to the order s. Since this scales up the approximation error, we need a precise quantification. The next result uses the logarithmic Sobolev inequality to bound higher derivatives of v.

Lemma 5.30. *Fix $s \in \mathbb{N}$. Let $\{i_1, ..., i_s\} \subset \{1, ..., d\}^s$. For the k-th order partial derivative of v we have that*

$$\frac{\partial^s}{\partial x_{i_1}, ..., \partial x_{i_s}} v_t^j(x) \lesssim \log(\sigma_{\min}^{-1}) \sigma_{\min}^{-s+1}.$$

For the derivatives with respect to t, we obtain the following result.

Lemma 5.31. *Fix $s \in \mathbb{N}$. Assume that $\log(\sigma_{\min}) \geq s$. For every $k \in \{1, ..., s\}$ we have that*

$$\frac{\partial^k}{\partial t^k} v_t^j(x) \lesssim \operatorname{polylog}(\sigma_{\min}^{-1}) \operatorname{polylog}(n) \sigma_{\min}^{-s-2}.$$

The bound on mixed derivatives follows exactly the same lines as the proofs of Lemma 5.30 and Lemma 5.31. Now we are ready to prove Theorem 5.21.

Proof of Theorem 5.21.

We scale the function down by $\sigma_{\min}^{-s-2} \log^2(n)$ and approximate $v \cdot \sigma_{\min}^{-s-2} \log^2(n)$ instead of v. After that, we add another layer to scale the output of the neural net up. The approximation error, ε, will also scale up by $\sigma_{\min}^{-s-2} \log^2(n)$. The maximal weight in the inner network will then no longer depend on this bound, but the largest weight might increase. Hence we can set

$$W = C\big(\varepsilon^{-\frac{1}{s-2}} \vee \sigma_{\min}^{-s-2} \log^2(n)\big).$$

Inserting everything into Theorem 5.20 and setting $\tilde{d} = d + 1$, we obtain for $a < 1$ with a probability of $1 - \frac{1}{n}$ for n big enough

$$\begin{aligned}
&\mathbb{E}_{\substack{t \sim \mathcal{U}[0,1] \\ X_t \sim p_t}} \big[|v_t(X_t) - \tilde{v}_t(X_t)|^2 \big] \\
&\lesssim + \int_0^1 \int_{\mathbb{R}^d \setminus [-\log(n), \log(n)]^d} |v_t(x) - \tilde{v}_t(x)|^2 p_t(x) \, \mathrm{d}x \, \mathrm{d}t + (2+a)\tau + a \log(n)^4 + \varepsilon \sigma_{\min}^{-s-2} \log^2(n) \\
&\quad + \frac{a^{-1} \log(n)^7}{n} \varepsilon^{-\frac{\tilde{d}}{s-1}} \log\big(2\big(\tau^{-1} \log(n)^7 L(B \vee 1)(\varepsilon^{-\frac{1}{s-1}} \vee \sigma_{\min}^{-s-2} \log^2(n))\big)\big) \\
&\quad + (2+a)\tau + a \log(n)^6 + \varepsilon \sigma_{\min}^{-s-2} \log^2(n). \qquad (5.45)
\end{aligned}$$

Note that B and L will depend on ε, thus we are restricting feasible choices of ε to choices that grow at most polynomial in n^{-1}. Ignoring logarithmic terms, we first solve for a

$$\frac{a^{-1} \varepsilon^{-\frac{\tilde{d}}{s-1}}}{n} \asymp a \quad \Longleftrightarrow \quad a \asymp \Big(\frac{\varepsilon^{-\frac{\tilde{d}}{s-1}}}{n}\Big)^{\frac{1}{2}}.$$

τ can be set such that the corresponding term is of the same order, it suffices to set

$$\tau \asymp \frac{\Big(\frac{\varepsilon^{-\frac{\tilde{d}}{s-1}}}{n}\Big)^{\frac{1}{2}}}{2 + \Big(\frac{\varepsilon^{-\frac{\tilde{d}}{s-1}}}{n}\Big)^{\frac{1}{2}}}.$$

This choice, as well as the choice of ε, influence the first term in (5.45) only logarithmically. Now

we solve for ε

$$\Big(\frac{\varepsilon^{-\frac{\tilde{d}}{s-1}}}{n}\Big)^{\frac{1}{2}} \asymp \varepsilon\sigma_{\min}^{-s-2} \quad \iff \quad \varepsilon \asymp n^{-\frac{s-1}{\tilde{d}+2s-2}}\sigma_{\min}^{(s+2)\frac{2s-2}{\tilde{d}+2s-2}}.$$

Restricting $\sigma_{\min}$ to polynomials in n, we ensure ε is polynomial in n. Thus, ignoring logarithmic factors in n, we obtain with a probability of $1-\frac{1}{n}$ for n big enough

$$\mathbb{E}_{\substack{t\sim\mathcal{U}[0,1]\\ X_t\sim p_t}}\big[|v_t(X_t)-\tilde{v}_t(X_t)|^2\big] \lesssim \operatorname{polylog}(n)n^{-\frac{s-1}{\tilde{d}+2s-2}}\sigma_{\min}^{(s+2)\left(\frac{2s-2}{\tilde{d}+2s-2}-1\right)}.$$

To choose $\sigma_{\min}$ such that is is optimal in the setting of Theorem 5.5, we have to set it such that

$$\sigma_{\min}^{1+\alpha} \asymp \Big(n^{-\frac{s-1}{\tilde{d}+2s-2}}\sigma_{\min}^{(s+2)\left(\frac{2s-2}{\tilde{d}+2s-2}-1\right)}\Big)^{\frac{1}{2}} \quad \iff \quad \sigma_{\min} \asymp n^{-\frac{1}{\frac{s+2}{s-1}\tilde{d}+(2\alpha+2)\left(\frac{\tilde{d}}{s-1}+2\right)}}.$$

This choice is a polynomial of n^{-1} and hence a feasible choice for $\sigma_{\min}$. For $n \geq 1$ and $\sigma_{\min} < 1$, we know that the network approximates the Lipschitz constant of v with precision $\varepsilon\sigma_{\min}^{-s-2}$. Plugging in the choice of $\sigma_{\min}$ leads to

$$\varepsilon\sigma_{\min}^{-s-2} \asymp n^{-\frac{2+2\alpha}{\frac{s+2}{s-1}\tilde{d}+(2\alpha+2)(\frac{\tilde{d}}{s-1}+2)}} \leq 1.$$

Hence we obtain for the Lipschitz constant $\hat{L}_t$ of $\hat{v}_t$

$$e^{\int_0^1 \hat{L}_t \,\mathrm{d}t} \leq e^{\int_0^1 \Gamma_t+1 \,\mathrm{d}t} \leq e^{C+1}.$$

Note that we cannot let $s \to \infty$, as this will blow up the constant. However, for every fixed $\eta > 0$, we can choose s such that

$$s = \Big\lceil (5+2\alpha)\frac{\tilde{d}}{\eta}+1\Big\rceil.$$

This leads to the final choice

$$\sigma_{\min} \asymp n^{-\frac{1}{\tilde{d}+4\alpha+4+\eta}}.$$

Now we use the results of Lemma 5.27, set $\delta = \frac{1}{2n}$ and use the union bound. Combining this with Theorem 5.5, Theorem 5.17, Theorem 4.3 and $\tilde{d} = d+1$, we obtain with a probability of $1-\frac{1}{n}$

$$\mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{\hat{\psi}_1(Z)}) \lesssim \operatorname{polylog}(n)n^{-\frac{1+\alpha}{d+4\alpha+5+\eta}}.$$

The number of hidden layers L of f_{NN} is bounded by $c\cdot\log(n)$ and the number of nonzero weights S is bounded by $c\cdot n^{c(d,\alpha,\eta)}\cdot\log^2(n)$, where c is a constant independent of n. □

Additional proofs of Section 5.6.4

Proof of Lemma 5.25. By the definition of v_t and $v_t(\cdot|\cdot)$, we get

$$v_t(x) = \int \Big(\frac{\sigma_t'}{\sigma_t}(x-\mu_t(y))+\mu_t'(y)\Big)\frac{p_t(x|y)}{p_t(x)}p^*(y)\,\mathrm{d}y,$$

where $'$ indicates the derivative with respect to t. For the Jacobian, we get using $\mu_t(y) = t^\gamma y$

$$\begin{aligned}
D_x v_t(x) &= D_x \int \big(\frac{\sigma_t'}{\sigma_t}(x - \mu_t(y)) + \mu_t'(y)\big)\frac{p_t(x|y)}{p_t(x)}p^*(y)\,\mathrm{d}y \\
&= D_x \frac{\sigma_t'}{\sigma_t}\frac{x}{p_t(x)}\int p_t(x|y)p^*(y)\,\mathrm{d}y - D_x\frac{\sigma_t'}{\sigma_t}\frac{t^\gamma}{p_t(x)}\int y p_t(x|y)p^*(y)\,\mathrm{d}y \\
&\quad + D_x \frac{\gamma t^{\gamma-1}}{p_t(x)}\int y p_t(x|y)p^*(y)\,\mathrm{d}y \\
&= \frac{\sigma_t'}{\sigma_t} I_d + \big(\gamma t^{\gamma-1} - \frac{\sigma_t' t^\gamma}{\sigma_t}\big) D_x \frac{\int y p_t(x|y)p^*(y)\,\mathrm{d}y}{p_t(x)}.
\end{aligned}$$

For the derivative with respect to x_i of the $j-th$ coordinate function, we get using the dominated convergence theorem

$$\begin{aligned}
&\frac{\partial}{\partial x_i}\frac{\int y_j p_t(x|y)p^*(y)\,\mathrm{d}y}{\int p_t(x|y)p^*(y)\,\mathrm{d}y} \\
&= \frac{(\int y_j \frac{\partial}{\partial x_i} p_t(x|y)p^*(y)\,\mathrm{d}y)(\int p_t(x|y)p^*(y)\,\mathrm{d}y) - (\int y_j p_t(x|y)p^*(y)\,\mathrm{d}y)(\int \frac{\partial}{\partial x_i} p_t(x|y)p^*(y)\,\mathrm{d}y)}{(\int p_t(x|y)p^*(y)\,\mathrm{d}y)^2} \\
&= \frac{(\int y_j (-\frac{x_i - t^\gamma y_i}{\sigma_t^2}) p_t(x|y)p^*(y)\,\mathrm{d}y)(\int p_t(x|y)p^*(y)\,\mathrm{d}y)}{(\int p_t(x|y)p^*(y)\,\mathrm{d}y)^2} \\
&\quad - \frac{(\int y_j p_t(x|y)p^*(y)\,\mathrm{d}y)(\int (-\frac{x_i - t^\gamma y_i}{\sigma_t^2}) p_t(x|y)p^*(y)\,\mathrm{d}y)}{(\int p_t(x|y)p^*(y)\,\mathrm{d}y)^2} \\
&= \frac{t^\gamma \int y_j y_i p_t(x|y)p^*(y)\,\mathrm{d}y}{\sigma_t^2 \int p_t(x|y)p^*(y)\,\mathrm{d}y} - \frac{t^\gamma (\int y_j p_t(x|y)p^*(y)\,\mathrm{d}y)(\int y_i p_t(x|y)p^*(y)\,\mathrm{d}y)}{\sigma_t^2 (\int p_t(x|y)p^*(y)\,\mathrm{d}y)^2}.
\end{aligned}$$

For fixed x and t let $Y^{x,t}$ be a random variable with density $\frac{p_t(x|\cdot)p^*(\cdot)}{\int p_t(x|y)p^*(y)\,\mathrm{d}y}$. Then for $t \in [0,1)$

$$\frac{\partial}{\partial x_i}\frac{\int y_j p_t(x|y)p^*(y)\,\mathrm{d}y}{\int p_t(x|y)p^*(y)\,\mathrm{d}y} = \frac{t^\gamma}{\sigma_t^2}(\mathbb{E}[Y_i^{x,t}Y_j^{x,t}] - \mathbb{E}[Y_i^{x,t}]\mathbb{E}[Y_j^{x,t}]) = \frac{t^\gamma}{\sigma_t^2}\operatorname{Cov}(Y^{x,t})_{ji}.$$

We obtain for the derivative of with respect to x_i of the $j-th$ coordinate function

$$\frac{\partial}{\partial x_i} v_t^j(x) = \frac{\sigma_t'}{\sigma_t}\mathbb{1}_{i=j} + \Big(\gamma t^{\gamma-1} - \frac{\sigma_t' t^\gamma}{\sigma_t}\Big)\frac{t^\gamma}{\sigma_t^2}\operatorname{Cov}(Y^{x,t})_{ji}. \qquad \square$$

Proof of Lemma 5.26. Let t^* be such that $\sigma_{t^*} = \frac{1}{\vartheta}$. Then

$$\begin{aligned}
&\int_0^{t^*} \Big|\frac{\sigma_t'}{\sigma_t}\mathbb{1}_{i=j} + \Big(\gamma t^{\gamma-1} - \frac{\sigma_t' t^\gamma}{\sigma_t}\Big)\frac{t^\gamma}{\sigma_t^2}\operatorname{Cov}(Y^{\cdot,t})_{ij}\Big|\,\mathrm{d}t \\
&\qquad\qquad \le \int_0^{t^*}\Big|\frac{\sigma_t'}{\sigma_t}\Big|\,\mathrm{d}t + \int_0^{t^*}\Big|\Big(\gamma t^{\gamma-1} - \frac{\sigma_t' t^\gamma}{\sigma_t}\Big)\frac{t^\gamma}{\sigma_t^2}\operatorname{Cov}(Y^{\cdot,t})_{ij}\Big|\,\mathrm{d}t.
\end{aligned}$$

For the first term, we can use that σ_t' is non positive and hence by change of variables

$$\int_0^{t^*}\Big|\frac{\sigma_t'}{\sigma_t}\Big|\,\mathrm{d}t = -\int_0^{t^*}\frac{\sigma_t'}{\sigma_t}\,\mathrm{d}t = \int_{\sigma_{t^*}}^1 \frac{1}{u}\,\mathrm{d}u = \log(\sigma_{t^*}^{-1}) = \log(\vartheta).$$

For the second integral, we can bound the Covariance term by

$$\big|\operatorname{Cov}(Y^{\cdot,t})_{ij}\big| \le C.$$

Now we get

$$\begin{aligned}\int_0^{t^*} \Big|\Big(\gamma t^{\gamma-1} - \frac{\sigma_t' t^\gamma}{\sigma_t}\Big)\frac{t^\gamma}{\sigma_t^2}\operatorname{Cov}(Y^{\cdot,t})_{ij}\Big|\,\mathrm{d}t &\le C\int_0^{t^*}\Big|\Big(\gamma t^{\gamma-1} - \frac{\sigma_t' t^\gamma}{\sigma_t}\Big)\frac{t^\gamma}{\sigma_t^2}\Big|\,\mathrm{d}t \\ &\le C\int_0^{t^*}\Big|\gamma t^{\gamma-1}\frac{t^\gamma}{\sigma_t^2}\Big|\,\mathrm{d}t + C\int_0^{t^*}\Big|\frac{\sigma_t' t^{2\gamma}}{\sigma_t^3}\Big|\,\mathrm{d}t \le \vartheta^2 C\int_0^{t^*}\gamma t^{\gamma-1}\,\mathrm{d}t \\ &\quad + \vartheta^2 C\int_0^{t^*}\Big|\frac{\sigma_t'}{\sigma_t}\Big|\,\mathrm{d}t \lesssim \vartheta^2(1+\log(\vartheta)).\end{aligned}$$

□

Proof of Lemma 5.27. First we abbreviate

$$B \coloneqq \mathbb{E}[g(\hat{v},Y)\mathbb{1}_{Y\notin A}] - \mathbb{E}[g(\tilde{v},Y)\mathbb{1}_{Y\notin A}] + \sum_{i=1}^n g(\tilde{v},X_i)\mathbb{1}_{X_i\notin A} - g(\hat{v},X_i)\mathbb{1}_{X_i\notin A}.$$

Both $|\tilde{v}|_\infty$ and $|\hat{v}|_\infty$ are bounded by $D'\log(n)^3$, where D' is a constant collecting all terms in (5.21). In the proof of Theorem 5.20, we showed that

$$g(v,y) \lesssim \log(n)^6 + \log(n)^2|y|^2 + \log(\sigma_{\min}^{-1})^2\log(n)^2|y|^2.$$

Thus, we obtain

$$\begin{aligned}\mathbb{E}_{X_i}[|B|] &\lesssim \mathbb{E}\Big[\Big|\mathbb{1}_{Y\notin A}\big(\log(n)^6 + \log(n)^2|Y|^2\big)\Big|\Big] \\ &= \log(n)^6\int_{\mathbb{R}^d\setminus A} p^*(y)\,\mathrm{d}y + \log(n)^2\int_{\mathbb{R}^d\setminus A}|y|^2 p^*(y)\,\mathrm{d}y.\end{aligned}$$

Using the structure of p^* and assuming $n \ge 4$, we obtain

$$\begin{aligned}\int_{\mathbb{R}^d\setminus A} p^*(y)\,\mathrm{d}y \le \int_{\mathbb{R}^d\setminus A}|y|^2 p^*(y)\,\mathrm{d}y &= \frac{\int_{\mathbb{R}^d\setminus A}|y|^2\exp\big(-\frac{|y|^2}{2} - a(y)\big)\,\mathrm{d}y}{\int\exp\big(-\frac{|y|^2}{2}-a(y)\big)\,\mathrm{d}y} \\ &\le e^{2L}\frac{\int_{\mathbb{R}^d\setminus A}|y|^2\exp\big(-\frac{|y|^2}{2}\big)\,\mathrm{d}y}{\int\exp\big(-\frac{|y|^2}{2}\big)\,\mathrm{d}y} = e^{2L}\left(2\frac{\int_{\log(n)}^\infty y_1^2\exp\big(-\frac{y_1^2}{2}\big)\,\mathrm{d}y_1}{\int\exp\big(-\frac{y_1^2}{2}\big)\,\mathrm{d}y_1}\right)^d.\end{aligned}$$

As

$$\begin{aligned}0 &\le \int_{\log(n)}^\infty y_1^2\exp\Big(-\frac{y_1^2}{2}\Big)\,\mathrm{d}y \\ &= \Big[-y_1\exp\Big(-\frac{y_1^2}{2}\Big)\Big]_{\log(n)}^\infty - \int_{\log(n)}^\infty\exp\Big(-\frac{y_1^2}{2}\Big)\,\mathrm{d}y_1 \le \log(n)\exp\Big(-\frac{\log(n)^2}{2}\Big),\end{aligned}$$

we can bound the above by

$$e^{2L}\left(2\frac{\int_{\log(n)}^{\infty} y_1^2 \exp\big(-\frac{y_1^2}{2}\big)\,\mathrm{d}y_1}{\int \exp\big(-\frac{y_1^2}{2}\big)\,\mathrm{d}y_1}\right)^d \lesssim \Big(\log(n)\exp\Big(-\frac{\log(n)^2}{2}\Big)\Big)^d.$$

For large n, this decays faster than any polynomial in n. We conclude

$$\mathbb{E}_{X_i}[|B|] \lesssim \log(n)^6\Big(\log(n)\exp\Big(-\frac{\log(n)^2}{2}\Big)\Big)^d.$$

Now Markov's inequality yields

$$\begin{aligned}\mathbb{P}(B \geq n^{-\frac{1}{2}}) &\lesssim \log(n)^6\Big(\log(n)\exp\Big(-\frac{\log(n)^2}{2}\Big)\Big)^d n^{\frac{1}{2}},\\ \mathbb{P}(B \leq -n^{-\frac{1}{2}}) &\lesssim \log(n)^6\exp\Big(-\frac{\log(n)^2}{2}\Big)\Big)^d n^{\frac{1}{2}}.\end{aligned}$$

Notice that we can shift the constant as a factor of the bound of B. As $\log(\sigma_{\min}^{-1}) \asymp \log(n)$, we need to choose n large enough such that

$$\log(n)^6\Big(\log(n)\exp(-\frac{\log(n)^2}{2})\Big)^d n^{\frac{1}{2}} \lesssim \frac{1}{3n} \quad \Longleftarrow \quad n \gtrsim e^{3(d+7+\frac{1}{2})},$$

where the implication stems from a loose upper bound using $\log(n)^{6+d} \leq n^{6+d}$. This finishes the proof. □

Proof of Lemma 5.28. From the proof of Theorem 5.20 and as $\log(\sigma_{\min}^{-1}) \asymp \log(n)$ we know that

$$|v_t(x)| \lesssim \log(n)^2|x| + \log(n), \quad |\tilde{v}_t(x)| \lesssim \log(n)^3.$$

Thus

$$\begin{aligned}\int_0^1 \int_{\mathbb{R}^d\setminus[-\log(n),\log(n)]^d} |v_t(x)-\tilde{v}_t(x)|^2 p_t(x)\,\mathrm{d}x\,\mathrm{d}t \lesssim \log(n)^6 \int_0^1 \int_{\mathbb{R}^d\setminus[-\log(n),\log(n)]^d} p_t(x)\,\mathrm{d}x\\ + \log(n)^4 \int_0^1 \int_{\mathbb{R}^d\setminus[-\log(n),\log(n)]^d} |x|^2 p_t(x)\,\mathrm{d}x.\end{aligned}$$

By the definition of p_t and due to the fact that the convolution of two densities is again a density, we know that

$$\begin{aligned}p_t(x) = \int p_t(x|y)p^*(y)\,\mathrm{d}y &= \frac{\int \exp\big(-\frac{|x-ty|^2}{2\sigma_t^2} - \frac{|y|^2}{2} - a(y)\big)\,\mathrm{d}y}{\int\int \exp\big(-\frac{|x-ty|^2}{2\sigma_t^2} - \frac{|y|^2}{2} - a(y)\big)\,\mathrm{d}y\,\mathrm{d}x}\\ &\leq e^{2L}\frac{\int \exp\big(-\frac{|x-ty|^2}{2\sigma_t^2} - \frac{|y|^2}{2}\big)\,\mathrm{d}y}{\int\int \exp\big(-\frac{|x-ty|^2}{2\sigma_t^2} - \frac{|y|^2}{2}\big)\,\mathrm{d}y\,\mathrm{d}x}.\end{aligned}$$

Now for the inner integrals, we obtain

$$\int \exp\Big(-\frac{|x-ty|^2}{2\sigma_t^2}-\frac{|y|^2}{2}\Big)\,\mathrm{d}y = \exp\Big(-\frac{|x|^2}{2\sigma_t^2}\Big)\int \exp\Big(-\frac{1}{2}\Big(1+\frac{t^2}{\sigma_t^2}\Big)|y|^2+\Big\langle\frac{t}{\sigma_t^2}x,y\Big\rangle\Big)\,\mathrm{d}y$$
$$= (2\pi)^{d/2}\Big(\frac{\sigma_t^2}{t^2+\sigma_t^2}\Big)^{d/2}\exp\Big(-\frac{|x|^2}{2(t^2+\sigma_t^2)}\Big).$$

Inserting this into the bound and collecting all factors of $|x|^2$, we obtain

$$p_t(x) \le e^{2L}\frac{\exp\big(-\frac{1}{2}\frac{1}{\sigma_t^2+t^2}|x|^2\big)}{\int\exp\big(-\frac{1}{2}\frac{1}{\sigma_t^2+t^2}|x|^2\big)\,\mathrm{d}x} = \frac{e^{2L}}{(2\pi(\sigma_t^2+t^2))^{\frac{d}{2}}}\exp\Big(-\frac{1}{2}\frac{1}{\sigma_t^2+t^2}|x|^2\Big).$$

Now we bound

$$\begin{aligned}
&\int_0^1\int_{\mathbb{R}^d\setminus[-\log(n),\log(n)]^d} p_t(x)\,\mathrm{d}x\,\mathrm{d}t\\
&\le \int_0^1\int_{\mathbb{R}^d\setminus[-\log(n),\log(n)]^d} |x|^2 p_t(x)\,\mathrm{d}x\,\mathrm{d}t\\
&\le \int_0^1 \frac{e^{2L}}{(2\pi(\sigma_t^2+t^2))^{\frac{d}{2}}}\int_{\mathbb{R}^d\setminus[-\log(n),\log(n)]^d}|x|^2\exp\Big(-\frac{1}{2}\frac{1}{\sigma_t^2+t^2}|x|^2\Big)\,\mathrm{d}x\,\mathrm{d}t\\
&= 2\int_0^1 \frac{e^{2L}}{(2\pi(\sigma_t^2+t^2))^{\frac{d}{2}}}\Big(\int_{\log(n)}^\infty x_1^2\exp\Big(-\frac{1}{2}\frac{1}{\sigma_t^2+t^2}x_1^2\Big)\,\mathrm{d}x_1\Big)^d\,\mathrm{d}t\\
&\le 2\int_0^1 \frac{e^{2L}}{(2\pi(\sigma_t^2+t^2))^{\frac{d}{2}}}\Big(\log(n)(\sigma_t^2+t^2)\exp\Big(-\frac{1}{2}\frac{1}{\sigma_t^2+t^2}\log(n)^2\Big)\Big)^d\,\mathrm{d}t,
\end{aligned}$$

where the last inequality follows from the same arguments as the bound in the proof of the first part of this Lemma. As $(\sigma_t^2+t^2)\le 2$, we can bound

$$\int_0^1\int_{\mathbb{R}^d\setminus[-\log(n),\log(n)]^d} p_t(x)\,\mathrm{d}x\,\mathrm{d}t \lesssim \log(n)^d\exp\Big(-\frac{1}{4}\log^2(n)\Big).$$

Overall, we need to choose n big enough such that

$$\log(n)^{6+d}\exp\Big(-\frac{1}{4}\log^2(n)\Big)\le n^{-\frac{1}{2}} \quad\Longleftarrow\quad n\ge e^{4(\frac{1}{2}+6+d)},$$

where again the implication stems from a loose upper bound using $\log(n)^{6+d}\le n^{6+d}$. □

Proof of Lemma 5.29. Similar to Yakovlev & Puchkin (2025, p. 44), we bound

$$\Big|\int_0^1\int|\bar v_t^1(x)-v_t(x|y)(x)|^2 p_t(x|y)\,\mathrm{d}x\,\mathrm{d}t - \int_0^1\int|\bar v_t^2(x)-v_t(x|y)(x)|^2 p_t(x|y)\,\mathrm{d}x\,\mathrm{d}t\Big|$$
$$\lesssim \int_0^1\int|\bar v_t^1(x)-\bar v_t^2(x)|(|\bar v_t^1(x)|+|\bar v_t^2(x)|+2|v_t(x|y)|)p_t(x|y)\,\mathrm{d}x\,\mathrm{d}t.$$

For $x\in\mathbb{R}^d$ and $y\in[-\log(n),\log(n)]^d$, we have

$$|v_t(x|y)| = \Big|\frac{\sigma_t'}{\sigma_t}(x+(\sigma_t-t)y)\Big| \le \Big|\frac{\sigma_t'}{\sigma_t}\Big|(|x|+|y|) \le \Big|\frac{\sigma_t'}{\sigma_t}\Big|(|x|+\sqrt{d}\log(n)).$$

By (5.43) all $\bar{v} \in \mathcal{M}$ are such that $|\bar{v}_t(x)| \lesssim \log(n)^3$. Therefore

$$\begin{aligned}\int_0^1 \int |\bar{v}_t^1(x) - \bar{v}_t^2(x)|(|\bar{v}_t^1(x)| + |\bar{v}_t^2(x)| + 2|v_t(x|y)|)p_t(x|y)\,\mathrm{d}x\,\mathrm{d}t \\ \lesssim \|\bar{v}^1 - \bar{v}^2\|_\infty(\log(n)^3 + \log(n)\int_0^1 \left|\frac{\sigma_t'}{\sigma_t}\right| \int |x|p_t(x|y)\,\mathrm{d}x\,\mathrm{d}t) \\ \lesssim \|\bar{v}^1 - \bar{v}^2\|_\infty \log(n)^4.\end{aligned}$$

The last inequality follows from Jensen's inequality and the same argument as in (5.44). This concludes the proof. □

Proof of Lemma 5.30. For an s-th order partial derivative and $\{i_1, ..., i_s\} \subset \{1, ..., d\}^s$, we know from the definition of v_t

$$\frac{\partial^s}{\partial x_{i_1}, ..., \partial x_{i_s}} v_t^j(x) = \mathbb{1}_{s=1}\frac{\sigma_t'}{\sigma_t} + \Big(1 - \frac{\sigma_t' t}{\sigma_t}\Big)\frac{\partial^s}{\partial x_{i_1}, ..., \partial x_{i_s}}\mathbb{E}_{q_{t,x}}[Y_j], \tag{5.46}$$

where

$$q_{t,x} \propto \exp\Big(-\frac{\|x - ty\|^2}{2\sigma_t^2} - \frac{\|y\|^2}{2} - a(y)\Big) \propto \exp(\langle \eta, Y\rangle)h(y),$$

where $\eta = \frac{t}{\sigma_t^2}x$ and $h(y) = \exp(-(\frac{t^2}{2\sigma_t^2} + \frac{1}{2})\|y\|^2 - a(y))$. Define

$$A(\eta) \coloneqq \log \int \exp(\langle \eta, Y\rangle)h(y)\,\mathrm{d}y.$$

In context of exponential families, A is typically called the log-partition function. Then

$$\frac{\partial}{\partial \eta_j}A(\eta) = \mathbb{E}_{q_{t,x}}[Y_j].$$

Thus,

$$\frac{\partial^s}{\partial x_{i_1}, ..., \partial x_{i_s}}\mathbb{E}_{q_{t,x}}[Y_j] = \frac{\partial^s}{\partial x_{i_1}, ..., \partial x_{i_s}}\frac{\partial}{\partial \eta_j}A(\eta) = \Big(\frac{t}{\sigma_t^2}\Big)^s \frac{\partial^s}{\partial \eta_{i_1}, ..., \partial \eta_{i_s}}\frac{\partial}{\partial \eta_j}A(\eta).$$

Further recall the definition of the cumulant generating function from Section 2.1.2,

$$K(\lambda) \coloneqq \log\big(\mathbb{E}[\exp(\langle \lambda, Y\rangle)]\big).$$

Then

$$\mathbb{E}[\exp(\langle \lambda, Y\rangle)] = \exp(A(\lambda + \eta) - A(\eta)), \quad K(\lambda) = A(\lambda + \eta) - A(\eta).$$

Differentiating in λ yields

$$\frac{\partial^{s+1}}{\partial \lambda_{i_1}, ..., \partial \lambda_{i_s}, \partial \lambda_j}K(\lambda)\Big|_{\lambda=0} = \frac{\partial^{s+1}}{\partial \lambda_{i_1}, ..., \partial \lambda_{i_s}, \partial \lambda_j}A(\lambda + \eta) - A(\eta)\Big|_{\lambda=0} = \frac{\partial^{s+1}}{\partial \eta_{i_1}, ..., \partial \eta_{i_s}, \partial \eta_j}A(\eta).$$

Hence,

$$\frac{\partial^s}{\partial x_{i_1}, ..., \partial x_{i_s}} v_t^j(x) = \mathbb{1}_{s=1}\frac{\sigma_t'}{\sigma_t} + \Big(1 - \frac{\sigma_t' t}{\sigma_t}\Big)\Big(\frac{t}{\sigma_t^2}\Big)^s \kappa(Y_{i_1}, ..., Y_{i_s}, Y_j),$$

where $\kappa(Y_{i_1}, ..., Y_{i_s}, Y_j)$ is the joint cumulant, of $Y_{i_1}, ..., Y_{i_s}, Y_j$, which is defined as exactly this partial derivative. For $s \geq 1$, this cumulant is shift invariant, hence

$$\kappa(Y_{i_1}, ..., Y_{i_s}, Y_j) = \kappa(Y_{i_1} - \mathbb{E}[Y_{i_1}], ..., Y_{i_s} - \mathbb{E}[Y_{i2}], Y_j - \mathbb{E}[Y_j]).$$

Using the formula from Leonov & Shiryaev (1959) we can express the cumulant using products of mixed moments

$$\begin{aligned} |\kappa(Y_{i_1} - \mathbb{E}[Y_{i_1}], ..., Y_{i_s} - \mathbb{E}[Y_{i_s}], Y_j - \mathbb{E}[Y_j])| &= \sum_{\pi} (|\pi| - 1)! \prod_{B \in \pi} \mathbb{E}\Big(\prod_{\ell \in B} |Y_\ell - \mathbb{E}[Y_\ell]| \Big) \\ &\leq \big(\sum_{\pi} (|\pi| - 1)! \big) \prod_{\ell=1}^{s+1} (\mathbb{E}[|Y_{B_\ell} - \mathbb{E}[Y_{B_\ell}]|^{s+1}])^{\frac{1}{s+1}}, \end{aligned} \tag{5.47}$$

where $\sum_\pi$ represents the sum over all partitions of $\{i_1, ..., i_s, j\}$ and B_ℓ is the ℓ-th element of $\{i_1, ..., i_s, j\}$.

From Theorem 2.15, we know that the distribution with density (5.30) satisfies the logarithmic Sobolev inequality with log-Sobolev constant $(1 + \frac{t^2}{\sigma_t^2})^{-1}$. By Ledoux (2001, Lemma 1.2) we know that a bounded perturbation will again only impact the log-Sobolev constant by an exponential term. Thus, we can bound the log-Sobolev constant of the distribution of $Y^{x,t}$ by

$$\lambda \leq e^{4L} \Big(1 + \frac{t^2}{\sigma_t^2} \Big)^{-1}. \tag{5.48}$$

Now we can use Theorem 2.26 to obtain

$$\mathbb{P}(|Y_i - \mathbb{E}[Y_i]| \geq r) \leq 2e^{-\frac{r^2}{2\lambda}}.$$

Now we can bound the $s+1$-th moment along the lines of Vershynin (2018, Proposition 2.5.2) via

$$\begin{aligned} \mathbb{E}[|Y_i - \mathbb{E}[Y_i]|^{s+1}] &= \int_0^\infty \mathbb{P}(|Y_i - \mathbb{E}[Y_i]|^{s+1} \geq u) \, \mathrm{d}u = \int_0^\infty \mathbb{P}(|Y_i - \mathbb{E}[Y_i]| \geq r) r^s (s+1) \, \mathrm{d}r \\ &\leq 2(s+1) \int_0^\infty e^{-\frac{r^2}{2\lambda}} r^s \, \mathrm{d}r = 2^{\frac{s+1}{2}} \lambda^{\frac{s+1}{2}} \Gamma\Big(\frac{s+1}{2} \Big). \end{aligned}$$

Taking the $s+1$-th root leads to the bound

$$\mathbb{E}[|Y_i - \mathbb{E}[Y_i]|^{s+1}]^{\frac{1}{s+1}} \lesssim \lambda^{\frac{1}{2}}.$$

Plugging in the log-Sobolev constant of Y from (5.48), we can control all moments via

$$\mathbb{E}[|Y_{i_j}^{x,t} - \mathbb{E}[Y_{i_j}^{x,t}]|^{s+1}]^{\frac{1}{s+1}} \lesssim \frac{\sigma_t}{\sqrt{\sigma_t^2 + t^2}} \leq \min\Big(1, \frac{\sigma_t}{t}\Big).$$

Thus we can bound (5.47) by

$$\prod_{\ell=1}^{s+1}(\mathbb{E}[|Y_{B_\ell}-\mathbb{E}[Y_{B_\ell}]|^{s+1}])^{\frac{1}{s+1}} \lesssim \min\Big(1,\frac{\sigma_t^{s+1}}{t^{s+1}}\Big).$$

We obtain for

$$\frac{\partial^s}{\partial x_{i_1},...,\partial x_{i_s}} v_t^j(x) \lesssim \mathbb{1}_{s=1}\frac{\sigma_t'}{\sigma_t} + \Big(1-\frac{\sigma_t' t}{\sigma_t}\Big)\Big(\frac{t}{\sigma_t^2}\Big)^s \min\Big(1,\frac{\sigma_t^{s+1}}{t^{s+1}}\Big).$$

Minimizing in σ_t and using $\frac{\sigma_t'}{\sigma_t} = \log(\sigma_{\min})$ leads to

$$\frac{\partial^s}{\partial x_{i_1},...,\partial x_{i_s}} v_t^j(x) \lesssim \log(\sigma_{\min})\sigma_{\min}^{-s+1}. \qquad \square$$

Proof of Lemma 5.31. Define

$$q_{x,t}(y) \coloneqq \frac{\exp\big(-\frac{|x-ty|^2}{2\sigma_t^2}-\frac{|y|^2}{2}-a(y)\big)}{\int \exp\big(-\frac{|x-ty|^2}{2\sigma_t^2}-\frac{|y|^2}{2}-a(y)\big)\,\mathrm{d}y}, \quad w_{x,t}(y) \coloneqq \exp\Big(-\frac{|x-ty|^2}{2\sigma_t^2}-\frac{|y|^2}{2}-a(y)\Big).$$

For $k=1$ we obtain using the dominated convergence theorem, the quotient rule and the derivative of the logarithm

$$\begin{aligned}\frac{\partial}{\partial t}v_t^j(x) &= \frac{\partial}{\partial t}\mathbb{E}_{Y\sim q_{x,t}}[v_t(x|Y)_j]\\ &= \mathbb{E}_{Y\sim q_{x,t}}\Big[\frac{\partial}{\partial t}v_t(x|Y)_j\Big] + \mathbb{E}_{Y\sim q_{x,t}}\Big[v_t(x|Y)_j\frac{\partial}{\partial t}\log(w_{x,t}(Y))\Big]\\ &\quad - \mathbb{E}_{Y\sim q_{x,t}}\Big[v_t(x|Y)_j\Big]\mathbb{E}_{Y\sim q_{x,t}}\Big[\frac{\partial}{\partial t}\log(w_{x,t}(Y))\Big]\\ &= \mathbb{E}_{Y\sim q_{x,t}}\Big[\frac{\partial}{\partial t}v_t(x|Y)_j\Big] + \mathrm{Cov}\Big(v_t(x|Y)_j,\frac{\partial}{\partial t}\log(w_{x,t}(Y))\Big).\end{aligned} \tag{5.49}$$

We can use the same structure to obtain higher derivatives. Thus, all terms occurring in an derivative of v_t of order k are either derivatives of $v_t(x|y)_j$, derivatives of $\log(w_{x,t}(Y))$ (including the functions themselves as 0-th order derivative) or products of these derivatives. Since

$$\begin{aligned}v_t(x|y)_j &= \log(\sigma_{\min})(x_j-ty_j)+y_j,\\ \frac{\partial}{\partial t}v_t(x|y)_j &= \log(\sigma_{\min})y_j,\end{aligned} \tag{5.50}$$

all higher derivatives of $v_t(\cdot|\cdot)$ vanish. For the second term occurring in (5.49), we have that

$$\frac{\partial}{\partial t}\log(w_{x,t}(Y)) = -\frac{\partial}{\partial t}\frac{1}{2}(|x-ty|^2)\sigma_t^{-2} = (\langle x,y\rangle - t|y|^2)\sigma_t^{-2} + \frac{1}{2}|x-ty|^2\cdot 2\log(\sigma_{\min})\sigma_t^{-2}. \tag{5.51}$$

We conclude that

$$\frac{\partial^k}{\partial t^k}\log(w_{x,t}(Y)) \lesssim \mathrm{poly}(x,y,t)\,\mathrm{polylog}(\sigma_{\min}^{-1})\sigma_t^{-2}, \tag{5.52}$$

where $\mathrm{poly}(x,y,t)$ is a again polynomial of finite degree in the components of x,y,t. Thus, bounds that are worse than σ_t^{-2} can only appear when $\frac{\partial}{\partial t}\log(w_{x,t}(Y))$ occurs raised to a power.

Looking at (5.49), we see that the highest power of $\frac{\partial}{\partial t}\log(w_{x,t}(Y))$ increases by 1 each time the derivative is taken via the covariance term. Hence in the k-th derivative, the highest order terms are of the form

$$\mathrm{Cov}\Big(f(Y)\Big(\frac{\partial}{\partial t}\log(w_{x,t}(Y))\Big)^{k-1}, \frac{\partial}{\partial t}\log(w_{x,t}(Y))\Big),$$

where either $f \equiv 1$ or f is of the form (5.50) or (5.52). First, let $f \equiv 1$. Then due to (5.51)

$$\begin{aligned}
&\mathrm{Cov}\Big(\Big(\frac{\partial}{\partial t}\log(w_{x,t}(Y))\Big)^{k-1}, \frac{\partial}{\partial t}\log(w_{x,t}(Y))\Big)\\
&\quad= \mathbb{E}\Big[\Big(\frac{\partial}{\partial t}\log(w_{x,t}(Y)) - \mathbb{E}\Big[\frac{\partial}{\partial t}\log(w_{x,t}(Y))\Big]\Big)^k\Big]\\
&\quad= \sigma_t^{-2k}\mathbb{E}\Big[\Big(\langle(1-t2\log(\sigma_{\min}))x, Y\rangle - \mathbb{E}[\langle(1-t2\log(\sigma_{\min}))x, Y\rangle]\\
&\qquad\qquad + (t^2 2\log(\sigma_{\min}) - t)|Y|^2 - \mathbb{E}[(t^2 2\log(\sigma_{\min}) - t)|Y^2|]\Big)^k\Big]\\
&\quad\lesssim \sigma_t^{-2k}\sum_{i=1}^d (1-t2\log(\sigma_{\min}))^k x^k \mathbb{E}\Big[\big(Y_i - \mathbb{E}[Y_i]\big)^k\Big]\\
&\qquad + \sum_{i=1}^d (t^2 2\log(\sigma_{\min})t - t)^k \mathbb{E}\Big[\big(Y_i^2 - \mathbb{E}[Y_i^2]\big)^k\Big],
\end{aligned}$$

where all expectations are taken with respect to $Y \sim q_{x,t}$. From the proof of Lemma 5.30 we know that

$$\mathbb{E}\Big[\big(Y_i - \mathbb{E}[Y_i]\big)^k\Big] \lesssim \Big(\frac{\sigma_t^2}{t^2+\sigma_t^2}\Big)^{\frac{k}{2}} \leq \min\Big(\frac{\sigma_t^2}{t^2}, 1\Big)^{\frac{k}{2}} = \min\Big(\frac{\sigma_t^k}{t^k}, 1\Big).$$

By Theorem 2.22 we know that if $Y_i - \mathbb{E}[Y_i]$ is subgaussian, which was shown in Lemma 5.30, then $(Y_i - \mathbb{E}[Y_i])^2$ is subexponential with the same order of subgaussian and subexponential norm.

$$\mathbb{E}\Big[\big(Y_i^2 - \mathbb{E}[Y_i^2]\big)^k\Big] = \mathbb{E}\Big[\big((Y_i - \mathbb{E}[Y_i])^2 - \mathbb{E}[(Y_i - \mathbb{E}[Y_i])^2] + 2\mathbb{E}[Y_i](Y_i - \mathbb{E}[Y_i])\big)^k\Big]$$

Note that $\mathbb{E}[Y_i] \lesssim \mathrm{polylog}(n)$ as $x \in [-\log(n), \log(n)]^d$. Hence $Y_i^2 - \mathbb{E}[Y_i^2]$ can be represented by the sum of a subexponential and a subgaussian variable. Furthermore, if $Y_i - \mathbb{E}[Y_i]$ is subgaussian then it is also subexponential and the subexponential norm differs only by a constant from the subgaussian norm. From Theorem 2.21 we learn that the sum of subexponential random variables is again subexponential and that we can bound the subexponential norm of the sum by the sum of the individual subexponential norms. Using that the subgaussian norm of $Y_i - \mathbb{E}[Y_i]$ is of order $\frac{\sigma_t}{\sqrt{\sigma_t^2+t^2}}$, we obtain using Theorem 2.20

$$\mathbb{E}\Big[\big(Y_i^2 - \mathbb{E}[Y_i^2]\big)^k\Big] \lesssim \Big(\frac{\sigma_t^2}{t^2+\sigma_t^2}\Big)^{\frac{k}{2}} \leq \min\Big(\frac{\sigma_t^k}{t^k}, 1\Big).$$

Now we consider the case that f is of the form (5.50) or (5.52). First, we note that as shown in the proof of Theorem 5.18, the distribution with density $q_{x,t}$ satisfies the Poincaré inequality with a constant

$$\rho \leq e^{2L}\Big(1 + \frac{t^2}{\sigma_t^2}\Big)^{-1}.$$

An immediate consequence of the Poincaré inequality, defined in Definition 2.12, is the following

bound for all smooth functions g, h

$$\operatorname{Cov}(g(Y), h(Y)) \le \rho\sqrt{\mathbb{E}[|\nabla g(Y)|^2]\mathbb{E}[|\nabla h(Y)|^2]}. \tag{5.53}$$

We are going to use this to bound

$$\operatorname{Cov}\Big(f(Y)\Big(\frac{\partial}{\partial t}\log(w_{x,t}(Y))\Big)^{k-1}, \frac{\partial}{\partial t}\log(w_{x,t}(Y))\Big).$$

Define

$$g(y) \coloneqq f(y)\Big(\frac{\partial}{\partial t}\log(w_{x,t}(y))\Big)^{k-1}.$$

Then

$$\nabla g(y) = (\nabla f(y))\Big(\frac{\partial}{\partial t}\log(w_{x,t}(y))\Big)^{k-1} + f(y)(k-1)\nabla\Big(\frac{\partial}{\partial t}\log(w_{x,t}(y))\Big)^{k-2}\Big(\nabla\frac{\partial}{\partial t}\log(w_{x,t}(y))\Big)$$

and

$$\begin{aligned}
|\nabla g(y)|^2 &\lesssim |\nabla f(y)|^2\Big|\frac{\partial}{\partial t}\log(w_{x,t}(y))\Big|^{2(k-1)} \\
&\quad + |f(y)|^2(k-1)^2\Big|\Big(\frac{\partial}{\partial t}\log(w_{x,t}(y))\Big)\Big|^{2(k-2)}\Big|\nabla\frac{\partial}{\partial t}\log(w_{x,t}(y))\Big|^2 \\
&\lesssim |\nabla f(y)|^2\Big|\sigma_t^{-4(k-1))}|x-ty|^{2(k-1)}|y|^{2(k-1)}\Big| && (5.54)\\
&\quad + \Big|\sigma_t^{-4(k-1)}\frac{1}{2}|x-ty|^{4(k-1)}(2\log(\sigma_{\min}))^{2(k-1)}\Big| && (5.55)\\
&\quad + |f(y)|^2\Big(\Big|\sigma_t^{-4(k-2)}|x-ty|^{2(k-2)}|y|^{2(k-2)}\Big| && (5.56)\\
&\qquad + \Big|\sigma_t^{-4(k-2)}\frac{1}{2^{2(k-1)}}|x-ty|^{4(k-2)}(2\log(\sigma_{\min}))^{2(k-2)}\Big|\Big) && (5.57)\\
&\qquad \cdot |\sigma_t^{-2}(x-2ty+t(x-ty)2\log(\sigma_{\min}))|^2. && (5.58)
\end{aligned}$$

All possible functions f are polynomials in y, hence the derivatives are also polynomials. As $v_t(\cdot|\cdot)$ is linear in y_i and $\frac{\partial}{\partial t}\log(w_{x,t})$ is quadratic in the components, the derivatives are of that degree or lower. We denote this by $\operatorname{poly}_2(x,y,t)$. In case f is of the form (5.52) ∇f and f are of order σ_t^{-2}, else the dependency on $\sigma_{\min}$ or σ_t is only logarithmically. For (5.54)

$$\begin{aligned}
&\mathbb{E}[|\nabla f(Y)|^2\sigma_t^{-4(k-1)}|x-tY|^{2(k-1)}|Y|^{2(k-1)}] \\
&= \int \operatorname{polylog}(\sigma_{\min})\operatorname{poly}_2(x,y,t)\sigma_t^{-4(k-1))}|x-ty|^{2(k-1)}|y|^{2(k-1)} \\
&\quad \cdot \frac{\exp\big(-\frac{|x-ty|^2}{2\sigma_t^2}-\frac{|y|^2}{2}-a(y)\big)}{\int\exp\big(-\frac{|x-ty|^2}{2\sigma_t^2}-\frac{|y|^2}{2}-a(y)\big)\,\mathrm{d}y}\,\mathrm{d}y \\
&\le \frac{e^{2L}\operatorname{polylog}(\sigma_{\min})}{\int\exp\big(-\frac{|z|^2}{2}-\frac{|\frac{x-\sigma_t z}{t}|^2}{2}\big)\,\mathrm{d}y} \\
&\quad \cdot \int \operatorname{poly}_2\Big(x,\frac{x-\sigma z}{t},t\Big)\sigma_t^{-4(k-1)}|\sigma_t z|^{2(k-1)}\Big|\frac{x-\sigma_t z}{t}\Big|^{2(k-1)}\exp\Big(-\frac{|z|^2}{2}-\frac{|\frac{x-\sigma_t z}{t}|^2}{2}\Big)\,\mathrm{d}y \\
&\lesssim \operatorname{polylog}(\sigma_{\min})\sigma_t^{-2k-2}|x|^{2(k-1)},
\end{aligned}$$

where the last inequality stems from similar arguments as in the proof of Theorem 5.18 and Theorem 5.20 for higher order moments. The exact calculations are omitted for the sake of brevity. The same calculation for the other terms in (5.55), (5.57), (5.56) and (5.58) yield

$$\mathbb{E}[|\nabla g(y)|^2] \lesssim \operatorname{polylog}(\sigma_{\min}) \log(n)^{4(k-1)} \sigma_t^{-2k-4},$$

where we used that $x \in [-\log(n), \log(n)]^d$ to combine the worst case dependencies. For h, we obtain in similar manner

$$\mathbb{E}[|\nabla h(y)|^2] \lesssim \sigma_t^{-4} \log(n)^4.$$

Inserting these bounds in (5.53), we obtain

$$\operatorname{Cov}(f(Y)(\frac{\partial}{\partial t} \log(w_{x,t}(Y)))^{k-1}, \frac{\partial}{\partial t} \log(w_{x,t}(Y))) \lesssim \operatorname{polylog}(\sigma_{\min}) \operatorname{polylog}(n) \frac{\sigma_t^2}{\sigma_t^2 + t^2} \sigma_t^{-k-4}.$$

Hence in all cases, the terms occurring in the k-th derivative are bounded by

$$\frac{\partial^k}{\partial t^k} v_t^j(x) \lesssim \operatorname{polylog}(\sigma_{\min}) \operatorname{polylog}(n) \sigma_{\min}^{-k-2}.$$

This concludes the proof. □

Helper results

Lemma 5.32. *Let $\sigma_t = \sigma_{\min}^t$ with $\sigma_{\min} \in (0,1)$. Then for $t \in (0,1)$*

$$\frac{t}{t^2 + \sigma_t^2} \leq \max(\log(\sigma_{\min}^{-1}), e^2).$$

Proof. First, let $t \geq \frac{1}{\log(\sigma_{\min}^{-1})}$. Then

$$\frac{t}{t^2 + \sigma_t^2} \leq \frac{1}{t} \leq \log(\sigma_{\min}^{-1}).$$

If $t < \frac{1}{\log(\sigma_{\min}^{-1})}$, then

$$\sigma_{\min}^{2t} = \exp(2 \log(\sigma_{\min}) t) \geq \exp(-2)$$

and thus

$$\frac{t}{t^2 + \sigma_t^2} \leq \frac{t}{t^2 + \exp(-2)} \leq \exp(2).$$ □

Lemma 5.33. *For p^* of the form (5.20) and any Lipschitz 1 function $f \colon \mathbb{R}^d \to \mathbb{R}$,*

$$\Big| \int \nabla f(z) p^*(x - z) \mathrm{d}z - \int \nabla f(z) p^*(y - z) \, \mathrm{d}z \Big| \lesssim |x - y|. \tag{5.59}$$

Proof. Using Hölders inequality we obtain

$$\Big| \int \nabla f(z) p^*(x - z) \mathrm{d}z - \int \nabla f(z) p^*(y - z) \, \mathrm{d}z \Big| \leq \|\nabla f\|_\infty \int |p^*(x - z) - p^*(y - z)| \, \mathrm{d}z$$

$$\leq \sqrt{d} \int \Big| \int_0^1 \langle x - y, \nabla p^*(x + t(y - x) - z) \rangle \, \mathrm{d}t \Big| \, \mathrm{d}z,$$

where the last inequality follows from the Lipschitz bound on f and a backwards-application of the multivariate chain rule. The ∇-operator is used with respect to the $\mathbb{R}^d$ valued input of p^*. Using the Cauchy-Schwarz inequality and changing the order of integration, we conclude

$$\begin{aligned}\int \Big| \int_0^1 \langle x - y, \nabla p^*(x + t(y - x) - z) \rangle \Big| \, \mathrm{d}z &\leq |x - y| \int \int_0^1 |\nabla p^*(x + t(y - x) - z)| \, \mathrm{d}t \, \mathrm{d}z \\ &= |x - y| \int_0^1 \int |\nabla p^*(w)| \, \mathrm{d}w \, \mathrm{d}t.\end{aligned}$$

We can bound the integral over the derivative via

$$\begin{aligned}\int |\nabla p^*(w)| \, \mathrm{d}w &= \int | - w - \nabla a(w)| \frac{\exp(-\frac{-|w|^2}{2} - a(w))}{\int \exp(-\frac{-|w|^2}{2} - a(w)) \, \mathrm{d}w} \, \mathrm{d}w \\ &\leq \mathbb{E}_{W \sim p^*}[|W|] + \sqrt{d}L.\end{aligned}$$

We can further bound

$$\mathbb{E}_{W \sim p^*}[|W|] \leq e^{2L} \mathbb{E}_{W \sim \mathcal{N}(0, I_d)}[|X|] \leq e^{2L} \sqrt{\mathbb{E}_{W \sim \mathcal{N}(0, I_d)}[|X|^2]} = e^{2L} \sqrt{d}.$$

Thus

$$\Big| \int \nabla f(z) p^*(x - z) \mathrm{d}z - \int \nabla f(z) p^*(y - z) \, \mathrm{d}z \Big| \leq d(e^{2L} + L)|x - y|. \qquad \square$$

Chapter 6

Conditional distribution estimation

In the previous chapters, we always estimated a distribution based on samples. In this chapter, we are going to assume that we have some additional information. Based on a sample containing pairs of observations and additional information, we want to estimate the conditional distribution. We are going to assume that the target is a d_Y-dimensional random variable Y and the covariable X is d_X-dimensional. This notation is in line with the literature, especially in the context of distributional regression. Note that the notation differs from the introduction, where we used X for the target.

Our goal is to estimate the conditional distribution $\mathbb{P}^*_{Y|X=x}$ for $x \in \mathbb{R}^{d_X}$ based on an i.i.d. sample $(X_1, Y_1), ..., (X_n, Y_n)$ of the joint distribution $\mathbb{P}^*_{X,Y}$ and evaluate this estimation using proper scoring rules, which are well established in the forecasting literature and specifically in the application in weather prediction, see for example Gneiting & Raftery (2007); Hersbach (2000); Rasp & Lerch (2018).

A variety of methods exist to estimate a distribution. Just like in the unconditional case, the superordinate question is which object should be estimated. In Section 4.1, we studied an estimator of the density, Chapter 3 and Chapter 5 focused on the estimation of a pushforward map that directly enables sampling from an estimated distribution. Another approach is to estimate the distribution function. Of course, these methods are closely connected, the ability to generate samples from an estimated distribution enables further estimates of either the density of the distribution function.

One common approach to estimating a conditional distribution follows immediately from the KDE discussed in Section 4.1. In case the joint distribution admits a density p with respect to the $d_Y + d_X$-dimensional Lebesgue measure, the conditional density is given by

$$p(y|x) = \frac{p(x,y)}{p_X(x)}.$$

where p_X is the marginal density in X. Estimating both, the joint and the marginal density using a KDE, with kernels $K^{\mathrm{y}}_{h_y}$ and $K^{\mathrm{x}}_{h_x}$ and bandwidths $h_y, h_x > 0$ respectively, we obtain the following estimator of the conditional density

$$\mathring{p}(y|x) = \frac{\sum_{i=1}^n K^{\mathrm{y}}_{h_y}(y - Y_i) K^{\mathrm{x}}_{h_x}(x - X_i)}{\sum_{i=1}^n K^{\mathrm{x}}_{h_x}(x - X_i)}, \tag{6.1}$$

which we call the *Nadaraya–Watson conditional kernel density estimator* (NW estimator), named after Nadaraya (1964); Watson (1964), who themselves considered the mean regression problem. A classical approach to model other characteristics than the mean is quantile regression (Koenker & Bassett, 1978) and advancements thereof, see for example Chernozhukov et al. (2010). The work of Yu & Jones (1998) motivated Hall et al. (1999), who proposed an estimator very close to (6.1). Asymptotic properties of (6.1) have been studied by Hyndman et al. (1996) with respect to the L_2 distance. Efromovich (2007) obtained anisotropic minimax rates for conditional density estimators in the L_2 distance. Li et al. (2022) studied minimax properties of a histogram-typed conditional density estimator in the total variation distance. The setting where for one set of covariates several observations of the target are available has been investigated by Bott & Kohler (2017).
One recent example for estimating the conditional distribution function in case of one-dimensional targets is isotonic distributional regression (Henzi et al., 2021). An even more recent approach, which also applies to $d_Y > 1$ is Engression (Shen & Meinshausen, 2025), who minimize the energy score using a generative model that approximates the pushforward directly.

Own contribution We are going to combine the concepts of Flow Matching and conditional distribution estimation. Initially, this results in a special case of a guided Flow Matching model (Zheng et al., 2023). While this model suffers from theoretical drawbacks, we introduce a minor adaptation that resolves this issue. Next, we demonstrate that this model is naturally connected to the NW estimator. We also show that there is a vector field that generates the NW estimator for fixed $x \in \mathbb{R}^{d_X}$.
Then we consider the evaluation of the NW estimator in the Fourier score. As noted in Section 2.3, the Fourier score is a scoring function that generalizes the energy score, which itself is the multivariate extension of the CRPS. In the beginning, we connect the classical concept of risk from statistical learning to proper scoring rules. Since there are, to the best of the author's knowledge, no comparable results besides for the one-dimensional case (Pic et al., 2023), we first derive a lower bound to assess our results. Then we derive an anisotropic rate of convergence for the NW estimator that matches the lower bound and is thus minimax optimal up to a logarithmic factor. Subsequently, we use the rate of convergence obtained for the NW estimator to derive a rate of convergence for the Flow Matching model in conditional distribution estimation. In the end we apply the Flow Matching conditional distribution estimator to forecasting datasets and weather prediction. Our results indicate that the estimator can keep up with state-of-the-art methods.

6.1. Flow Matching as a conditional distribution estimator

On the population level, we can extend the Flow Matching model as introduced in Chapter 5 directly.
To estimate a conditional density, the covariates are employed as additional arguments to the vector field and the density path. For fixed $y \in \mathbb{R}^{d_Y}$, let $v_t(\cdot|y)$ be a vector field that generates

$p_t(\cdot|y)$. For fixed $x \in \mathbb{R}^{d_X}$, define

$$p_{t,x}(z) = \int p_t(z|y)p^*(y|x)\,\mathrm{d}y, \qquad \text{and} \qquad v_{t,x}(z) = \int v_t(z|y)\frac{p_t(z|y)p^*(y|x)}{p_{t,x}(z)}\,\mathrm{d}y,$$

where $p^*(\cdot|x)$ is the density of $\mathbb{P}^*_{Y|X=x}$. To avoid notation conflicts and simultaneously maintain comparability with the previous chapter, we will use the lower case x to indicate the dependency on $x \in \mathbb{R}^{d_X}$. For fixed x the equivalence of

$$\mathbb{E}_{\substack{t\sim\mathcal{U}[0,1]\\ Z_t\sim p_{t,x}}}\big[|\tilde{v}_t(Z_t,x) - v_{t,x}(Z_t)|^2\big], \qquad \text{and} \qquad \mathbb{E}_{\substack{t\sim\mathcal{U}[0,1]\\ Y\sim p^*(\cdot|x)\\ Z_t\sim p_t(\cdot|Y)}}\left[\,|\tilde{v}_t(Z_t,x) - v_t\left(Z_t|Y\right)|^2\,\right]$$

follows in exactly the same way as in Lipman et al. (2023, Theorem 2). This adaptation is also shown in Zheng et al. (2023).

In practice, the true conditional density $p^*(\cdot|x)$ is unknown and hence inaccessible. Additionally, training a new model for every x of interest is computationally expensive. To learn $\tilde{v}$ simultaneously for the entire covariate space, the following empirical counterpart can be implemented

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\substack{t\sim\mathcal{U}[0,1]\\ Z_t\sim p_t(\cdot|Y_i)}}\left[\,|\tilde{v}_t(Z_t,X_i) - v_t\left(Z_t|Y_i\right)|^2\,\right]. \tag{6.2}$$

This corresponds to the adaptation made in conditional diffusions, see Tang et al. (2025).

The following result shows that analogous to Lipman et al. (2023, Theorem 2), we still obtain an equivalent optimization problem, however the random variable Z_t has no density with respect to the Lebesgue measure anymore and the vector field is only nonzero on a finite set. In slight abuse of notation, we identify a point $x \in \mathbb{R}^{d_X}$ with the set containing only this point.

Lemma 6.1. *Let* $\bar{p}\colon [0,1]\times\mathbb{R}^{d_Y}\times\mathbb{R}^{d_X}\to\mathbb{R}_{\geq 0}$, $\bar{v}_t\colon [0,1]\times\mathbb{R}^{d_Y}\times\mathbb{R}^{d_X}\to\mathbb{R}^{d_Y}$ *such that*

$$\bar{p}_t(z,x) \coloneqq \frac{1}{n}\sum_{i=1}^{n} p_t(z|Y_i)\delta_{X_i}(x), \quad \bar{v}_t(z,x) \coloneqq \frac{\sum_{i=1}^{n} v_t(z|Y_i)p_t(z|Y_i)\delta_{X_i}(x)}{\sum_{i=1}^{n} p_t(z|Y_i)\delta_{X_i}(x)},$$

where the functions are set to 0 *in case of* $\frac{0}{0}$. *Then for a fixed, measurable function* $\tilde{v}\colon [0,1]\times\mathbb{R}^{d_Y}\times\mathbb{R}^{d_X}\to\mathbb{R}^{d_Y}$

$$\mathbb{E}_{\substack{t\sim\mathcal{U}[0,1]\\ Z_t,X\sim\bar{p}_t}}\left[\,|\tilde{v}_t(Z_t,X) - \bar{v}_t\left(Z_t,X\right)|^2\,\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\substack{t\sim\mathcal{U}[0,1]\\ Z_t\sim p_t(\cdot|Y_i)}}\left[\,|\tilde{v}_t(Z_t,X_i) - v_t\left(Z_t|Y_i\right)|^2\,\right] + C,$$

where C *is a constant independent of* $\tilde{v}$.

The objective in (6.2) cannot capture any regularity in the influence of the covariates. Nevertheless, optimal rates of convergence profit from higher regularity in the covariates. To allow for rates exploiting this kind of smoothness, we adapt the objective function using a kernel

function $K^{\mathrm{x}}_{h_x} : \mathbb{R}^{d_X} \to \mathbb{R}_{\geq 0}$, where $h_x > 0$ is the bandwidth,

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\substack{t\sim\mathcal{U}[0,1]\\ Z_t\sim p_t(\cdot|Y_i)\\ X\sim K^{\mathrm{x}}_{h_x}(\cdot - X_i)}} \left[\, |\tilde{v}_t(Z_t, X) - v_t\left(Z_t|Y_i\right)|^2 \,\right]. \tag{6.3}$$

Note that $K^{\mathrm{x}}_{h_x}$ does not depend on t. For $h_x \to 0$ we obtain the objective (6.2), hence (6.3) generalizes (6.2). Analogously to Lemma 6.1 we obtain the following equivalence.

Lemma 6.2. *Let $p\colon [0,1] \times \mathbb{R}^{d_Y} \times \mathbb{R}^{d_X} \to \mathbb{R}_{\geq 0}, v\colon [0,1] \times \mathbb{R}^{d_Y} \times \mathbb{R}^{d_X} \to \mathbb{R}^{d_Y}$ such that*

$$p_t(z,x) \coloneqq \frac{1}{n}\sum_{i=1}^{n} p_t(z|Y_i) K^x_{h_x}(x - X_i), \quad v_t(z,x) \coloneqq \frac{\sum_{i=1}^n v_t(z|Y_i) p_t(z|Y_i) K^x_{h_x}(x - X_i)}{\sum_{i=1}^n p_t(z|Y_i) K^x_{h_x}(x - X_i)},$$

where the functions are set to 0 in case of $\frac{0}{0}$. Then for a fixed, measurable function $\tilde{v}\colon [0,1] \times \mathbb{R}^{d_Y} \times \mathbb{R}^{d_X} \to \mathbb{R}^{d_Y}$

$$\mathbb{E}_{\substack{t\sim\mathcal{U}[0,1]\\ Z_t, X\sim p_t}} \left[\, |\tilde{v}_t(Z_t, X) - v_t\left(Z_t, X\right)|^2 \,\right] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\substack{t\sim\mathcal{U}[0,1]\\ Z_t\sim p_t(\cdot|Y_i)\\ X\sim K^x_{h_x}(\cdot - X_i)}} \left[\, |\tilde{v}_t(Z_t, X) - v_t\left(Z_t|Y_i\right)|^2 \,\right] + C,$$

where C is a constant independent of $\tilde{v}$.

To obtain a generative estimator for the conditional distribution, we first fix a latent distribution $\mathbb{U}$ on $\mathbb{R}^{d_Y}$ that admits a density with respect to the Lebesgue measure.
Let $\mathcal{M}$ be some function class such that the following minimizing argument exists and all corresponding ODEs have a unique solution. Then we choose

$$\hat{v} \in \operatorname*{argmin}_{\tilde{v}\in\mathcal{M}} \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\substack{t\sim\mathcal{U}[0,1]\\ Z_t\sim p_t(\cdot|Y_i)\\ X\sim K^{\mathrm{x}}_{h_x}(\cdot - X_i)}} \left[\, |\tilde{v}_t(Z_t, X) - v_t\left(Z_t|Y_i\right)|^2 \,\right]. \tag{6.4}$$

For fixed $x \in \mathbb{R}^{d_X}$, we solve the ODE

$$\frac{\partial \psi_{t,x}(y)}{\partial t} = \hat{v}_t(\psi_{t,x}(y), x), \quad \psi_{0,x}(y) = y.$$

Using the solution $\hat{\psi}$ of this ODE to push forward a random variable $\zeta \sim \mathbb{U}$ we obtain for $t \in [0,1]$

$$\hat{\psi}_{t,x}(\zeta) \sim \hat{p}_{t,x}. \tag{6.5}$$

In line with Section 5.2, we now take a closer look at the functions that are approximated when minimizing (6.4). In case the kernel $K^{\mathrm{x}}_{h_x}$ is a density, integration over z and x shows that p_t from Lemma 6.2 itself is a joint density. For the conditional density, we obtain

$$p_{t,x}(z) = \frac{p_t(z,x)}{\int p_t(z,x)\,\mathrm{d}z} = \frac{\sum_{i=1}^n p_t(z|Y_i) K^{\mathrm{x}}_{h_x}(x - X_i)}{\sum_{i=1}^n K^{\mathrm{x}}_{h_x}(x - X_i)}.$$

For suitable $p_t(\cdot|Y_i)$, $p_{1,x}(z)$ coincides with the estimator in (6.1). Thus in the following, we

set $p_{t,x}(\cdot) = \mathring{p}_t(\cdot|x)$ for every $x \in \mathbb{R}^{d_X}$. Specifically, we can use a standard notation for the conditional density without causing a notation conflict. In case $t = 1$, we omit the subscript, if this causes no confusion.
The next result shows that for fixed $x \in \mathbb{R}^{d_X}$ the vector field in Lemma 6.2 generates a probability path $\mathring{p}_t(\cdot|x)$ such that $\mathring{p}_1(\cdot|x)$ is the NW estimator.

Lemma 6.3. *For every $y \in \mathbb{R}^{d_Y}$, let $v_t(\cdot|y)$ be such that it generates $p_t(\cdot|y)$. Define*

$$\mathring{p}_t(z|x) = \frac{\sum_{i=1}^n p_t(z|Y_i) K_{h_x}^x(x - X_i)}{\sum_{i=1}^n K_{h_x}^x(x - X_i)}, \quad \mathring{v}_t(y|x) = v_t(z, x).$$

Then for every $x \in \mathbb{R}^{d_X}$ the vector field $\mathring{v}_t(\cdot|x)$ generates $\mathring{p}_t(\cdot|x)$.

Thus, for fixed $x \in \mathbb{R}^p$, the solution $\mathring{\psi}_{t,x}$ of the ODE

$$\frac{\partial \mathring{\psi}_{t,x}(z)}{\partial t} = \mathring{v}_t(\mathring{\psi}_{t,x}(z)|x), \quad \mathring{\psi}_{0,x}(z) = z$$

used as a pushforward leads to

$$\mathring{\psi}_{t,x}(\zeta) \sim \mathring{p}_t(\cdot|x). \tag{6.6}$$

Hence the Flow Matching estimator (6.5) is naturally connected to the NW estimator.

6.2. Proper scoring rules and risk

In this section, we shortly connect the concepts of risk and proper scoring rules. To this end, we recall from Section 2.3 that a scoring rule is a function $S\colon \mathcal{P} \times \mathbb{R}^{d_Y} \to \overline{\mathbb{R}}$ defined over a class $\mathcal{P}$ of distributions that fulfills the further requirements of Section 2.3. We assume that the true conditional distribution function $\mathbb{P}_{Y|X=x}$ is a member of $\mathcal{P}$ for every $x \in \mathbb{R}^{d_X}$.
Based on a test sample $(\bar{X}_j, \bar{Y}_j)_{j=1,\dots,m}$ following the same distribution as (X_i, Y_i), the performance of an estimator $\hat{\mathbb{P}}_{Y|X}$ with respect to the scoring rule S can be evaluated via

$$R_m(\hat{\mathbb{P}}^{Y|X}) := \frac{1}{m} \sum_{j=1}^m S(\hat{\mathbb{P}}_{Y|\bar{X}_j}, \bar{Y}_j).$$

If $(\bar{X}_j, \bar{Y}_j)_{j=1,\dots,m}$ are identically distributed, the population counterpart to $R_m(\hat{\mathbb{P}}_{Y|X})$ is given by

$$\begin{aligned}
R(\hat{\mathbb{P}}_{Y|X}) &= \mathbb{E}_{(\bar{X}_j, \bar{Y}_j)}[R_m(\hat{\mathbb{P}}_{Y|X})] \\
&= \mathbb{E}_{(\bar{X}_j, \bar{Y}_j)}\Big[\frac{1}{m} \sum_{j=1}^m S(\hat{\mathbb{P}}_{Y|\bar{X}_j}, \bar{Y}_j)\Big] \\
&= \mathbb{E}_{(\bar{X}_1, \bar{Y}_1)}\big[S(\hat{\mathbb{P}}_{Y|\bar{X}_1}, \bar{Y}_1)\big] \\
&= \mathbb{E}_{(\bar{X}_1, \bar{Y}_1)}\big[S(\hat{\mathbb{P}}_{Y|\bar{X}_1}, \bar{Y}_1) - S(\mathbb{P}_{Y|\bar{X}_1}, \bar{Y}_1)\big] + \mathbb{E}_{(\bar{X}_1, \bar{Y}_1)}\big[S(\mathbb{P}_{Y|\bar{X}_1}, \bar{Y}_1)\big] \\
&= \mathbb{E}_{\bar{X}_1}\big[\mathbb{E}_{\bar{Y}_1|\bar{X}_1}\big[S(\hat{\mathbb{P}}_{Y|\bar{X}_1}, \bar{Y}_1) - S(\mathbb{P}_{Y|\bar{X}_1}, \bar{Y}_1)\big]\big] + \mathbb{E}_{\bar{X}_1}\big[\mathbb{E}_{\bar{Y}_1|\bar{X}_1}\big[S(\mathbb{P}_{Y|\bar{X}_1}, \bar{Y}_1)\big]\big].
\end{aligned} \tag{6.7}$$

This is exactly the decomposition into divergence and entropy function presented in Section 2.3 integrated over $\bar{X}_1$. The entropy function does not depend on the estimator $\hat{\mathbb{P}}_{Y|X}$.
In case S is proper, we can now connect the excess risk to the divergence function: the target of any estimation method $\hat{\mathbb{P}}$, where we suppress the dependency to indicate the integration over the conditional variable, is

$$\mathcal{E}(\hat{\mathbb{P}}) := R(\hat{\mathbb{P}}) - \min_{\mathbb{P}\in\mathcal{P}} R(\mathbb{P}) = \mathbb{E}_{\bar{X}_1}\left[\mathbb{E}_{\bar{Y}_1|\bar{X}_1}\left[S(\hat{\mathbb{P}}_{Y|\bar{X}_1}, \bar{Y}_1) - S(\mathbb{P}_{Y|\bar{X}_1}, \bar{Y}_1)\right]\right] \tag{6.8}$$

where the minimum is taken over all $\mathbb{P} \in \mathcal{P}$. In case of a strictly proper scoring rule, the minimal argument is unique.

6.3. Rate of convergence in Fourier score

In this section, we want to analyze both, the NW-estimator and the Flow Matching estimator for conditional distribution estimation in the risk associated with the Fourier score.
Since the risk associated with the Fourier score has, to the best of the author's knowledge, not been studied before, a natural first step is to establish a lower bound on this rate for any estimator. To this end, we introduce the smoothness class that is considered in the entire section. To measure the smoothness of the conditional density, we first define the fractional Sobolev ellipsoid on $(0,1)^{d_Y}$ for some $s \in \mathbb{R}$

$$H^s(\Gamma) := \Big\{ f\colon (0,1)^{d_Y} \to \mathbb{R} \,\Big|\, \int_{(0,1)^d} (1+|u|^2)^s |\mathcal{F}f(u)|^2 \,\mathrm{d}u \le \Gamma \Big\}. \tag{6.9}$$

Further, we define the fractional Sobolev norm by

$$\|f\|_{H^s(\mathbb{R}^d)} = \Big(\int_{(0,1)^d} \left(1+|u|^2\right)^s |\mathcal{F}f(u)|^2 \,\mathrm{d}u \Big)^{1/2} \tag{6.10}$$

for some $\Gamma > 0$. Note that for $s \in \mathbb{N}$, this space coincides with the Sobolev space $W^{s,2}$ defined in Section 2.1. Using this, we define the set $\mathcal{P}_{s,\alpha}$ as the set of all distributions $\mathbb{P}_{X,Y}$ on $(0,1)^{d_X+d_Y}$ that admit densities with respect to the Lebesgue measure, for the conditional densities it holds that $p_{Y|X=x} \in H^s(\Gamma)$ for all $x \in \mathbb{R}^{d_X}$ and for an $\alpha \in (0,1]$

$$\sup_{x,x'\in\mathcal{X}, x\neq x'} \frac{\|\varphi_x - \varphi_{x'}\|_\gamma}{|x-x'|^\alpha} \le L.$$

6.3.1. Lower bound

For this class $\mathcal{P}_{s,\alpha}$ we obtain a lower bound for the estimation of the conditional distribution in Fourier score.

Theorem 6.4. *If the marginal density p_X is lower bounded by some constant $c > 0$, then we*

have

$$\inf_{\hat{f}_n} \sup_{f \in \mathcal{P}_{s,\alpha}} \mathbb{E}\Big[\int \int \frac{|\hat{\varphi}_x(u) - \varphi_x(u)|^2}{|u|^\gamma} \, \mathrm{d}u \, p_X(x) \, \mathrm{d}x \Big] \gtrsim \begin{cases} n^{-\frac{2}{\frac{d_X}{\alpha}+2}}, & \gamma \geq d_Y, \\ n^{-\frac{2+\frac{\gamma}{s}}{2+\frac{d_Y}{s}+\frac{d_Y}{\alpha}+\frac{d_X\gamma}{2\alpha s}}}, & \gamma < d_Y. \end{cases} \tag{6.11}$$

The proof follows the classical structure of Tsybakov (2009) for proving nonparametric lower bounds.

We particularly observe that in case of $\gamma > d$, the lower bound coincides with the lower bounds for the mean regression problem evaluated in the weighted L_2 distance, see Györfi (2002, Theorem 3.2). In case of $\gamma < d$ the lower bound on the rate is sensitive to smoothness in the target space and the covariate space and is hence anisotropic. In case of $d_Y = 1$, we recover the same lower bound as Pic et al. (2023), thus our result directly extends their work to higher dimensions and other scales γ.

6.3.2. Upper bound for the NW estimator

Next, we want derive a rate of convergence for the NW estimator in the risk associated to the Fourier score. We start from the definition of the NW estimator in (6.1) and denote the kernel in the target variable by K, which will ease notation in the proof. In order to enable transfer to the Flow Matching estimator, we denote the bandwidth of this kernel by $\sigma_{\min} > 0$. We denote the characteristic function of the NW density estimator $\mathring{p}(\cdot|x)$ for $x \in \mathbb{R}^{d_X}$ with $\mathring{\varphi}_x$.

Further, we impose an order assumption on the kernel K. To this end, we first define a kernel of order ℓ.

Definition 6.5. *Let $\ell \in \mathbb{N}$. A function $K \colon \mathbb{R}^d \to \mathbb{R}$ with $\|K\|_1 < \infty$ is a kernel of order ℓ if the functions $u \mapsto u^j K(u)$, where j is a multiindex such that $1 \leq |j| \leq \ell - 1$, are absolutely integrable with respect to the Lebesgue measure and*

$$\int K(u) \, \mathrm{d}u = 1, \quad \int u^j K(u) \, \mathrm{d}u = 0, \quad \int |u|^\ell |K(u)| \, \mathrm{d}u < \infty.$$

In the Fourier domain, this means that for a kernel of order $\ell \in \mathbb{N}$, we have that

$$|\varphi_K(\xi) - 1| \lesssim |\xi|^\ell, \quad \text{for } \xi \to 0,$$

where φ_K is the Fourier transform of K. Then we can upper bound the risk of the Nadaraya-Watson estimator (6.1).

Theorem 6.6. *Assume that $p^* \in \mathcal{P}_{s,\alpha}$ and that $p^*_{Y|X}$ admits finite conditional moments up to order 2. Let K be a symmetric kernel of order $\lceil s + \gamma/2 \rceil$. Then for $\gamma < d_Y + 2$ there are suitable*

choices of h_x and $\sigma_{\min}$ such that

$$\mathbb{E}_{(X_1,Y_1),\dots,(X_n,Y_n),\bar{X}_1}\Big[\int \frac{|\mathring{\varphi}_{\bar{X}_1}(u)-\varphi_{\bar{X}_1}(u)|^2}{|u|^\gamma}\,\mathrm{d}u\Big] \lesssim \begin{cases} n^{-\frac{2}{2+\frac{d_X}{\alpha}}}, & \gamma > d_Y, \\ n^{-\frac{2}{2+\frac{d_X}{\alpha}}}\log(n), & \gamma = d_Y, \\ n^{-\frac{2+\frac{\gamma}{s}}{2+\frac{d_Y}{s}+\frac{d_X}{\alpha}+\frac{d_X\gamma}{2\alpha s}}}, & \gamma < d_Y. \end{cases}$$

The bounds match the lower bounds of Theorem 6.4 up to the logarithmic factor in case $\gamma = d_Y$, thus we conclude that the rate is minimax optimal up to the logarithmic factor for the class $\mathcal{P}_{s,\alpha}$. The logarithmic factor in case $\gamma = d$ is typical for boundary cases between different convergence regimes. This can also be seen in Corollary 3.8 and the subsequent results, when moving from the squared parametric regime to the high-dimensional regime.
Notably, in case $\gamma > d_Y$ the fast rate coincides with the rate that can be attained for the mean regression problem. Further note that the energy score is in the upper, fast rate regime. Besides the easy closed form, this is another indication why the energy score is a favorable score for the evaluation of high-dimensional conditional distributions.

6.3.3. Upper bound for Flow Matching estimator

In order to evaluate the Flow Matching method based on the objective (6.3), we want to exploit the connection to the NW estimator, which follows from Lemma 6.2 combined with Lemma 6.3. If we condition the probability path corresponding to the vector field approximated in the smoothed Flow Matching objective on x, then we obtain the density of the NW estimator. While this holds in a general setting, the evaluation requires knowledge of the specific model construction. Therefore, we need to choose the function $v(\cdot|\cdot)$. We assume the following:

Assumption 6.7.

1. *We consider the following choices of $\sigma\colon [0,1]\times\mathbb{R}^d \to \mathbb{R}_{>0}$ and $\mu\colon [0,1]\times\mathbb{R}^d\to\mathbb{R}^d$:*

$$\sigma_t = 1-(1-\sigma_{\min})t \quad \text{and} \quad \mu_t(y) = ty.$$

2. *We choose $\mathbb{U} = \mathcal{N}(0, I_{d_Y})$ for the latent distribution.*

3. *We use the Gaussian kernel for the smoothing in the covariates.*

Further, we restrict the result to the risk corresponding to the energy score. This enables the use of ReLU networks, which have been the focus of this thesis. The characteristic function at time $t=1$ of (6.5) conditioned on $x\in\mathbb{R}^{d_X}$ is denoted by $\hat{\varphi}_x$.

Theorem 6.8. *Let $\gamma = d_Y+\beta$ for $\beta\in(0,2)$ and $\frac{\gamma}{2}+s\le 2$. Then under the assumptions of Theorem 6.6 and $p_X(x)\le C$ for some $C>0$, there is a finite ReLU network such that*

$$\mathbb{E}_{(X_1,Y_1),\dots,(X_n,Y_n),\bar{X}_1}\Big[\int \frac{|\hat{\varphi}_{\bar{X}_1}(u)-\varphi_{\bar{X}_1}(u)|^2}{|u|^\gamma}\mathrm{d}u\Big] \lesssim n^{-\frac{2}{2+\frac{d_X}{\alpha}}}.$$

The proof combines Theorem 6.6 with the strategy from Section 5.4. Specifically, we can exploit the properties of the energy score to apply Grönwall's lemma.
In contrast to Theorem 6.6, we cannot use kernels of order greater than 2. This is due to the fact that the kernel is the density of the latent distribution. Changing the proof of Lemma 6.2 slightly, we can loosen the assumption on γ.

Corollary 6.9. *In the setting of Theorem 6.8, assume $s \geq \frac{\gamma}{2}$, $s \leq 2$ instead of $\frac{\gamma}{2} + s \leq 2$. Then we still obtain the rate*

$$\mathbb{E}_{(X_1,Y_1),\dots,(X_n,Y_n),\bar{X}_1}\Big[\int \frac{|\mathring{\varphi}_x(u) - \varphi_x(u)|^2}{|u|^\gamma}\,\mathrm{d}u\Big] \lesssim n^{-\frac{2}{2+d_X}}.$$

Combined with the restriction to the energy score, the result of Theorem 6.8 and Corollary 6.9 is limited to very low dimensions. Results for smaller choices of γ, which in turn allow for higher dimensions d_Y, are of course interesting. The next lemma shows that for $\gamma < d_Y$, we can bound the risk corresponding to the Fourier score in a way that allows for the application for Grönwall's lemma.

Lemma 6.10. *Assume that for two distributions $\mathbb{P}$ and $\mathbb{Q}$ on $\mathbb{R}^{d_X}$ there exists a $\tau \in (0,2)$ such that*

$$\int |u|^\tau |\varphi_{\mathbb{P}}(u)|^{2-\tau}\,\mathrm{d}u < C, \quad \int |u|^\tau |\varphi_{\mathbb{Q}}(u)|^{2-\tau}\,\mathrm{d}u < C$$

for a constant C. Further assume that the β-th moment of $\mathbb{P}$ and $\mathbb{Q}$ is finite. Then for $\gamma \leq d + \beta$

$$d_{\mathrm{FS}}(\mathbb{P},\mathbb{Q}) \lesssim d_{\mathrm{ES}}(\mathbb{P},\mathbb{Q}) + \mathbb{E}[|X-Y|]^\tau.$$

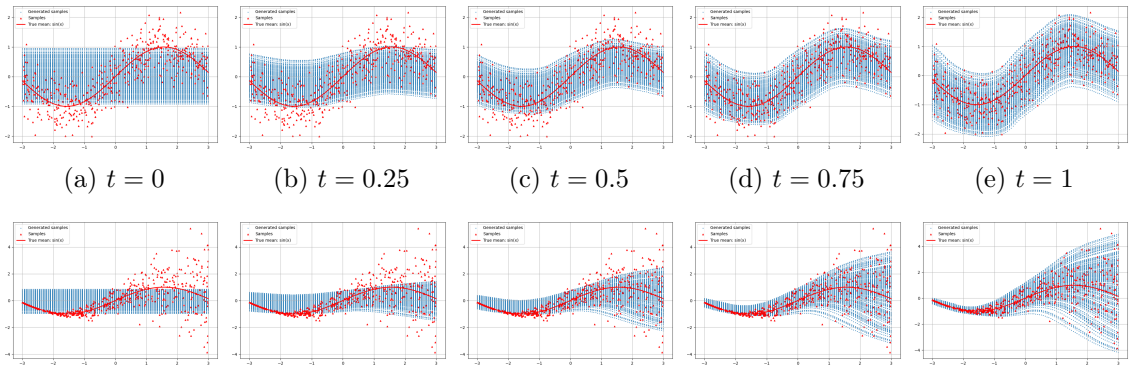
In order to apply Lemma 6.10, we need to guarantee enough regularity of the distributions. In case of $\mathbb{P}^*$, this is just a further assumption on the conditional distribution. In case of the Flow Matching estimator, this depends on the regularity of the network class used. Using ReLU networks, the highest possible regularity is Lipschitz continuity. Therefore, we would need to use smooth networks, such as ReQU networks, resulting from using the the activation function $\max(0,x)^2$ in (2.22), which allow for higher order derivatives. Then, a combination of Theorem 6.6, Corollary 6.9, Lemma 6.10 and a higher order approximation result such as the result by Belomestny et al. (2023, Theorem 2) leads to rates of convergence that permit the case $\gamma < d_Y$.

6.4. Numerical experiments

In this section, we want to give a first impression of how the proposed Flow Matching estimator performs in practice. All implementations of the Flow Matching model in this section are based on the implementation by Tong et al. (2024), which is available on github. Note that the code has been significantly altered and adapted to the conditional setting. We also use Poli et al. (2025) to solve the neural ODE. More precisely, we use the solver `dopri5`, the sensitivity `adjoint` and set atol = rtol = 10^{-5}.

6.4.1. Illustration

First, we want to illustrate the Flow Matching model for conditional distribution estimation in a simple, one-dimensional setting. To this end, let $X \sim \mathcal{U}[-3,3]$ and $Y \sim \mathcal{N}(\sin(X), (\tau(X))^2)$, for $\tau\colon [-3,3] \to (0,\infty)$. We employ a network architecture with 3 hidden SeLU layers with width 64. For a definition of this activation function, we refer to Klambauer et al. (2017). Furthermore, we conduct 10000 training iterations using a batch size of 128. For the optimization, we use the Adam optimizer (Kingma & Ba, 2014) with the standard parameters $lr = 0.001$, $\beta_1 = 0.9$ and $\beta_2 = 0.999$. For the kernel of the covariates, we employ the Gaussian kernel and a bandwidth of $h_x = 0.1$. We use the Epanechnikov kernel as the latent distribution and set $\sigma_{\min} = 10^{-4}$.
Figure 6.1 shows the flow at distinct times $t \in [0,1]$ for a constant function τ, resulting in a homoscedastic model, and a varying function τ, leading to a heteroscedastic model. We can see that the Flow Matching model can adapt to both, the homo- and the heteroscedastic setting.



(a) $t = 0$ (b) $t = 0.25$ (c) $t = 0.5$ (d) $t = 0.75$ (e) $t = 1$

(f) $t = 0$ (g) $t = 0.25$ (h) $t = 0.5$ (i) $t = 0.75$ (j) $t = 1$

Figure 6.1.: Model based on 500 samples. Flow for distinct $t \in [0,1]$. 200 latent samples are chosen once and then put through the model for different values of x. The red triangles are the observations. Top row $\tau \equiv 0.5$, bottom row $\tau(x) = 0.5 \cdot \left(\frac{x}{3} + 1\right)^2$).

Using fewer samples and refraining from calculating each iteration on a small batches reveals the impact the smoothing in the covariates has. In Figure 6.2, we can see that the smoothing prevents the model from overfitting. To this end, we increased the number of training iterations to 20000 while keeping all other setting as in the previous experiment. As expected, the smoothing shifts the observations slightly from their actual position in each training iteration. Thus, the sharp concentration of the estimated distribution's mass around the observation gets flattened. The same behavior is expected when using different batches in each training iteration.
We also observe this phenomenon when the true distribution is concentrated in specific regions of the covariate space. While overfitting is undesirable, the smoothing of the covariates hinders the model from improving the vector field in areas of sharp concentration in the latter case.

6.4.2. Regression datasets

To provide an initial indication of how the Flow Matching estimator compares to other methods, we apply it to classical regression datasets, more precisely, the Boston Housing dataset (Harrison & Rubinfeld, 1978), the Concrete Compressive Strength dataset (Yeh, 1998), the Energy

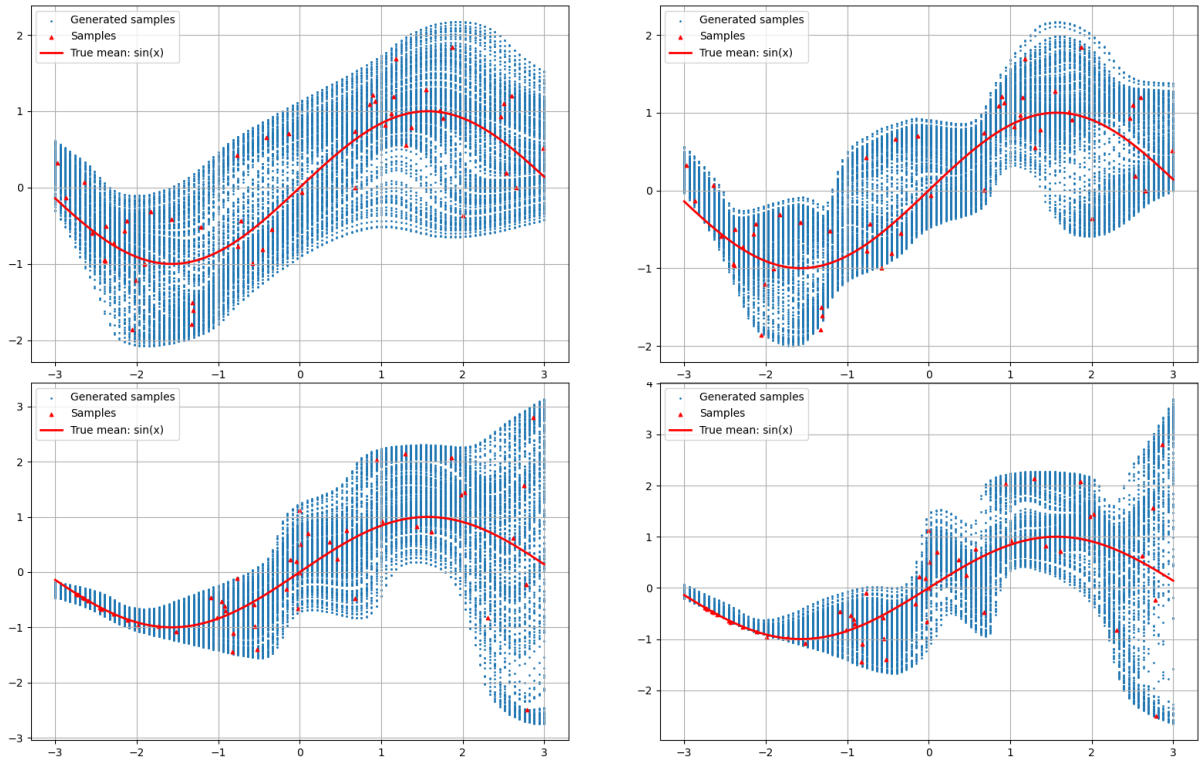


Figure 6.2.: Model based on 50 samples. Left column with conditional smoothing ($h_x = 0.1$), right column without conditional smoothing. 200 latent samples are chosen once and then put through the model for different values of x. The red triangles are the observations. Top row $\tau \equiv 0.5$, bottom row $\tau(x) = 0.5 \cdot \left(\frac{x}{3} + 1\right)^2$).

Efficiency dataset (Tsanas & Xifara, 2012), and the Kin8nm dataset by the University of Toronto. The Boston dataset contains 13 features that are expected to influence the median value of owner-occupied homes in the Boston area. The Concrete dataset contains eight features that are supposed to impact the compressive strength of concrete. The Energy dataset consists of eight features and two targets that relate building properties to heating and cooling loads. We use heating load as the target variable. The Kin8nm dataset uses eight settings of a robot arm to estimate the distance between the robot and an external object.

For this first indication, we compare the performance of the Flow Matching estimator to Walz et al. (2024) who performed their experiments on the same datasets. Note that their study also incorporates additional datasets. Although the overall task is the same, some of the models in Walz et al. (2024) require a separate neural network model to obtain a single-value output and operate on this output. Other models learn the distribution directly. In Section 6.4.3 we apply the Flow Matching estimator to single-value outputs of a preceding model. Hence in this section, we focus on cases where the distribution is learned directly. Additionally, Flow Matching learns an entire flow, which limits the explanatory power of directly comparing the number of training iterations to the size of the networks. We keep the network size the same as in the largest case of Walz et al. (2024), using networks with two hidden layers and a width of 50. In addition to 4000 training iterations, we also run our experiments for 20000 training iterations. In case of 4000 iterations, we use a learning rate of the Adam optimizer Kingma & Ba (2014) of 0.01 and

in case of 20000 iterations, we decrease this rate to 0.001.

We use the SeLU activation function and a batch size of 455 for the Energy, the Boston and the Kin8nm dataset and a batch size of 512 for the Concrete dataset. In all cases, we set $h_x = 0.001$ and used the Gaussian kernel for the covariates. We employ $\mathcal{N}(0, 1)$ as latent distribution. For the Boston and the Energy dataset, we used $\beta_1 = 0.95$, $\beta_2 = 0.999$ and a weight decay of 0.01. For the Concrete and the Kin8nm datasets, we used $\beta_1 = 0.9$ and $\beta_2 = 0.999$ and no weight decay. For the Concrete dataset, we choose $\sigma_{\min} = 0.0005$, for the other datasets, we set $\sigma_{\min} = 0.0001$.

Note that we do not fine-tune the hyperparameters, which are instead kept directly from the first implementation. With this setup, we can compare the behavior of the objective value of the Flow Matching algorithm, hereafter referred to as the loss, during training to that of the CRPS. We use random training and test splits with a test sample size of 10% to perform ten independent runs.

Interestingly, a decrease in loss does not correspond to a decrease in mean CRPS. Figure 6.3 illustrates this for the Boston dataset with 20000 iterations. While the additional training iterations decreased the loss in all ten runs from 5000 to 20000, there was no such relation in the CRPS. This effect was weaker in other datasets. It would be interesting to see whether using the Hyvärinen score aligns the loss and the score. As we saw in Section 5.1.2, the Hyvärinen score is directly connected to the objective of diffusion models, which, in turn, is closely related to the Flow Matching objective in the case of the Gaussian probability path.

The results of our experiments are collected in Table 6.1 and Table 6.2. Since we did not optimize anything in our model towards the CRPS, we report both, the mean CRPS after the entire training and the minimal mean CRPS after $\{5000, 10000, 15000, 20000\}$ or $\{1000, 2000, 3000, 4000\}$ training iterations. To calculate the mean CRPS, we use 100 draws from the latent distribution $\mathcal{N}(0, 1)$ and apply the model to the sample for every observation of the covariates in the test dataset. Using the true target observations, we calculate the empirical CRPS and average it over all 100 draws.

To compensate for the fact that we did not optimize the Flow Matching estimator towards the CRPS, we compare the minimal CRPS reported to the $2L$ models in Walz et al. (2024, Table 6), which use networks of the same size. Looking at the minimal achievable CRPS in Table 6.1, we see that the Flow Matching estimator running on 4000 training iterations is capable of beating all 7 methods considered in Walz et al. (2024) on the Boston dataset and the Concrete dataset, is in third place on the Energy dataset and in fourth place (shared with three other methods). Taking the mean over the 10 runs, the minimal mean CRPS of the Flow Matching estimator places second on the Boston dataset, last on the concrete dataset, sixth on the Energy dataset and last on the Kin8nm dataset. Considering the longer training runs as a compensation for the higher amount of information the Flow Matching estimator outputs, we see that the minimal achievable CRPS in Table 6.2 beats all models on all dataset except for the Energy datasets, where it places second with two other models. The minimal mean CRPS places second on the Boston dataset, seventh on the Concrete dataset, fifth shared with one other model on the Energy dataset, and second shared with two others on the Kin8nm dataset.

		Boston	Concrete	Energy	Kin8nm
Minimal	Mean	1.55	2.50	0.267	0.0473
Mean	Min	1.16	2.13	0.237	0.0427
CRPS	Max	1.79	3.15	0.320	0.0506
Mean CRPS	Mean	1.73	2.54	0.286	0.0478
after 4000	Min	1.19	2.19	0.237	0.0427
iterations	Max	2.16	3.31	0.380	0.0515

Table 6.1.: Mean CRPS values based on 100 draws from the latent distribution evaluated on the test dataset. The top row reports the mean, the minimum and the maximum over the minimal mean CRPS based on evaluations at $\{1000, 2000, 3000, 4000\}$ training iterations based on 10 independent runs. The bottom row uses only the mean CRPS values after 4000 runs.

		Boston	Concrete	Energy	Kin8nm
Minimal	Mean	1.54	2.57	0.256	0.0416
Mean	Min	1.17	2.11	0.241	0.0385
CRPS	Max	1.83	3.1	0.278	0.0428
Mean CRPS	Mean	1.68	2.61	0.265	0.0419
after 20000	Min	1.28	2.13	2.11	0.0385
iterations	Max	2.00	3.14	0.283	0.0439

Table 6.2.: Mean CRPS values based on 100 draws from the latent distribution evaluated on the test dataset. The top row reports the mean, the minimum and the maximum over the minimal mean CRPS based on evaluations at $\{5000, 10000, 15000, 20000\}$ training iterations based on 10 independent runs. The bottom row uses only the mean CRPS values after 20000 runs.

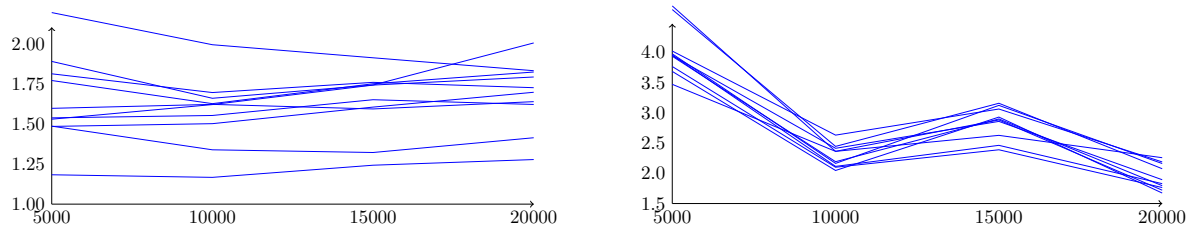
Overall, our results suggest that the Flow Matching estimator is a promising approach for distribution regression. Of course, an exact comparison should be conducted on the same computer using the same network construction, optimizers, and hyperparameter tuning.

6.4.3. Probabilistic Weather forecasting using Flow Matching

In this section, we apply the Flow Matching estimator to point forecasts from the WeatherBench2 dataset (Rasp et al., 2024), which is a benchmark dataset for weather prediction. Our goal is to learn a conditional predictive distribution from single-valued point forecasts and their realizations. Walz et al. (2024) took the same approach, performing their evaluation on an earlier version of the WeatherBench dataset.

We use the Integrated Forecast System (IFS) numerical prediction model in its High Resolution Configuration (HRES) from the European Center for Medium-Range Weather Forecasts (ECMWF). The forecast includes 16 variables and covers a 0.25 latitude-longitude grid from January 2016 to January 2023. Forecasts are issued every 12 hours for lead times ranging from 0 to 240 hours in 6 hour intervals.

Our experiments are based on temperature at a height of two meters. We use data from 2016 to 2021 for training and data from 2022 for evaluation. Furthermore, we use forecasts for two, three,



(a) Mean CRPS vs. number of training iterations.

(b) Loss vs. number of training iterations.

Figure 6.3.: Behavior of the mean CRPS and the loss compared to the number of training iterations when running the experiment for the Boston dataset with 20000 iterations. The mean CRPS was calculated based on 100 draws from the latent distribution evaluated on the test dataset. Note that the lines connect the points at $\{5000, 10000, 15000, 20000\}$.

four, and five days ahead. These correspond to lead times of 48, 72, 96, and 120 hours. Due to computational restrictions, we focus on locations in Europe between latitudes 42°N and 60°N and longitudes 10°W and 30°E, selecting 10 places from a uniform distribution over a broader 5.625 latitude-longitude grid. The resulting locations are illustrated in Figure 6.4.

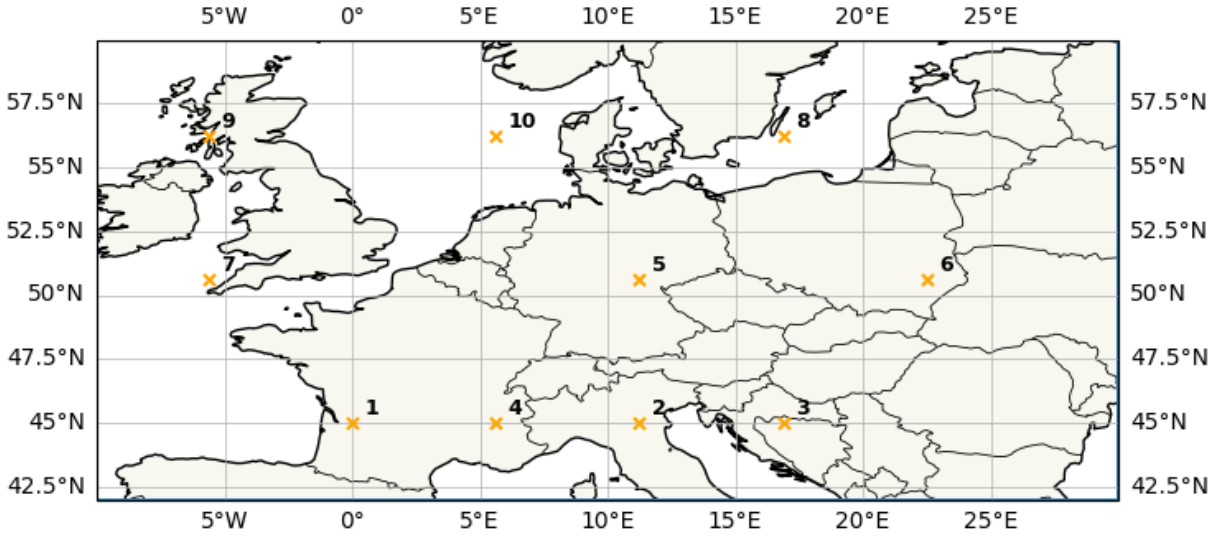


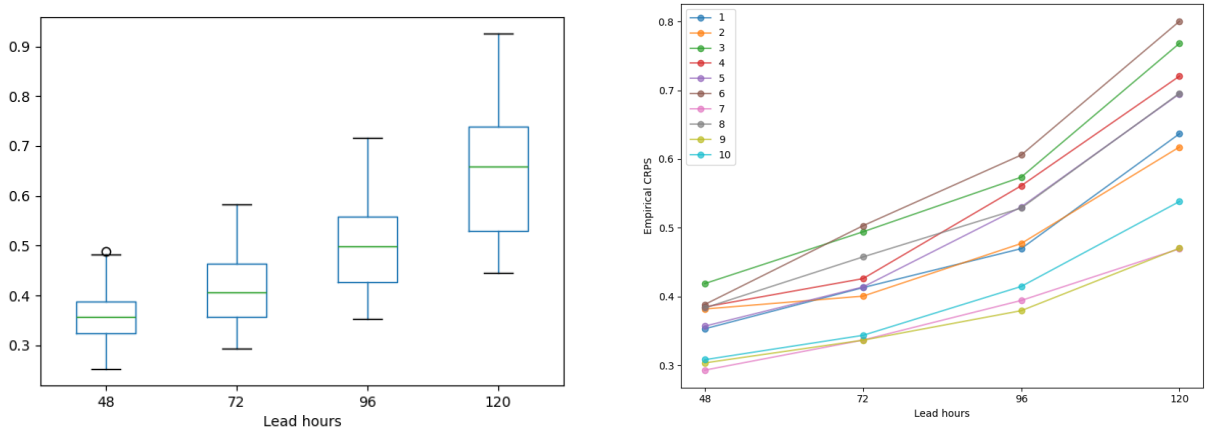
Figure 6.4.: 10 places drawn from a uniform distribution over a 5.625 latitude-longitude grid covering Europe. The map was created using Met Office (2015) and made with free vector and raster map data from Natural Earth.

As proposed by Rasp et al. (2024) and Lam et al. (2023), we use the IFS HRES forecast with a lead time of 0 as the true realization.

First, we train and evaluate each location and lead time individually. Then, we use the Flow Matching estimator to estimate the conditional distribution. The covariate kernel is a Gaussian kernel with a bandwidth of 0.01. The latent distribution is $\mathcal{N}(0, 1)$ and we set $\sigma_{\min} = 1$. Note that we scaled only the forecasts, not the true observations, which explains the large variance. The network architecture consists of two hidden SeLU layers with a width of 50. We conduct 10000 training iterations. In each training iteration, we train the model with a batch size of 455 samples. For the optimization, we employ the Adam optimizer Kingma & Ba (2014) with the

parameters $lr = 0.001$, $\beta_1 = 0.95$, $\beta_2 = 0.999$ and a weight decay of 0.01. For every location, we conduct 10 independent runs. After the training in each run, we draw 100 observations from the latent distribution $\mathcal{N}(0, 1)$ and apply the model to every forecast for days in the year 2022. Thus, we obtain a collection of 100 observations sampled from the estimated distribution for every forecast. Using these observations and the true realization, we calculate the CRPS and average over the test dataset.

The results can be seen in Figure 6.5. Compared to the state-of-the-art models based on functional generative networks (Alet et al., 2025), the model GenCast (Price et al., 2025) and the IFS ensemble version (ENS), our experiments indicate that the Flow Matching model performs surprisingly well. While the other models are trained on a broad range of input data using much larger and more sophisticated networks, the Flow Matching model is trained on simple point forecasts. For an easy comparison for the year 2022, we refer to this[1] illustration, which can be set to `Europe`. However, note that we could not assess the exact alignment of the region. Additionally, the only WeatherBench2 application we can access is the data from the WeatherBench website linked above. The cited publications are based on an earlier version of the WeatherBench database.



(a) Mean empirical CRPS over all runs at all locations.

(b) Mean empirical CRPS for different locations. The locations refer to the labels in Figure 6.4.

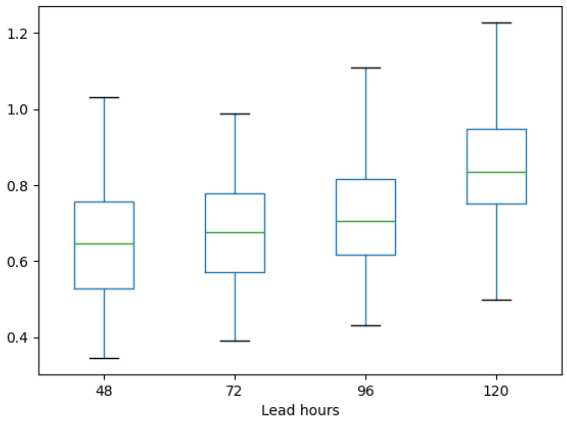
Figure 6.5.: Results of the Flow Matching estimator using separate models for different locations.

A comparison to Walz et al. (2024, Table 2), who also built their model based solely on point forecasts, indicates that the Flow Matching estimator should be investigated more comprehensively. Although our estimator's mean CRPS is much lower for all lead times, an exact comparison is hindered by the fact that the analysis in Walz et al. (2024) is based on an earlier WeatherBench dataset.

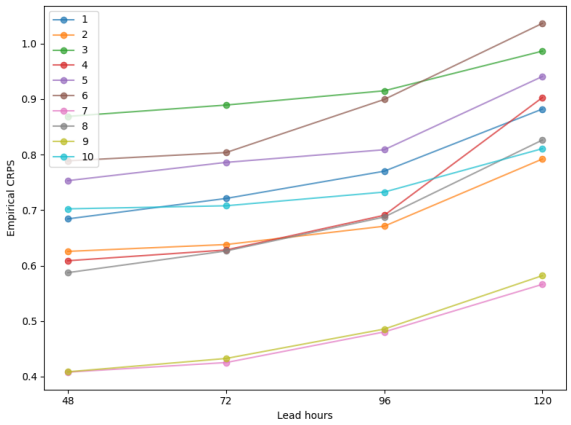
As we saw in Section 6.4.2, the Flow Matching model is not restricted to the univariate case. Therefore, we can conduct the same experiment using a single model for all locations rather than separate models. The results are summarized in Figure 6.6. While keeping the network size unmodified leads to inferior results, the predictions are still quite accurate considering the size of the network. By increasing the network to four hidden SeLU layers with a width of 128 and the

[1] `https://sites.research.google/gr/weatherbench/probabilistic-scores/`

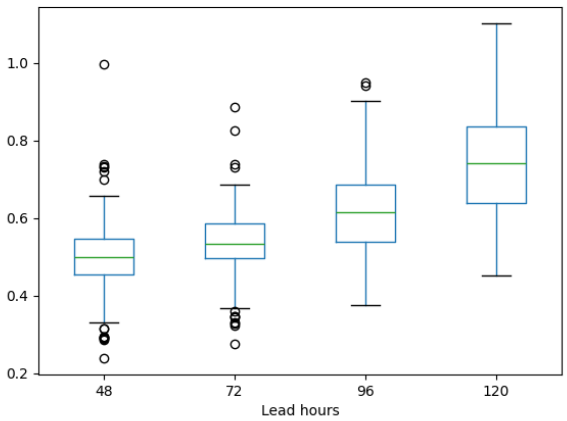
number of training iterations to 20000, we can improve the mean CRPS again while drastically reducing the overall size and computation time compared to the separate models. However, we observe that this comes at the cost of more outliers.



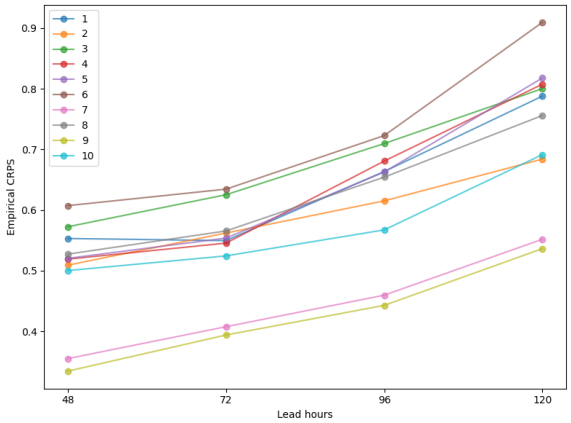
(a) Mean empirical CRPS over all runs at all locations, two layer network.



(b) Mean empirical CRPS for different locations, two layer network.



(c) Mean empirical CRPS over all runs at all locations, four layer network.



(d) Mean empirical CRPS for different locations, four layer network.

Figure 6.6.: Results of the Flow Matching estimator using one joint model for all locations.

6.5. Proofs

6.5.1. Proof of Section 6.1

Proof of Lemma 6.1. For fixed $t \in [0,1]$ we have

$$|\tilde{v}_t(z,x) - \bar{v}_t(z,x)|^2 = |\tilde{v}_t(z,x)|^2 - 2\langle \tilde{v}_t(z,x), \bar{v}_t(z,x)\rangle + |\bar{v}_t(z,x)|^2,$$
$$|\tilde{v}_t(z,x) - v_t(z|Y_i)|^2 = |\tilde{v}_t(z,x)|^2 - 2\langle \tilde{v}_t(z,x), v_t(z|Y_i)\rangle + |v_t(z|Y_i)|^2.$$

The last term does not influence the minimal argument in $\tilde{v}$. For the first two we have

$$\mathbb{E}_{Z_t,X\sim\bar{p}_t}[|\tilde{v}_t(Z_t,X)|^2] = \int\int |\tilde{v}_t(z,x)|^2 \frac{1}{n}\sum_{i=1}^{n} p_t(z|Y_i)\delta_{X_i}(x)\,\mathrm{d}z\,\mathrm{d}x$$

$$= \frac{1}{n}\sum_{i=1}^{n} \int\int |\tilde{v}_t(z,x)|^2 p_t(z|Y_i)\delta_{X_i}(x)\,\mathrm{d}z\,\mathrm{d}x$$
$$= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{Z_t\sim p_t(\cdot|Y_i)}[|\tilde{v}_t(z,X_i)|^2]$$

and

$$\begin{aligned}
&\mathbb{E}_{Z_t,X\sim\bar{p}_t}[\langle \tilde{v}_t(Z_t,X), \bar{v}_t(Z_t,X)\rangle] \\
&\quad= \int\int \langle \tilde{v}_t(z,x), \bar{v}_t(z,x)\rangle \bar{p}(z,x)\,\mathrm{d}z\,\mathrm{d}x \\
&\quad= \int\int \Big\langle \tilde{v}_t(z,x), \frac{\sum_{i=1}^n v_t(z|Y_i)p_t(z|Y_i)\delta_{X_i}(x)}{\sum_{i=1}^n p_t(z|Y_i)\delta_{X_i}(x)} \Big\rangle \frac{1}{n}\sum_{i=1}^{n} p_t(z|Y_i)\delta_{X_i}(x)\,\mathrm{d}z\,\mathrm{d}x \\
&\quad= \frac{1}{n}\sum_{i=1}^{n} \int\int \langle \tilde{v}_t(x), v_t(z|Y_i)\rangle p_t(z|Y_i)\delta_{X_i}(x)\,\mathrm{d}z\,\mathrm{d}x \\
&\quad= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{Z_t\sim p_t(\cdot|Y_i)}[\langle \tilde{v}_t(Z_t,X_i), v_t(Z_t|Y_i)\rangle]. \qquad \square
\end{aligned}$$

Proof of Lemma 6.2. The proof is nearly identical to the proof of Lemma 6.1. For fixed $t \in [0,1]$ we have

$$|\tilde{v}_t(z,x) - v_t(z,x)|^2 = |\tilde{v}_t(z,x)|^2 - 2\langle \tilde{v}_t(z,x), v_t(z,x)\rangle + |v_t(z,x)|^2,$$
$$|\tilde{v}_t(z,x) - v_t(z|Y_i)|^2 = |\tilde{v}_t(z,x)|^2 - 2\langle \tilde{v}_t(z,x), v_t(z|Y_i)\rangle + |v_t(z|Y_i)|^2.$$

The last term does not influence the minimal argument in $\tilde{v}$. For the first two we have

$$\begin{aligned}
\mathbb{E}_{Z_t,X\sim p_t}[|\tilde{v}_t(Z_t,X)|^2] &= \int\int |\tilde{v}_t(z,x)|^2 \frac{1}{n}\sum_{i=1}^{n} p_t(z|Y_i)K^{\mathrm{x}}_{h_x}(x-X_i)\,\mathrm{d}z\,\mathrm{d}x \\
&= \frac{1}{n}\sum_{i=1}^{n} \int\int |\tilde{v}_t(z,x)|^2 p_t(z|Y_i)K^{\mathrm{x}}_{h_x}(x-X_i)\,\mathrm{d}z\,\mathrm{d}x \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\substack{Z_t\sim p_t(\cdot|Y_i)\\ X\sim K^{\mathrm{x}}_{h_x}(\cdot-X_i)}}[|\tilde{v}_t(z,X)|^2]
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}_{Z_t,X\sim p_t}[\langle \tilde{v}_t(Z_t,X), v_t(Z_t,X)\rangle] &= \int\int \langle \tilde{v}_t(z,x), v_t(z,x)\rangle \bar{p}(z,x)\,\mathrm{d}z\,\mathrm{d}x \\
&= \int\int \Big\langle \tilde{v}_t(z,x), \frac{\sum_{i=1}^n v_t(z|Y_i)p_t(z|Y_i)K^{\mathrm{x}}_{h_x}(x-X_i)}{\sum_{i=1}^n p_t(z|Y_i)K^{\mathrm{x}}_{h_x}(x-X_i)} \Big\rangle \\
&\qquad \cdot \frac{1}{n}\sum_{i=1}^{n} p_t(z|Y_i)K^{\mathrm{x}}_{h_x}(x-X_i)\,\mathrm{d}z\,\mathrm{d}x \\
&= \frac{1}{n}\sum_{i=1}^{n} \int\int \langle \tilde{v}_t(x), v_t(z|Y_i)\rangle p_t(z|Y_i)K^{\mathrm{x}}_{h_x}(x-X_i)\,\mathrm{d}z\,\mathrm{d}x
\end{aligned}$$

$$= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\substack{Z_t\sim p_t(\cdot|Y_i)\\ X\sim K^{\mathrm{x}}_{h_x}(\cdot - X_i)}} [\langle \tilde{v}_t(Z_t, X), v_t(Z_t|Y_i)\rangle]. \qquad \square$$

Proof of Lemma 6.3. The proof follows along the same lines as the proof of Lemma 5.2. We have that

$$\begin{aligned}
\frac{d}{dt}\mathring{p}_t(y|x) &= \frac{\sum_{i=1}^n \frac{d}{dt} p_t(y|Y_i) K^{\mathrm{x}}_{h_x}(x - X_i)}{\sum_{i=1}^n K^{\mathrm{x}}_{h_x}(x - X_i)} \\
&= \frac{-\sum_{i=1}^n \operatorname{div}(p_t(y|Y_i) v_t(y|Y_i)) K^{\mathrm{x}}_{h_x}(x - X_i)}{\sum_{i=1}^n K^{\mathrm{x}}_{h_x}(x - X_i)} \\
&= -\operatorname{div}\Big(\frac{\sum_{i=1}^n p_t(y|Y_i) v_t(y|Y_i) K^{\mathrm{x}}_{h_x}(x - X_i)}{\sum_{i=1}^n K^{\mathrm{x}}_{h_x}(x - X_i)}\Big) \\
&= -\operatorname{div}\Big(\frac{\sum_{i=1}^n p_t(z|Y_i) K^{\mathrm{x}}_{h_x}(x - X_i)}{\sum_{i=1}^n K^{\mathrm{x}}_{h_x}(x - X_i)} \cdot \frac{\sum_{i=1}^n p_t(y|Y_i) v_t(y|Y_i) K^{\mathrm{x}}_{h_x}(x - X_i)}{\sum_{i=1}^n p_t(z|Y_i) K^{\mathrm{x}}_{h_x}(x - X_i)}\Big) \\
&= -\operatorname{div}(\mathring{p}_t(z|x)\mathring{v}_t(z, x)). \qquad \square
\end{aligned}$$

6.5.2. Proofs of Section 6.3

Proof of Theorem 6.4. Our proof is similar to Li et al. (2022, Theorem 4.1), who use the classical approach of a nonparametric lower bound as in Tsybakov (2009). We first construct the following set of densities. Choose a function $h \in C^\infty([0,1])$ such that $\int h(x)\,\mathrm{d}x = 0$, $\int h^2(x)\,\mathrm{d}x = 1$, $\|h\|_\infty < \infty$, and $\|\varphi_h\|_\gamma < \infty$, where φ_h is the Fourier transform of h. Let $r, m \in \mathbb{N}$ and define the density of a disturbed uniform distribution via

$$p^{\Delta}_{Y|X}(y|x) = 1 + \sum_{\bar{i}}\sum_{\bar{j}} \Delta_{\bar{i},\bar{j}} \prod_{k\in[d_Y]} h_{i_k,r}(y_k) \prod_{k\in[d_X]} h_{j_k,m}(x_k),$$

where $\bar{i}$ is short for $\bar{i} \in \{1, ..., r\}^{d_Y}$ and $\bar{j}$ is short for $\bar{j} \in \{1, ..., m\}^{d_X}$, $\Delta_{\bar{i},\bar{j}} \in \{\pm 1\}$ and for a $\rho > 0$

$$\begin{aligned}
h_{i_k,r}(y_k) &= \rho\sqrt{r} h(r y_k - i_k + 1), \\
h_{j_k,m}(x_k) &= \rho h(m x_k - j_k + 1).
\end{aligned} \tag{6.12}$$

First, we show that the functions $p^{\Delta}_{Y|X}$ belong to the class $\mathcal{P}_{s,\alpha}$.

Lemma 6.11.

1. *Let*
$$\rho r^{d_Y/2} \lesssim \frac{1}{2}.$$
Then $p^{\Delta}_{Y|X}$ is a density and $p^{\Delta}_{Y|X}(y|x) \geq \frac{1}{2}$ for all $y \in \mathbb{R}^{d_Y}$ and $x \in \mathbb{R}^{d_X}$.

2. *Let $r \geq 1$. Then $p^{\Delta}_{Y|X} \in \mathcal{H}^s(\Gamma)$ where*
$$\Gamma \asymp 1 + \rho r^{s + d_Y/2}.$$

3. Let φ_x^Δ be the characteristic function of $p_{Y|X}^\Delta$. Then

$$\sup_{x,x' \in \mathcal{X}, x \neq x'} \frac{\left\| \varphi_x^\Delta - \varphi_{x'}^\Delta \right\|_\gamma}{|x - x'|^\alpha} \leq L$$

for

$$L \asymp \rho m^\alpha r^{d_Y/2 - \gamma/2}.$$

Next, we are going to bound the Kullback-Leibler divergence between $p_{Y|X}^\Delta$ and $p_{Y|X}^{\Delta'}$.

Lemma 6.12. *Under the condition of Lemma 6.11 No. 1, then we have that*

$$\mathsf{KL}(p^\Delta \mid p^{\Delta'}) \lesssim \rho^2 r^{d_Y}.$$

Next, we show that there is a set of hypotheses large enough, that for two elements out of this set our distance is lower bounded. For ease of notation, we identify a distribution with its density.

Lemma 6.13. *There is a set T of densities $p_{Y|X}^\Delta$ such that $|T| \geq 2^{r^{d_Y} m^{d_X}/8}$ with $d_H(\Delta, \Delta') \geq r^{d_Y} m^{d_X}/8$, where d_H is the Hamming distance, and*

$$\int \int \frac{|\varphi_x^\Delta(u) - \varphi_x^{\Delta'}(u)|^2}{|u|^\gamma} \, \mathrm{d}u \, p_X(x) \, \mathrm{d}x \gtrsim r^{d_Y - \gamma} \rho^2$$

for every $p_{Y|X}^\Delta, p_{Y|X}^{\Delta'} \in T$ with $\Delta \neq \Delta'$.

In the following, let T be a set as constructed in Lemma 6.13. To apply Tsybakov (2009, Theorem 2.7), we need to ensure

$$\mathsf{KL}((p^\Delta)^{\otimes n} \big| (p^{\Delta'})^{\otimes n}) = n \mathsf{KL}(p^\Delta, p^{\Delta'}) \lesssim n \rho^2 r^{d_Y} \lesssim \log(|T|), \tag{6.13}$$

where $(p^\Delta)^{\otimes n}$ is the product measure corresponding to the n i.i.d. samples based on the density p^Δ. The condition (6.13) holds true if

$$n \rho^2 r^{d_Y} \lesssim r^{d_Y} m^{d_X} \quad \text{or equivalently} \quad \rho \lesssim n^{-\frac{1}{2}} m^{d_X/2}. \tag{6.14}$$

Now we can combine (6.14) with Lemma 6.12 and Lemma 6.11. We obtain using Tsybakov (2009, Theorem 2.7) and $r \geq 1$

$$\begin{aligned}
\inf_{\hat{f}_n} \sup_{f \in \mathcal{P}_{s,\alpha}} \mathbb{E}\Big[\int \int \frac{|\hat{\varphi}_x(u) - \varphi_x(u)|^2}{|u|^\gamma} \, \mathrm{d}u \, p_X(x) \, \mathrm{d}x \Big] \\
&\gtrsim r^{d_Y - \gamma} \min\left(n^{-1} m^{d_X}, r^{-d_Y}, r^{-d_Y - 2s}, m^{-2\alpha} r^{-d_Y + \gamma}\right) \\
&\geq \min\left(r^{d_Y - \gamma} n^{-1} m^{d_X}, r^{-\gamma}, r^{-\gamma - 2s}, m^{-2\alpha}\right) \\
&= \min\left(r^{d_Y - \gamma} n^{-1} m^{d_X}, r^{-\gamma - 2s}, m^{-2\alpha}\right).
\end{aligned}$$

If $\gamma \geq d_Y$, then we can fix $r = 1$. Then we can choose $m \asymp n^{\frac{1}{2\alpha+d_X}}$, which yields the rate

$$\inf_{\hat{f}_n} \sup_{f \in \mathcal{P}_{s,\alpha}} \mathbb{E}\Big[\int \int \frac{|\hat{\varphi}_x(u) - \varphi_x(u)|^2}{|u|^\gamma} \,\mathrm{d}u \, p_X(x) \,\mathrm{d}x \Big] \gtrsim n^{-\frac{2}{2+\frac{d_X}{\alpha}}}.$$

In case $\gamma < d_Y$, we choose $m \asymp r^{\frac{\gamma+2s}{2\alpha}}$ and $r \asymp n^{-\frac{1}{2s+d_Y+d_X\frac{\gamma+2s}{2\alpha}}}$. In this case

$$\inf_{\hat{f}_n} \sup_{f \in \mathcal{P}_{s,\alpha}} \mathbb{E}\Big[\int \int \frac{|\hat{\varphi}_x(u) - \varphi_x(u)|^2}{|u|^\gamma} \,\mathrm{d}u \, p_X(x) \,\mathrm{d}x \Big] \gtrsim n^{-\frac{2+\frac{\gamma}{s}}{2+\frac{d_Y}{s}+\frac{d_Y}{\alpha}+\frac{d_X\gamma}{2\alpha s}}}.$$

This concludes the proof.

□

Proof of Theorem 6.6. First, we define

$$\mathring{\varphi}_x(u) \coloneqq \mathcal{F}K(\sigma_{\min} u) \sum_{i=1}^n e^{\mathrm{i}\langle u, Y_i\rangle} w_i(x), \qquad w_i(x) \coloneqq \frac{K^{\mathrm{x}}_{h_x}(x - X_i)}{\sum_{i=1}^n K^{\mathrm{x}}_{h_x}(x - X_i)},$$

where $\mathcal{F}K$ is the Fourier transformation of K as defined in (2.10). Here and in the following, i is the complex unit, while i is the index. Further, let $\varphi_x(u) \coloneqq \mathbb{E}_{Y|X=x}[e^{\mathrm{i}\langle u, Y\rangle}]$ be the true characteristic function of the conditional distribution. Then

$$\begin{aligned}
\mathring{\varphi}_x(u) - \varphi_x(u) &= \sum_{i=1}^n w_i(x)(\mathcal{F}K(\sigma_{\min} u) e^{\mathrm{i}\langle u, Y_i\rangle} - 1) \\
&= \mathcal{F}K(\sigma_{\min} u) \sum_{i=1}^n w_i(x)(e^{\mathrm{i}\langle u, Y_i\rangle} - \varphi_{X_i}(u)) && (6.15) \\
&\quad + \mathcal{F}K(\sigma_{\min} u) \sum_{i=1}^n w_i(x)(\varphi_{X_i}(u) - \varphi_x(u)) && (6.16) \\
&\quad + \sum_{i=1}^n w_i(x)(\mathcal{F}K(\sigma_{\min} u)\varphi_x(u) - \varphi_x(u)). && (6.17)
\end{aligned}$$

For the first term (6.15), we obtain integrating over $\frac{1}{|u|^\gamma}$ and taking the expectation with respect to (X_i, Y_i) and $(\bar{X}_1)$

$$\begin{aligned}
&\mathbb{E}_{(X_i,Y_i),\bar{X}_1}\Big[\int \frac{|\mathcal{F}K(\sigma_{\min} u) \sum_{i=1}^n w_i(\bar{X}_1)(e^{\mathrm{i}\langle u, Y_i\rangle} - \varphi_{X_i}(u))|^2}{|u|^\gamma} \,\mathrm{d}u \Big] \\
&\quad = \int \frac{|\mathcal{F}K(\sigma_{\min} u)|^2}{|u|^\gamma} \mathbb{E}_{(X_i,Y_i),\bar{X}_1}\Big[\Big| \sum_{i=1}^n w_i(\bar{X}_1)(e^{\mathrm{i}\langle u, Y_i\rangle} - \varphi_{X_i}(u)) \Big|^2 \Big] \,\mathrm{d}u \\
&\quad \leq \int \frac{|\mathcal{F}K(\sigma_{\min} u)|^2}{|u|^\gamma} \mathbb{E}_{X_i}\Big[\sum_{i=1}^n \mathbb{E}_{\bar{X}_1}[w_i^2(\bar{X}_1)] \mathbb{E}_{Y_i|X_i}\big[|e^{\mathrm{i}\langle u, Y_i\rangle} - \varphi_{X_i}(u)|^2 \big] \Big] \,\mathrm{d}u \\
&\quad = \int \frac{|\mathcal{F}K(\sigma_{\min} u)|^2}{|u|^\gamma} \mathbb{E}_{X_i}\Big[\sum_{i=1}^n \mathbb{E}_{\bar{X}_1}[w_i^2(\bar{X}_1)](1 - |\varphi_{X_i}(u)|^2) \Big] \,\mathrm{d}u
\end{aligned}$$

$$
\begin{aligned}
&= \sigma_{\min}^{\gamma-d_Y} \int \frac{|\mathcal{F}K(t)|^2}{|t|^\gamma} \mathbb{E}_{X_i}\Big[\sum_{i=1}^n \mathbb{E}_{\bar{X}_1}[w_i^2(\bar{X}_1)](1-|\varphi_{X_i}(\frac{t}{\sigma_{\min}})|^2)\Big]\,\mathrm{d}u \\
&= \sigma_{\min}^{\gamma-d_Y} \int_{|t|<1} \frac{|\mathcal{F}K(t)|^2}{|t|^\gamma} \mathbb{E}_{X_i}\Big[\sum_{i=1}^n \mathbb{E}_{\bar{X}_1}[w_i^2(\bar{X}_1)](1-|\varphi_{X_i}(\frac{t}{\sigma_{\min}})|^2)\Big]\,\mathrm{d}u \\
&\quad + \sigma_{\min}^{\gamma-d_Y} \int_{|t|\geq 1} \frac{|\mathcal{F}K(t)|^2}{|t|^\gamma} \mathbb{E}_{X_i}\Big[\sum_{i=1}^n \mathbb{E}_{\bar{X}_1}[w_i^2(\bar{X}_1)](1-|\varphi_{X_i}(\frac{t}{\sigma_{\min}})|^2)\Big]\,\mathrm{d}u.
\end{aligned}
$$

For the last term, we obtain using the standard observation from Nadaraya-Watson estimation $\mathbb{E}_{\bar{X}_1}[w_i^2(\bar{X}_1)] \lesssim \frac{1}{nh_x^{d_X}}$ c.f. Tsybakov (2009, Lemma 1.3)

$$
\begin{aligned}
&\sigma_{\min}^{\gamma-d_Y} \int_{|t|\geq 1} \frac{|\mathcal{F}K(t)|^2}{|t|^\gamma} \mathbb{E}_{X_i}\Big[\sum_{i=1}^n \mathbb{E}_{\bar{X}_1}[w_i^2(\bar{X}_1)](1-|\varphi_{X_i}(\frac{t}{\sigma_{\min}})|^2)\Big]\,\mathrm{d}u \\
&\qquad \leq \sigma_{\min}^{\gamma-d_Y} \int_{|t|\geq 1} |\mathcal{F}K(t)|^2 \mathbb{E}_{X_i}\Big[\sum_{i=1}^n \mathbb{E}_{\bar{X}_1}[w_i^2(\bar{X}_1)]\Big]\,\mathrm{d}u \\
&\qquad \lesssim \sigma_{\min}^{\gamma-d_Y} n^{-1} h_x^{-d_X}.
\end{aligned}
$$

For the first term, we start with the case $\gamma < d_Y$. Then, analogously to the last term, we obtain

$$
\sigma_{\min}^{\gamma-d_Y} \int_{|t|<1} \frac{|\mathcal{F}K(t)|^2}{|t|^\gamma} \mathbb{E}_{X_i}\Big[\sum_{i=1}^n \mathbb{E}_{\bar{X}_1}[w_i^2(\bar{X}_1)](1-|\varphi_{X_i}(\frac{t}{\sigma_{\min}})|^2)\Big]\,\mathrm{d}u \lesssim \sigma_{\min}^{\gamma-d_Y} n^{-1} h_x^{-d_X}.
$$

If $\gamma \in (d_Y, d_Y+2)$, we can use the following result.

Lemma 6.14. *(Székely et al., 2005, Lemma 1) For $0<\alpha<2$ and all $y \in \mathbb{R}^{d_Y}$*

$$
\int \frac{1-\cos(t,y)}{|t|^{d+\alpha}}\,\mathrm{d}t = C(d,\alpha)|y|^\alpha,
$$

where $t \in \mathbb{R}^{d_Y}$, and $C(d,\alpha)>0$ is a constant depending only on d and α.

Additionally,

$$
|\varphi_{X_i}(u)|^2 = \mathbb{E}_{Y_i|X_i}[\cos(u, Y-Y')],
$$

where Y, Y' are conditionally independent. Then

$$
\begin{aligned}
&\sigma_{\min}^{\gamma-d_Y} \int_{|t|<1} \frac{|\mathcal{F}K(t)|^2}{|t|^\gamma} \mathbb{E}_{X_i}\Big[\sum_{i=1}^n \mathbb{E}_{\bar{X}_1}[w_i^2(\bar{X}_1)]\Big(1-\Big|\varphi_{X_i}\Big(\frac{t}{\sigma_{\min}}\Big)\Big|^2\Big)\Big]\,\mathrm{d}u \\
&\qquad = \sigma_{\min}^{\gamma-d_Y} \mathbb{E}_{X_i}\Big[\sum_{i=1}^n \mathbb{E}_{\bar{X}_1}[w_i^2(\bar{X}_1)] \mathbb{E}_{Y_i|X_i}\Big[\int_{|t|<1} \frac{1-\cos(\langle \frac{t}{\sigma_{\min}}, Y_i - Y_i'\rangle)}{|t|^\gamma}\,\mathrm{d}t\Big]\Big] \\
&\qquad \leq \mathbb{E}_{X_i}\Big[\sum_{i=1}^n \mathbb{E}_{\bar{X}_1}[w_i^2(\bar{X}_1)] \mathbb{E}_{Y_i|X_i}\Big[\int \frac{1-\cos(\langle u, Y_i - Y_i'\rangle)}{|u|^\gamma}\,\mathrm{d}u\Big]\Big] \\
&\qquad \lesssim \mathbb{E}_{X_i}\Big[\sum_{i=1}^n \mathbb{E}_{\bar{X}_1}[w_i^2(\bar{X}_1)] \mathbb{E}_{Y_i|X_i}\Big[|Y_i - Y_i'|^{\gamma-d_Y}\Big]\Big] \\
&\qquad \lesssim n^{-1} h_x^{-d_X}.
\end{aligned}
$$

In the last inequality we used the assumption that the conditional expectation is bounded. Finally, if $\gamma = d_Y$, we use that for $y \in \mathbb{R}^{d_Y}$

$$\begin{aligned}\int_{|t|<1} \frac{1-\cos(\langle \frac{t}{\sigma_{\min}}, y\rangle)}{|t|^\gamma}\,\mathrm{d}t &= \sigma_{\min}^{d_Y-\gamma}\int_{|u|<\sigma_{\min}^{-1}} \frac{1-\cos(\langle u, y\rangle)}{|u|^\gamma}\,\mathrm{d}u \\ &= \sigma_{\min}^{d_Y-\gamma}\int_{|u|<|y|^{-1}} \frac{1-\cos(\langle u, y\rangle)}{|u|^\gamma}\,\mathrm{d}u \\ &\quad + \sigma_{\min}^{d_Y-\gamma}\int_{|y|^{-1}\le|u|<\sigma_{\min}^{-1}} \frac{1-\cos(\langle u, y\rangle)}{|u|^\gamma}\,\mathrm{d}u.\end{aligned}$$

If either of the domains is empty, we set the integral to 0. Further, if $y = 0$, then both integrals are 0. Thus let $y \neq 0$. Then for the first integral, we can use that $1-\cos(\langle u, y\rangle) \le \frac{\langle u,y\rangle^2}{2} \le \frac{1}{2}|u|^2|y|^2$. Thus, using d_Y-dimensional polar coordinates

$$\int_{|u|<|y|^{-1}} \frac{1-\cos(\langle u, y\rangle)}{|u|^\gamma}\,\mathrm{d}u \lesssim |y|^2 \int_0^{|y|^{-1}} r^{2-d} r^{d-1}\,\mathrm{d}r = \frac{1}{4}.$$

For the second integral, we obtain again by d_Y-dimensional polar coordinates

$$\begin{aligned}\int_{|y|^{-1}\le|u|<\sigma_{\min}^{-1}} \frac{1-\cos(\langle u, y\rangle)}{|u|^\gamma}\,\mathrm{d}u &\le 2\int_{|y|^{-1}\le|u|<\sigma_{\min}^{-1}} \frac{1}{|u|^\gamma}\,\mathrm{d}u \\ &\lesssim \int_{|y|^{-1}}^{\sigma_{\min}^{-1}} r^{d-1} r^{-d}\,\mathrm{d}r \\ &= \log\Big(\frac{|y|}{\sigma_{\min}}\Big).\end{aligned}$$

Using the concavity of the logarithm and Jensen's inequality, we obtain the following bound in case $\gamma = d_Y$

$$\sigma_{\min}^{\gamma-d_Y}\int_{|t|<1} \frac{|\mathcal{F}K(t)|^2}{|t|^\gamma}\mathbb{E}_{X_i}\Big[\sum_{i=1}^n \mathbb{E}_{\bar{X}_1}[w_i^2(\bar{X}_1)]\Big(1-\Big|\varphi_{X_i}\Big(\frac{t}{\sigma_{\min}}\Big)\Big|^2\Big)\Big]\,\mathrm{d}u \lesssim n^{-1}h_x^{-d_X}(1+\log(\sigma_{\min}^{-1})).$$

Thus, the bound for the first term is given by

$$\mathbb{E}_{(X_i,Y_i),\bar{X}_1}\Big[\int \frac{\mathcal{F}K(\sigma_{\min}u)\sum_{i=1}^n w_i(\bar{X}_1)(e^{\mathrm{i}\langle u,Y_i\rangle}-\varphi_{X_i}(u))}{|u|^\gamma}\,\mathrm{d}u \lesssim \begin{cases}\sigma_{\min}^{\gamma-d_Y} n^{-1}h_X^{-d_X}, & \gamma < d_Y,\\ n^{-1}h_X^{-d_X}\log(\sigma_{\min}^{-1}), & \gamma = d_Y,\\ n^{-1}h_X^{-d_X}, & \gamma > d_Y.\end{cases}$$

For the second term (6.16), we obtain using Jensen's inequality for every $x \in \mathbb{R}^{d_X}$

$$\begin{aligned}\int \frac{|\sum_{i=1}^n w_i(x)\mathcal{F}K(\sigma_{\min}u)(\varphi_{X_i}(u)-\varphi_x(u))|^2}{|u|^\gamma}\,\mathrm{d}u &\le \sum_{i=1}^n w_i(x)\int \frac{|\varphi_{X_i}(u)-\varphi_x(u)|^2}{|u|^\gamma}\,\mathrm{d}u \\ &\le \sum_{i=1}^n w_i(x)L^2|X_i-x|^{2\alpha} \\ &\lesssim h_x^{2\alpha},\end{aligned}$$

where the last inequality used the compact support of the kernel K^x.
For the third term (6.17), we obtain for every $x \in \mathbb{R}^{d_X}$

$$\begin{aligned}\int \frac{|\mathcal{F}(K(\sigma_{\min}u)-1)\varphi_x(u)|^2}{|u|^\gamma}\,\mathrm{d}u &= \int \frac{|\mathcal{F}(K(\sigma_{\min}u)-1)|^2}{|u|^\gamma}\frac{|\varphi_x(u)|^2(1+|u|^2)^s}{(1+|u|^2)^s}\,\mathrm{d}u\\ &\le \Gamma \sup_{u\in\mathbb{R}\setminus\{0\}} |\mathcal{F}(K(\sigma_{\min}u)-1)|^2|u|^{-\gamma}(1+|u|^2)^{-s}\\ &= \Gamma\sigma_{\min}^{\gamma+2s}\sup_{\frac{z}{\sigma_{\min}}\in\mathbb{R}\setminus\{0\}} |\mathcal{F}(K(z)-1)|^2|z|^{-\gamma}(\sigma_{\min}^s+|z|^2)^{-s}\\ &\le \Gamma\sigma_{\min}^{\gamma+2s}\sup_{\frac{z}{\sigma_{\min}}\in\mathbb{R}\setminus\{0\}} |\mathcal{F}(K(z)-1)|^2|z|^{-\gamma-2s}.\end{aligned}$$

If K is of order $\lceil s+\gamma/2\rceil$, then

$$\int \frac{|\mathcal{F}(K(\sigma_{\min}u)-1)\varphi_x(u)|^2}{|u|^\gamma}\,\mathrm{d}u \lesssim \Gamma\sigma_{\min+\gamma}^{\gamma+2s}.$$

Collecting all terms leads to the bound

$$\mathbb{E}_{(X_i,Y_i),\bar{X}_1}\Big[\int \frac{|\mathring{\varphi}_x(u)-\varphi_x(u)|^2}{|u|^\gamma}\,\mathrm{d}u\Big] \lesssim \sigma_{\min}^{2s+\gamma} + h_x^{2\alpha} + \begin{cases}\sigma_{\min}^{\gamma-d_Y}n^{-1}h_X^{-d_X}, & \gamma<d_Y,\\ n^{-1}h_X^{-d_X}\log(\sigma_{\min}^{-1}), & \gamma=d_Y,\\ n^{-1}h_X^{-d_X}, & \gamma>d_Y.\end{cases}$$

In case $\gamma < d$, we choose $\sigma_{\min} \asymp h_x^{2\alpha}$ and $h_x \asymp n^{-\frac{\gamma+2s}{2(d_X s+d_Y\alpha)+d_X\gamma+2\alpha s}}$. In case $\gamma > d$, we set $\sigma_{\min} > 0$ arbitrarily small and $h_x \asymp n^{-\frac{1}{2\alpha+d_X}}$. In case $\gamma = d$, we set $\sigma_{\min} \asymp n^{-\frac{1}{2d+\gamma}}$ and $h_x \asymp n^{-\frac{1}{2\alpha+d_X}}$. Then

$$\mathbb{E}_{(X_i,Y_i),\bar{X}_1}\Big[\int \frac{|\mathring{\varphi}_x(u)-\varphi_x(u)|^2}{|u|^\gamma}\,\mathrm{d}u\Big] \lesssim \begin{cases} n^{-\frac{2}{2+\frac{d_X}{\alpha}}}, & \gamma>d_Y,\\ n^{-\frac{2}{2+\frac{d_X}{\alpha}}}\log(n), & \gamma=d_Y,\\ n^{-\frac{2+\frac{\gamma}{s}}{2+\frac{d_Y}{s}+\frac{d_X}{\alpha}+\frac{d_X\gamma}{2\alpha s}}}, & \gamma<d_Y.\end{cases}$$

This concludes the proof. □

Proof of Theorem 6.8. In order to avoid confusion with the conditional probability path $p_t(\cdot|\cdot)$, we will denote the time depended conditional density of the Flow Matching estimator by $\hat{p}_{t,x}$ and the NW estimator by $\mathring{p}_{t,x}$ for $x \in \mathbb{R}^{d_X}$. These are the densities of (6.5) and (6.6). For $t = 1$ we obtain the density whose characteristic functions are $\hat{\varphi}_x$ and $\mathring{\varphi}_x$, respectively.
We start using the triangle inequality

$$\begin{aligned}&\mathbb{E}_{(X_i,Y_i)}\Big[\mathbb{E}_{\bar{X}_1}\Big[\int \frac{|\hat{\varphi}_{\bar{X}_1}(u)-\varphi_{\bar{X}_1}(u)|^2}{|u|^\gamma}\,\mathrm{d}u\Big]\Big]\\ &\lesssim \mathbb{E}_{(X_i,Y_i)}\Big[\mathbb{E}_{\bar{X}_1}\Big[\int \frac{|\hat{\varphi}_{\bar{X}_1}(u)-\mathring{\varphi}_{\bar{X}_1}(u)|^2}{|u|^\gamma}\,\mathrm{d}u\Big]\Big] + \mathbb{E}_{(X_i,Y_i)}\Big[\mathbb{E}_{\bar{X}_1}\Big[\int \frac{|\mathring{\varphi}_{\bar{X}_1}(u)-\varphi_{\bar{X}_1}(u)|^2}{|u|^\gamma}\,\mathrm{d}u\Big]\Big].\end{aligned}$$

The first term corresponds to the difference between the Flow Matching model and the Nadaraya-

Watson estimator, making it an approximation error. The second term is the risk of the Nadaraya-Watson estimator itself. For the second term, we can use Theorem 6.6. However, we need to take into account that the Gaussian kernel, which corresponds to the choice of $\mathbb{U} = \mathcal{N}(0, I_{d_Y})$ for the latent distribution, is only a kernel of order 2. We will come back to this issue in Corollary 6.9.

In order to bound the approximation error by the term minimized in (6.4), we first need to derive a bound which is accessible to Grönwalls lemma. In case of the energy score, we have that

$$\begin{aligned}
&\mathbb{E}_{(X_i,Y_i)}\Big[\mathbb{E}_{\bar{X}_1}\Big[\int \frac{|\hat{\varphi}_{\bar{X}_1}(u) - \mathring{\varphi}_{\bar{X}_1}(u)|^2}{|u|^\gamma}\,\mathrm{d}u\Big]\Big] \\
&\quad= \mathbb{E}_{(X_i,Y_i)}\Big[\mathbb{E}_{\bar{X}_1}\Big[\mathbb{E}_{\substack{Y\sim\hat{p}_{1,\bar{X}_1}\\ Z\sim\mathring{p}_{1,\bar{X}_1}}}[|Y-Z|^\beta] - \frac{1}{2}\mathbb{E}_{Y,Y'\overset{\text{iid}}{\sim}\hat{p}_{1,\bar{X}_1}}[|Y-Y'|^\beta] \\
&\qquad - \frac{1}{2}\mathbb{E}_{Z,Z'\overset{\text{iid}}{\sim}\mathring{p}_{1,\bar{X}_1}}[|Z-Z'|^\beta]\Big]\Big] \\
&\quad\leq \mathbb{E}_{(X_i,Y_i)}\Big[\mathbb{E}_{\bar{X}_1}\Big[\mathbb{E}_{\substack{Y\sim\hat{p}_{1,\bar{X}_1}\\ Z\sim\mathring{p}_{1,\bar{X}_1}}}[|Y-Z|^\beta]\Big]\Big].
\end{aligned}$$

As $\beta \in (0,2)$, we can use Jensen's inequality to obtain

$$\mathbb{E}_{\substack{Y\sim\hat{p}_{1,\bar{X}_1}\\ Z\sim\mathring{p}_{1,\bar{X}_1}}}[|Y-Z|^\beta] \leq \mathbb{E}_{\substack{Y\sim\hat{p}_{1,\bar{X}_1}\\ Z\sim\mathring{p}_{1,\bar{X}_1}}}[|Y-Z|^2]^{\frac{\beta}{2}}.$$

If the functions in the set $\mathcal{M}$ are Lipschitz continuous in the d_Y-dimensional component for fixed $t \in [0,1]$ and $x \in \mathbb{R}^{d_X}$ with Lipschitz constant $\hat{\Gamma}_t$, we can proceed analogously to the proof of Theorem 5.8 and get

$$\mathbb{E}_{\substack{Y\sim\hat{p}_{1,\bar{X}_1}\\ Z\sim\mathring{p}_{1,\bar{X}_1}}}[|Y-Z|^2]^{\frac{\beta}{2}} \leq \Big(2e^{\int_0^1 1+2\hat{\Gamma}_t \mathrm{d}t}\int_0^1\int \left|\hat{v}\left(z,\bar{X}_1\right) - \mathring{v}_t\left(z,\bar{X}_1\right)\right|^2 \mathring{p}_{t,\bar{X}_1}(z)\,\mathrm{d}z\,\mathrm{d}t\Big)^{\frac{\beta}{2}}.$$

Further,

$$\begin{aligned}
&\mathbb{E}_{\bar{X}_1}\Big[\Big(\int_0^1\int \left|\hat{v}\left(z,\bar{X}_1\right) - \mathring{v}_t\left(z,\bar{X}_1\right)\right|^2 \mathring{p}_{t,\bar{X}_1}(z)\,\mathrm{d}z\,\mathrm{d}t\Big)^{\frac{\beta}{2}}\Big] \\
&\qquad\leq \Big(\mathbb{E}_{\bar{X}_1}\Big[\int_0^1\int \left|\hat{v}\left(z,\bar{X}_1\right) - \mathring{v}_t\left(z,\bar{X}_1\right)\right|^2 \frac{\mathring{p}_t(z,\bar{X}_1)}{\frac{1}{n}\sum_{i=1}^n K^{\mathrm{x}}_{h_x}(\bar{X}_1 - X_i)}\,\mathrm{d}z\,\mathrm{d}t\Big]\Big)^{\frac{\beta}{2}} \\
&\qquad= \Big(\int\int_0^1\int |\hat{v}(z,x) - \mathring{v}_t(z,x)|^2 \frac{\mathring{p}_t(z,x)}{\frac{1}{n}\sum_{i=1}^n K^{\mathrm{x}}_{h_x}(x - X_i)}\,\mathrm{d}z\,\mathrm{d}t\,p_X(x)\,\mathrm{d}x\Big)^{\frac{\beta}{2}}.
\end{aligned}$$

Since we chose the Gaussian kernel for the covariates, the previous expression is well-defined. Then due to the support of

$$\frac{1}{n}\sum_{i=1}^n K^{\mathrm{x}}_{h_x}(x - X_i) \geq K^{\mathrm{x}}_{h_x}(\vec{1}) = \exp\Big(-\frac{d^2}{2h_x^2}\Big),$$

where $\vec{1}$ is the $\mathbb{R}^p$-valued vector with entry 1 in every component. Thus,

$$\Big(\int\int_0^1\int |\hat{v}\,(z,x)-\mathring{v}_t\,(z,x)|^2\,\frac{\mathring{p}_t(z,x)}{\sum_{i=1}^n K_{h_x}^{\mathrm{x}}(x-X_i)}\,\mathrm{d}z\,\mathrm{d}t\,p_X(x)\,\mathrm{d}x\Big)^{\frac{\beta}{2}}$$
$$\lesssim \exp\Big(\frac{d^2}{2h_x^2}\cdot\frac{\beta}{2}\Big)\Big(\int\int_0^1\int |\hat{v}\,(z,x)-\mathring{v}_t\,(z,x)|^2\,\mathring{p}_t(z,x)\,\mathrm{d}z\,\mathrm{d}t\,\mathrm{d}x\Big)^{\frac{\beta}{2}}.$$

By the choice of $\hat{v}$ via (6.4) and Lemma 6.2, we know that for every $\tilde{v}\in\mathcal{M}$ and any $a>0$

$$\int\int_0^1\int |\hat{v}\,(z,x)-\mathring{v}_t\,(z,x)|^2\,\mathring{p}_t(z,x)\,\mathrm{d}z\,\mathrm{d}t\,\mathrm{d}x \le \int\int_0^1\int |\tilde{v}\,(z,x)-\mathring{v}_t\,(z,x)|^2\,\mathring{p}_t(z,x)\,\mathrm{d}z\,\mathrm{d}t\,\mathrm{d}x$$
$$\le |\tilde{v}-\mathring{v}|_{\infty,[0,1]\times[-a,a]^{d_Y+d_X}} + \int_0^1\int_{\mathbb{R}^d\setminus[-a,a]^{d_Y+d_X}} |\tilde{v}\,(z,x)-\mathring{v}_t\,(z,x)|^2\,\mathring{p}_t(z,x)\,\mathrm{d}(z,x)\,\mathrm{d}t. \tag{6.18}$$

For the set $\mathcal{M}$ we only consider functions whose Lipschitz constants are not too far from the Lipschitz constant of $\mathring{v}$ on $[-a,a]^{d_Y+d_X}$. Outside of this area, we cut the network at the maximal absolute value of $\mathring{v}$ on $[-a,a]^{d_Y+d_X}$. To bound the first term and to show that there are ReLU functions satisfying this assumption, we are going to use Theorem 2.6. For the second term we are going to exploit the tail-behavior of $\mathring{p}_t$ combined with the boundedness of $\tilde{v}$ and $\mathring{v}$.
For both, we first need to evaluate the corresponding properties of $\mathring{v}$.

Lemma 6.15. *For every $z\in(-a,a)^{d_Y}$ and $x\in\mathbb{R}^{d_X}$ we have that*

$$|\mathring{v}_t(z,x)| \le \frac{\sqrt{d_Y}(1+a)}{1-(1-\sigma_{\min})t} \le \frac{\sqrt{d_Y}(1+a)}{\sigma_{\min}}.$$

Further, for every $x\in\mathbb{R}^{d_X}$ and every $t\in[0,1]$

$$\mathrm{Lip}(\mathring{v}_t(\cdot,x)) \le \frac{1}{\sigma_t}+\frac{2d}{\sigma_t^3}.$$

Now we can determine the value of a needed for the second term in (6.18). Using the first bound in Lemma 6.15 for a supremum norm bound of both vector fields, we obtain

$$\int_0^1\int_{\mathbb{R}^d\setminus[-a,a]^{d_Y+d_X}} |\tilde{v}\,(z,x)-\mathring{v}_t\,(z,x)|^2\,\mathring{p}_t(z,x)\mathrm{d}(z,x)\mathrm{d}t$$
$$\lesssim \int_0^1\int_{\mathbb{R}^d\setminus[-a,a]^{d_Y+d_X}} \frac{|z|^2+t^2d}{\sigma_t^2} f_t(z,x)\mathrm{d}(z,x)\mathrm{d}t.$$

For the integral, we first note that for general $b,c>0$ by the change of variables formula

$$\int_{|z|\ge b}\exp\Big(-\frac{|z|^2}{c}\Big)\mathrm{d}z = \frac{2\pi^{d/2}}{\Gamma(d/2)}\int_b^\infty r^{d-1}\exp\Big(-\frac{r^2}{c}\Big)\,\mathrm{d}r$$
$$= \frac{2\pi^{d/2}}{\Gamma(d/2)}\frac{c^{d/2}}{2}\int_{b^2/c}^\infty u^{\frac{d}{2}-1}e^{-u}\,\mathrm{d}u$$
$$= \frac{\pi^{d/2}}{\Gamma(d/2)}c^{d/2}\Gamma\left(\frac{d}{2},\frac{b^2}{c}\right),$$

where Γ is the upper incomplete Gamma function as defined in Gabcke (1979, Satz 4.4.3). Due to the universal notation of the Gamma function, we accept this double usage of Γ for both the Lipschitz constant and the Gamma function. Similarly

$$\int_{|z|\geq b} |z|^2 \exp\Big(-\frac{|z|^2}{c}\Big)\mathrm{d}z = \frac{\pi^{d/2}}{\Gamma(d/2)} \cdot c^{\frac{d+2}{2}}\Gamma\big(\frac{d}{2}+1, \frac{b^2}{c}\big).$$

For $i \in \{1, ..., n\}$, we obtain for the integration over x

$$\begin{aligned}
\int_{|x|_\infty > a} K^{\mathrm{x}}_{h_x}(x - X_i)\,\mathrm{d}x &= \int_{|x|_\infty > a} \frac{\exp\big(-\frac{|x-X_i|^2}{2h_x^2}\big)}{(2\pi h_x^2)^{d_X/2}}\,\mathrm{d}x \\
&\leq \int_{|x|_\infty > a-1} \frac{\exp\big(-\frac{|x|^2}{2h_x^2}\big)}{(2\pi h_x^2)^{d_X/2}}\,\mathrm{d}x \\
&\leq \frac{1}{(2\pi h_x^2)^{d_X/2}} \int_{|x|>a-1} \exp\Big(-\frac{|x|^2}{2h_x^2}\Big)\,\mathrm{d}x \\
&\leq \frac{1}{\Gamma(d_X/2)}\Gamma\Big(\frac{d_X}{2}, \frac{(a-1)^2}{2h_x^2}\Big).
\end{aligned}$$

For the integration over z, we get

$$\begin{aligned}
\int_{|z|_\infty > a} \frac{|z|^2 + t^2 d}{\sigma_t^2} \frac{\exp\big(-\frac{|z-Y_i|^2}{2\sigma_t^2}\big)}{(2\pi\sigma_t^2)^{d_Y/2}}\,\mathrm{d}z &= \int_{|u+Y_i|_\infty > a} \frac{|u+Y_i|^2 + t^2 d}{\sigma_t^2} \frac{\exp\big(-\frac{|u|^2}{2\sigma_t^2}\big)}{(2\pi\sigma_t^2)^{d_Y/2}}\,\mathrm{d}u \\
&\leq 2\int_{|u|_\infty > a-1} \frac{|u|^2 + 1 + t^2 d}{\sigma_t^2} \frac{\exp\big(-\frac{|u|^2}{2\sigma_t^2}\big)}{(2\pi\sigma_t^2)^{d_Y/2}}\,\mathrm{d}u \\
&\leq \frac{4}{\Gamma(d_Y/2)}\Gamma\Big(\frac{d_Y}{2}+1, \frac{(a-1)^2}{2\sigma_t^2}\Big) \\
&\quad + 2\frac{t^2 d}{\sigma_t^2}\frac{1}{\Gamma(d_Y/2)}\Gamma\Big(\frac{d_Y}{2}, \frac{(a-1)^2}{2\sigma_t^2}\Big).
\end{aligned}$$

If $\min(\frac{(a-1)^2}{2\sigma_t^2}, \frac{(a-1)^2}{2h_x^2}) \geq \max(\frac{d_X}{2}, \frac{d_Y}{2}+1)$ we can use Gabcke (1979, Satz 4.4.3), to bound

$$\begin{aligned}
\Gamma\Big(\frac{d_X}{2}, \frac{(a-1)^2}{2h_x^2}\Big) &\leq \frac{d_X}{2} e^{-\frac{(a-1)^2}{2h_x^2}} \Big(\frac{(a-1)^2}{2h_x^2}\Big)^{\frac{d_X}{2}-1}, \\
\Gamma\Big(\frac{d_Y}{2}, \frac{(a-1)^2}{2\sigma_t^2}\Big) &\leq \frac{d_Y}{2} e^{-\frac{(a-1)^2}{2\sigma_t^2}} \Big(\frac{(a-1)^2}{2\sigma_t^2}\Big)^{\frac{d_Y}{2}-1}, \\
\Gamma\Big(\frac{d_Y}{2}+1, \frac{(a-1)^2}{2\sigma_t^2}\Big) &\leq \Big(\frac{d_Y}{2}+1\Big) e^{-\frac{(a-1)^2}{2\sigma_t^2}} \Big(\frac{(a-1)^2}{2\sigma_t^2}\Big)^{\frac{d_Y}{2}}.
\end{aligned}$$

Hence we obtain for $a > 1$

$$\begin{aligned}
&\exp\Big(\frac{d^2}{2h_x^2}\cdot\frac{\beta}{2}\Big)\Big(\int_0^1 \int_{\mathbb{R}^d\setminus[-a,a]^{d_Y+d_X}} |\tilde{v}(z,x) - \mathring{v}_t(z,x)|^2 \mathring{p}_t(z,x)\,\mathrm{d}(z,x)\,\mathrm{d}t\Big)^{\beta/2} \\
&\qquad \lesssim \exp\Big(\frac{d^2}{2h_x^2}\cdot\frac{\beta}{2}\Big)\Big(\int_0^1 e^{-\frac{(a-1)^2}{2h_x^2}}\Big(\frac{(a-1)^2}{2h_x^2}\Big)^{\frac{d_X}{2}-1} e^{-\frac{(a-1)^2}{2\sigma_t^2}}\Big(\frac{(a-1)^2}{2\sigma_t^2}\Big)^{\frac{d_Y}{2}-1}
\end{aligned}$$

$$+ e^{-\frac{(a-1)^2}{2h_x^2}}\Big(\frac{(a-1)^2}{2h_x^2}\Big)^{\frac{d_X}{2}-1}\frac{1}{\sigma_t^2}e^{-\frac{(a-1)^2}{2\sigma_t^2}}\Big(\frac{(a-1)^2}{2\sigma_t^2}\Big)^{\frac{d_Y}{2}}\,\mathrm{d}t\Big)^{\beta/2}$$
$$\lesssim \exp\Big(\frac{d^2}{2h_x^2}\cdot\frac{\beta}{2}\Big)\Big(\int_0^1 e^{-\frac{(a-1)^2}{2h_x^2}-\frac{(a-1)^2}{2\sigma_t^2}}\frac{(a-1)^{d_X+d_Y-2}}{h_x^{d_X-2}\sigma_t^{d_Y}}\,\mathrm{d}t\Big)^{\beta/2}.$$

Since

$$\begin{aligned}\int_0^1 e^{-\frac{(a-1)^2}{2\sigma_t^2}}\frac{1}{\sigma_t^{d_Y}}\,\mathrm{d}t &= \int_1^{\frac{1}{\sigma_{\min}}}\exp\Big(-\frac{(a-1)^2u}{2}\Big)\frac{u^{d_Y-2}}{1-\sigma_{\min}}\,\mathrm{d}u\\ &\leq \int_1^{\frac{1}{\sigma_{\min}}}\exp\Big(-\frac{(a-1)^2u^2}{2}\Big)u^{d_Y/2-2}\,\mathrm{d}u\\ &\lesssim (a-1)^{-\frac{d_Y-2}{2}}\Gamma\Big(\frac{d_Y-2}{4},\frac{(a-1)^2}{2}\Big)\\ &\lesssim (a-1)^{-\frac{d_Y-2}{2}}e^{-\frac{(a-1)^2}{2}}\\ &\leq e^{-\frac{(a-1)^2}{2}},\end{aligned}$$

we can further bound for $d \leq a-1$

$$\begin{aligned}&\exp\Big(\frac{d^2}{2h_x^2}\cdot\frac{\beta}{2}\Big)\Big(\int_0^1 e^{-\frac{(a-1)^2}{2h_x^2}-\frac{(a-1)^2}{2\sigma_t^2}}\frac{(a-1)^{d_X+d_Y-2}}{h_x^{d_X-2}\sigma_t^{d_Y}}\mathrm{d}t\Big)^{\beta/2}\\ &\qquad\lesssim \exp\Big(\frac{d^2-(a-1)^2}{2h_x^2}\cdot\frac{\beta}{2}\Big)e^{-\frac{(a-1)^2\beta}{4}}\Big(\frac{(a-1)^{d_X+d_Y-2}}{h_x^{d_X-2}}\Big)^{\beta/2}\\ &\qquad\leq e^{-\frac{(a-1)^2\beta}{4}}\Big(\frac{(a-1)^{d_X+d_Y-2}}{h_x^{d_X-2}}\Big)^{\beta/2}.\end{aligned}$$

For the approximation error, we thus obtain

$$\begin{aligned}&\mathbb{E}_{\bar{X}_1}\Big[\int|\hat{\varphi}_{\bar{X}_1}(u)-\mathring{\varphi}_{\bar{X}_1}(u)|^2\frac{\mathrm{d}u}{|u|^\gamma}\Big]\\ &\quad\lesssim e^{\frac{\beta}{2}\int_0^1 2\hat{\Gamma}_t\mathrm{d}t}\Big(\exp\Big(\frac{d^2}{2h_x^2}\cdot\frac{\beta}{2}\Big)|\tilde{v}-\mathring{v}|^{\beta/2}_{\infty,[0,1]\times[-a,a]^{d_Y+d_X}}+e^{-\frac{(a-1)^2\beta}{4}}\Big(\frac{(a-1)^{d_X+d_Y-2}}{h_x^{d_X-2}}\Big)^{\beta/2}\Big).\end{aligned}$$

From Lemma 6.15 we know that the Lipschitz constant of $\mathring{v}$ in the d_Y-dimensional component is bounded by $\frac{2d+1}{\sigma_t^3}$. Then

$$\int_0^1\frac{1}{\sigma_t^3}\,\mathrm{d}t=\int_0^1\frac{1}{(1-(1-\sigma_{\min})\,t)^3}\,\mathrm{d}t\leq\frac{1}{2\,(1-\sigma_{\min})\,\sigma_{\min}^2}-\frac{1}{2\,(1-\sigma_{\min})}\leq\frac{1}{\sigma_{\min}^2}.$$

To bound the tail aiming for a rate n^{-C}, we need to assure that

$$e^{\frac{\beta(4d+2)}{2\sigma_{\min}^2}-\frac{(a-1)^2\beta}{4}}\Big(\frac{(a-1)^{d_X+d_Y-2}}{h_x^{d_X-2}}\Big)^{\beta/2}\leq n^{-C},$$

which is the case if

$$\frac{4d+2}{2\sigma_{\min}^2}+\frac{2}{\beta}C\log(n)-(d_X-2)\log(h_x)\leq\frac{(a-1)^2}{2}-(d_X+d_Y-2)\log(a-1).$$

For large a and small $\sigma_{\min}$ we can choose

$$a \gtrsim \max(\log(n) + \sigma_{\min}^{-1} - \log(h_x), d-1).$$

Now we need to find a neural network such that $|\tilde{v} - v|_{\infty,[0,1]\times[-a,a]^{d_Y+d_X}} \leq n^{-C}$, which is cut off at the boundaries of $[-a,a]^d$ and which approximates the Lipschitz constant of v simultaneously with the same error. We proceed like in the proof of Theorem 5.8 but need to bound the second partial derivatives of $\mathring{v}$ in z, x and t. In view of Lemma 6.15, all of them are bounded $\sigma_{\min}^{-m_1} h_x^{-m_2}$, where $m_1, m_2 \in \mathbb{N}$ independent of n. Thus we need to choose the approximation error of the network such that

$$e^{\frac{\beta(4d+2)}{2\sigma_{\min}^2} + \frac{d^2}{2h_x^2}\cdot\frac{\beta}{2}} \varepsilon^{\beta/2} \leq n^{-C} \cdot \sigma_{\min}^{m_1} h_x^{m_2} \quad \text{equivalent to} \quad \varepsilon \leq n^{-\frac{2C}{\beta}} \sigma_{\min}^{\frac{2m_1}{\beta}} h_x^{\frac{2m_2}{\beta}} e^{-\frac{4d+2}{\sigma_{\min}^2} - \frac{d^2}{2h_x^2}}.$$

We choose $\sigma_{\min}$ and h_x as in Theorem 6.6. Hence, setting $\log(\sigma_{\min}) \asymp \log(n)$ and $\log(h_x) \asymp \log(n)$, by Theorem 2.6, there is a ReLU network with

$$\begin{aligned} L &\gtrsim \operatorname{poly}(n) \cdot (\log^2(n) + 1), \\ M &\gtrsim \operatorname{poly}(n) \cdot e^{\operatorname{poly}(n)} \cdot (\log^2(n) + 1), \end{aligned}$$

where $\operatorname{poly}(n)$ is a polynomial in n. Combined with Theorem 6.6, we obtain the result.

□

Proof of Corollary 6.9. In order to maintain coherence when comparing to the proof of Theorem 6.6, we go back to the notation of the kernel K instead of $p_t(\cdot|\cdot)$. First we note that for $\ell \in [0,1]$ and $z \in \mathbb{R}^{d_Y}$

$$\begin{aligned} |\mathcal{F}K(z) - 1| &= \Big| \int (e^{\mathrm{i}\langle t,z\rangle} - 1) K(t) \, \mathrm{d}t \Big| \\ &\leq \int |e^{\mathrm{i}\langle t,z\rangle} - 1| K(t) \, \mathrm{d}t \\ &\leq \int 2 \big| \sin(\frac{\langle t,z\rangle}{2}) \big| K(t) \, \mathrm{d}t \\ &\leq 2|z|^\ell \int |t|^\ell K(t) \, \mathrm{d}t. \end{aligned}$$

The integral is finite since we chose the Gaussian kernel for K. Then for $w \in \mathbb{R}$ by Taylor

$$e^{\mathrm{i}w} = 1 + \mathrm{i}w + \int_0^w (w-t)(-e^{\mathrm{i}t}) \, \mathrm{d}t,$$

which implies

$$\big| e^{\mathrm{i}w} - 1 - \mathrm{i}w \big| \leq \int_0^{|w|} (|w| - t) \, \mathrm{d}t = \frac{|w|^2}{2}.$$

Additionally

$$\big| e^{\mathrm{i}w} - 1 - \mathrm{i}w \big| \leq 2 + |w|.$$

Thus, for $\ell \in (1,2]$, using $\int uK(t)\,\mathrm{d}u = 0$,

$$\begin{aligned}
|\mathcal{F}K(z) - 1| &= \Big| \int (e^{\mathrm{i}\langle t,z\rangle} - 1)K(t)\,\mathrm{d}t \Big| \\
&= \Big| \int (e^{\mathrm{i}\langle t,z\rangle} - 1 - \mathrm{i}t^\top u)K(t)\,\mathrm{d}t \Big| \\
&\leq 3 \int |\langle t,z\rangle|^\ell K(t)\,\mathrm{d}t \\
&\leq 3|z|^\ell \int |t|^\ell K(t)\,\mathrm{d}t.
\end{aligned}$$

If $s - \frac{\gamma}{2} \geq 0$ and $s \leq 2$, then

$$\begin{aligned}
\int \frac{|(\mathcal{F}K(\sigma_{\min}u) - 1)\varphi_x(u)|^2}{|u|^\gamma}\,\mathrm{d}u &\lesssim \sigma_{\min}^{2s} \int |u|^{2s-\gamma}|\varphi_x(u)|^2\,\mathrm{d}u \\
&\leq \sigma_{\min}^{2s} \int (1+|u|^2)^{s-\frac{\gamma}{2}}|\varphi_x(u)|^2\,\mathrm{d}u \\
&\leq \sigma_{\min}^{2s} \int (1+|u|^2)^{s}|\varphi_x(u)|^2\,\mathrm{d}u \\
&\lesssim \sigma_{\min}^{2s}.
\end{aligned}$$

The rest of the proof follows analogously to Theorem 6.6 and Theorem 6.8. □

6.5.3. Additional proofs of Section 6.5

Proof of Lemma 6.11.

1. Since $\int h(x)\,\mathrm{d}x = 0$, we only need to ensure that the supremum norm of the perturbations are bounded by 1. However, it will be useful to bound $p_{Y|X}$ from below, hence we bound the supremum norm of the perturbations by $\frac{1}{2}$. In fact, by the disjoint support of the bumps, we can bound

$$\Big| \rho \sum_{\bar{i}} \sum_{\bar{j}} \Delta_{\bar{i}\bar{j}} \prod_{k\in[d_Y]} h_{i_k,r}(y_k) \prod_{k\in[d_X]} h_{j_k,m}(x_k) \Big|_\infty \leq \rho r^{d_Y/2} \|h\|_\infty^{d_Y+d_X}.$$

2. First, we calculate the absolute value of the Fourier transform $\varphi_{\bar{i},r}$ of $\prod_{k\in[d_Y]} h_{i_k,r}(y_k)$ for fixed r and fixed $\bar{i} \in \{1,...,r\}^{d_Y}$,

$$\begin{aligned}
|\varphi_{\bar{i},r}(u)| &= \Big| \int e^{\mathrm{i}\langle u,y\rangle} \prod_{k\in[d_Y]} h_{i_k,r}(y_k)\,\mathrm{d}y \Big| \\
&= r^{d_Y/2} \Big| \prod_{k\in[d_Y]} \int e^{\mathrm{i}u_k,y_k} h_{i_k,r}(y_k)\,\mathrm{d}y_k \Big| \\
&= r^{-d_Y/2} \big|\Phi\big(\frac{u}{r}\big)\big|,
\end{aligned}$$

where $\Phi(v) \coloneqq \prod_{k\in[d_Y]} \varphi_h(v_k)$ for $v \in \mathbb{R}^{d_Y}$.

Since for $r \geq 1$

$$\begin{aligned}\int (1+|u|^2)^s |\varphi_{\bar{i},r}(u)|^2 \,\mathrm{d}u &= r^{-d_Y} \int (1+|u|^2)^s \big|\Phi\big(\frac{u}{r}\big)\big|^2 \,\mathrm{d}u \\ &= \int (1+|zr|^2)^s \big|\Phi(z)\big|^2 \,\mathrm{d}z \\ &\leq r^{2s} \int (1+|z|^2)^s \big|\Phi(z)\big|^2 \,\mathrm{d}z,\end{aligned}$$

where the last term is bounded by a constant by assumption on h. Then for fixed $x \in \mathbb{R}^{d_X}$

$$\begin{aligned}\|p^{\Delta}_{Y|X}(\cdot|x)\|^2_{\mathcal{H}^s} &= \big\|\mathbb{1}_{[0,1]^{d_Y}} + \rho \sum_{\bar{i}} \sum_{\bar{j}} \Delta_{\bar{i},\bar{j}} \prod_{k\in[d_X]} h_{j_k,m}(x_k) \prod_{k\in[d_Y]} h_{i_k,r}\big\|^2_{\mathcal{H}^s} \\ &\leq \big\|\mathbb{1}_{[0,1]^{d_Y}}\big\|^2_{\mathcal{H}^s} + \|\rho \sum_{\bar{i}} \sum_{\bar{j}} \Delta_{\bar{i},\bar{j}} \prod_{k\in[d_X]} h_{j_k,m}(x_k) \prod_{k\in[d_Y]} h_{i_k,r}\big\|^2_{\mathcal{H}^s} \\ &\lesssim 1 + \rho^2 \sum_{\bar{i}} \sum_{\bar{j}} \big\| \prod_{k\in[d_X]} h_{j_k,m}(x_k)\big\|^2_{\mathcal{H}^s} r^{2s} \\ &\lesssim 1 + \rho^2 r^{2s+d_Y}.\end{aligned}$$

where the penultimate and the last inequality follow from the disjoint support.

3. Again, due to the disjoint support and the bound on $|\varphi_{\bar{i},r}|$ from the previous part of the proof, we have for $x, x' \in \mathbb{R}^{d_X}$

$$\|\varphi^{\Delta}_x - \varphi^{\Delta}_{x'}\|^2_{\gamma} = \int \frac{\rho^2}{|u|^{\gamma}} \sum_{\bar{i}} \sum_{\bar{j}} \big| \prod_{k\in[d_x]} h_{j_k,m}(x_k) - \prod_{k\in[d_x]} h_{j_k,m}(x'_k)\big|^2 r^{-d_Y} \big|\Phi\big(\frac{u}{r}\big)\big|^2 \,\mathrm{d}u.$$

Due to the disjoint support, we know that there are only two indices $\bar{j}$ such that the summands are nonzero. Let j^* be one of those indices. Then

$$\begin{aligned}&\Big| \prod_{k\in[d_X]} h_{j^*_k,m}(x_k) - \prod_{k\in[d_X]} h_{j^*_k,m}(x'_k)\Big| \\ &\qquad = \sum_{k=1}^{d_X} \Big(\prod_{i=1}^{k-1} h_{j^*_i,m}(x_i)\Big)(h_{j^*_k,m}(x_k) - h_{j^*_k,m}(x'_k))\Big(\prod_{i=k+1}^{d_X} h_{j^*_i,m}(x'_i)\Big) \\ &\qquad \leq \|h\|^{d_X-1}_{\infty} \sum_{k=1}^{d_X} (h_{j^*_k,m}(x_k) - h_{j^*_k,m}(x'_k)) \\ &\qquad \leq \|h\|^{d_X-1}_{\infty} \sum_{k=1}^{d_X} (2\|h\|_{\infty} \wedge \|h'\|_{\infty} m|x_k - x'_k|) \\ &\qquad \leq \|h\|^{d_X-1}_{\infty} \sum_{k=1}^{d_X} (2\|h\|_{\infty} \vee \|h'\|_{\infty})(1 \wedge m|x_k - x'_k|) \\ &\qquad \leq \|h\|^{d_X-1}_{\infty} \sum_{k=1}^{d_X} (2\|h\|_{\infty} \vee \|h'\|_{\infty}) m^{\alpha}|x_k - x'_k|^{\alpha} \\ &\qquad \lesssim m^{\alpha}|x - x'|^{\alpha}.\end{aligned}$$

The bound for the other index follows analogously. Thus we can further bound

$$\begin{aligned}\|\varphi_x^\Delta - \varphi_{x'}^\Delta\|_\gamma^2 &\lesssim \rho^2 m^{2\alpha}|x-x'|^{2\alpha}\int \frac{\left|\Phi\left(\frac{u}{r}\right)\right|^2}{|u|^\gamma}\,\mathrm{d}u \\ &= \rho^2 m^{2\alpha}|x-x'|^{2\alpha} r^{d_Y-\gamma}\int \frac{\left|\Phi(z)\right|^2}{|z|^\gamma}\,\mathrm{d}z.\end{aligned}$$

The last integral is finite by construction. Thus

$$\|\varphi_x^\Delta - \varphi_{x'}^\Delta\|_\gamma \lesssim \rho m^\alpha r^{d_Y/2-\gamma/2}|x-x'|^\alpha. \qquad \square$$

Proof of Lemma 6.12. For the Kullback-Leibler divergence between the conditional densities $p_{X|Y}^\Delta$ and $p_{X|Y}^{\Delta'}$, we can use the fact that the χ^2-divergence bounds the Kullback-Leibler divergence. For a definition of the χ^2-divergence see Tsybakov (2009, p. 86) and for the relation to the Kullback-Leibler divergence see Tsybakov (2009, Lemma 2.7). This leads to

$$\begin{aligned}\mathsf{KL}(p_{X|Y}^\Delta | p_{X|Y}^{\Delta'}) &\leq \int_{[0,1]^{d_Y}} \frac{(p_{X|Y}^\Delta(y|x) - p_{X|Y}^{\Delta'}(y|x))^2}{p_{X|Y}^{\Delta'}(y|x)}\,\mathrm{d}y \\ &\leq 2\int_{[0,1]^{d_Y}} (p_{X|Y}^\Delta(y|x) - p_{X|Y}^{\Delta'}(y|x))^2\,\mathrm{d}y \\ &\leq 2\rho^2 \sum_{\bar{i}}\sum_{\bar{j}} (\Delta_{\bar{i},\bar{j}} - \Delta'_{\bar{i},\bar{j}}) \prod_{k\in[d_X]} h_{j_k,m}^2(x_k) \int \prod_{k\in d_Y} h_{j_k,r}^2(y_k)\,\mathrm{d}y,\end{aligned}$$

where the last inequality follows from the assumption of Lemma 6.11 No. 1 and the disjoint support of the disturbance functions. Furthermore

$$\begin{aligned}\int \prod_{k\in d_Y} h_{j_k,r}^2(y_k)\,\mathrm{d}y &= \prod_{k\in d_Y}\int h_{j_k,r}^2(y_k)\,\mathrm{d}y_k \\ &= r^{d_Y}\prod_{k\in d_Y}\int h^2(ry_k - i_k + 1)\,\mathrm{d}y_k \\ &= \prod_{k\in d_Y}\int h^2(z)\,\mathrm{d}z_k \\ &= 1.\end{aligned}$$

Thus

$$\mathsf{KL}(p_{X|Y}^\Delta | p_{X|Y}^{\Delta'}) \leq 8\rho^2 r^{d_Y}\sum_{\bar{j}}\prod_{k\in[d_X]} h_{j_k,m}^2(x_k).$$

Integration yields

$$\begin{aligned}\mathsf{KL}(p^\Delta | p^{\Delta'}) &\leq 8\rho^2 r^{d_Y}\int \sum_{\bar{j}}\prod_{k\in[d_X]} h_{j_k,m}^2(x_k) p_X(x)\,\mathrm{d}x \\ &\lesssim \rho^2 r^{d_Y},\end{aligned}$$

where the last inequality follows from the support of the functions $j_{j_k,m}$ and the assumption that

$\|h\|_\infty < \infty$. □

Proof of Lemma 6.13. By the Varshamov-Gilbert construction (Tsybakov, 2009, Lemma 2.9) we know that there exists a set T of densities $p^{\Delta}_{Y|X}$ with $\Delta \in \{\pm 1\}^{d_Y+d_X}$ such that $|T| \geq 2^{r^{d_Y} m^{d_X}/8}$ and $d_H(\Delta, \Delta') \geq \frac{r^{d_y} m^{d_z}}{8}$ for $p^{\Delta}_{Y|X}, p^{\Delta'}_{Y|X} \in T$.
For the difference between the two densities $p^{\Delta}_{Y|X}, p^{\Delta'}_{Y|X} \in T$ we obtain

$$p^{\Delta}_{Y|X}(y|x) - p^{\Delta'}_{Y|X}(x|y) = \rho \sum_{\bar{i}} \sum_{\bar{j}} (\Delta_{\bar{i},\bar{j}} - \Delta'_{\bar{i},\bar{j}}) \prod_{k\in[d_Y]} h_{i_k,r}(y_k) \prod_{k\in[d_X]} h_{j_k,m}(x_k),$$

where $\Delta_{\bar{i},\bar{j}} - \Delta'_{\bar{i},\bar{j}} \in \{\pm 2, 0\}$. We abbreviate $b_{\bar{j}}(x) \coloneqq \prod_{k\in[d_X]} h_{j_k,m}(x_k)$ and $a_{\bar{i}}(y) \coloneqq \prod_{k\in[d_Y]} h_{i_k,r}(y_k)$. Then

$$\begin{aligned}
&|\varphi_x(u)^{\Delta} - \varphi_x(u)^{\Delta'}|^2 \\
&= \rho^2 \Big(\sum_{\bar{i}} \sum_{\bar{j}} (\Delta_{\bar{i},\bar{j}} - \Delta'_{\bar{i},\bar{j}}) \mathcal{F}(a_{\bar{i}})(u) b_{\bar{j}}(x) \Big) \cdot \Big(\sum_{\bar{i}} \sum_{\bar{j}} (\Delta_{\bar{i},\bar{j}} - \Delta'_{\bar{i},\bar{j}}) \overline{\mathcal{F}(a_{\bar{i}})(u)} b_{\bar{j}}(x) \Big).
\end{aligned}$$

In each sum, only the terms with indices in S are nonzero. Additionally, due to the disjoint support of the different bumps, only the product of the $(\bar{i}, \bar{j})$ term with itself can be nonzero. Thus we obtain

$$\begin{aligned}
&\int\int \frac{|\varphi^{\Delta}_x(u) - \varphi^{\Delta'}_x(u)|^2}{|u|^\gamma} \,\mathrm{d}u \; p_X(x) \,\mathrm{d}x \\
&\qquad = \rho^2 \sum_{(\bar{i},\bar{j})\in S} (\Delta_{\bar{i},\bar{j}} - \Delta'_{\bar{i},\bar{j}})^2 \int (b_{\bar{j}}(x))^2 \int \frac{|\mathcal{F}^{-1}(a_{\bar{i}})(u)|^2}{|u|^\gamma} \,\mathrm{d}u \; p_X(x) \,\mathrm{d}x.
\end{aligned}$$

For the integral in x,

$$\begin{aligned}
\int (b_{\bar{j}}(x))^2 p_X(x) \,\mathrm{d}x &= \int \prod_{k\in[d_x]} h^2(mx_k - i_k + 1) p_X(x) \,\mathrm{d}x \\
&\geq \frac{1}{c} \int \prod_{k\in[d_x]} h^2(mx_k - i_k + 1) \,\mathrm{d}x \\
&= m^{-d_X} c^{-1} \int \prod_{k\in[d_x]} h^2(w_k) \,\mathrm{d}w \\
&= m^{-d_X} c^{-1}.
\end{aligned}$$

For the integral in u, we first note that

$$\begin{aligned}
\int e^{\mathrm{i}\langle u,y\rangle} \prod_{y\in[d_Y]} h_{i_k,r}(y_k) \,\mathrm{d}y &= r^{\frac{d_Y}{2}} \prod_{k\in[d_Y]} \int e^{\mathrm{i}u_k y_k} h(ry_k - i_k + 1) \,\mathrm{d}y_k \\
&= r^{\frac{d_Y}{2}} e^{\mathrm{i}\langle \frac{u}{r}, i - \vec{1}\rangle} \prod_{k\in[d_Y]} \varphi_h\big(\frac{u_k}{r}\big).
\end{aligned}$$

Then

$$\int \frac{|\mathcal{F}(a_i^-)(u)|^2}{|u|^\gamma}\,\mathrm{d}u = r^{-d_Y}\int \frac{|\prod_{k\in[d_Y]}\varphi_h(\frac{u_k}{r})|^2}{|u|^\gamma}\,\mathrm{d}u$$
$$= r^{-\gamma}\int \frac{|\prod_{k\in[d_Y]}\varphi_h(w_k)|^2}{|w|^\gamma}\,\mathrm{d}w.$$

The last term is again a constant. Thus, we obtain

$$\int\int \frac{|\varphi_x^\Delta(u)-\varphi_x^{\Delta'}(u)|^2}{|u|^\gamma}\,\mathrm{d}u\,p_X(x)\,\mathrm{d}x \gtrsim |S|\rho^2 r^{-\gamma}m^{-d_X} \gtrsim r^{d_Y-\gamma}\rho^2. \qquad \square$$

Proof of Lemma 6.10. By Székely (2003, Proposition 2) we have that in case $\mathbb{E}_\mathbb{P}[|Y|^\beta]<\infty$ and $\mathbb{E}_\mathbb{Q}[|Z|^\beta]<\infty$ for $0<\beta\le 2$,

$$\mathbb{E}_{Y\sim\mathbb{P},Z\sim\mathbb{Q}}[|Y-Z|^\beta] - \frac{1}{2}\mathbb{E}_{Y,Y'\overset{\text{iid}}{\sim}\mathbb{P}}[|Y-Y'|^\beta] - \frac{1}{2}\mathbb{E}_{Z,Z'\overset{\text{iid}}{\sim}\mathbb{Q}}[|Z-Z'|^\beta]$$
$$= \frac{\alpha 2^\beta\Gamma\left(\frac{d+\beta}{2}\right)}{4\pi^{d/2}\Gamma(1-\beta/2)}\int \frac{|\varphi_\mathbb{P}(u)-\varphi_\mathbb{Q}(u)|^2}{|u|^{d+\beta}}\,\mathrm{d}u.$$

We can separate the integral

$$\int \frac{|\varphi_\mathbb{P}(u)-\varphi_\mathbb{Q}(u)|^2}{|u|^\gamma}\,\mathrm{d}u = \int_{|u|\le 1}\frac{|\varphi_\mathbb{P}(u)-\varphi_\mathbb{Q}(u)|^2}{|u|^\gamma}\,\mathrm{d}u + \int_{|u|>1}\frac{|\varphi_\mathbb{P}(u)-\varphi_\mathbb{Q}(u)|^2}{|u|^\gamma}\,\mathrm{d}u.$$

For the first term we can use the condition $\gamma < d+\beta$ to see that

$$\int_{|u|\le 1}\frac{|\varphi_\mathbb{P}(u)-\varphi_\mathbb{Q}(u)|^2}{|u|^\gamma}\,\mathrm{d}u \le \int_{|u|\le 1}\frac{|\varphi_\mathbb{P}(u)-\varphi_\mathbb{Q}(u)|^2}{|u|^{d+\beta}}\,\mathrm{d}u.$$

For the second term, we have for every $\tau\in(0,2)$

$$\begin{aligned}
&\int_{|u|>1}\frac{|\varphi_\mathbb{P}(u)-\varphi_\mathbb{Q}(u)|^2}{|u|^\gamma}\,\mathrm{d}u\\
&\quad\le \int_{|u|>1}|\varphi_\mathbb{P}(u)-\varphi_\mathbb{Q}(u)|^2\,\mathrm{d}u\\
&\quad= \int_{|u|>1}|\mathbb{E}[e^{\mathrm{i}\langle u,X\rangle}-e^{\mathrm{i}\langle u,Y\rangle}]|^\tau|\varphi_\mathbb{P}(u)-\varphi_\mathbb{Q}(u)|^{2-\tau}\,\mathrm{d}u\\
&\quad= \int_{|u|>1}|\mathbb{E}[\cos(uX)-\cos(uY)+\mathrm{i}(\sin(uX)-\sin(uY))]|^\tau|\varphi_\mathbb{P}(u)-\varphi_\mathbb{Q}(u)|^{2-\tau}\,\mathrm{d}u\\
&\quad\le \int_{|u|>1}\mathbb{E}[|\cos(uX)-\cos(uY)|+|\sin(uX)-\sin(uY)|]^\tau|\varphi_\mathbb{P}(u)-\varphi_\mathbb{Q}(u)|^{2-\tau}\,\mathrm{d}u\\
&\quad\le 2^\tau\mathbb{E}[|X-Y|]^\tau\int_{|u|>1}|u|^\tau(|\varphi_\mathbb{P}(u)|+|\varphi_\mathbb{Q}(u)|)^{2-\tau}\,\mathrm{d}u.
\end{aligned}$$

By assumption, the integral is finite. Hence we conclude

$$\int \frac{|\varphi_{\mathbb{P}}(u)-\varphi_{\mathbb{Q}}(u)|^2}{|u|^\gamma}\,\mathrm{d}u \lesssim \mathbb{E}_{Y\sim\mathbb{P},Z\sim\mathbb{Q}}[|Y-Z|^\beta] - \frac{1}{2}\mathbb{E}_{Y,Y'\overset{\text{iid}}{\sim}\mathbb{P}}[|Y-Y'|^\beta] - \frac{1}{2}\mathbb{E}_{Z,Z'\overset{\text{iid}}{\sim}\mathbb{Q}}[|Z-Z'|^\beta] + \mathbb{E}[|X-Y|]^\tau. \qquad \square$$

Proof of Lemma 6.15. For the supremum norm bound, observe that for every $z \in \mathbb{R}^{d_Y}$ and $x \in \mathbb{R}^{d_X}$

$$\begin{aligned}
|\mathring{v}_t(z,x)| &= \Big|\sum_{i=1}^n v_t(z|Y_i)\frac{p_t(z|Y_i)K^{\mathrm{x}}_{h_x}(x-X_i)}{\sum_{i=1}^n p_t(z|Y_i)K^{\mathrm{x}}_{h_x}(x-X_i)}\Big| \\
&\le \sum_{i=1}^n |v_t(z|Y_i)|\frac{p_t(z|Y_i)K^{\mathrm{x}}_{h_x}(x-X_i)}{\sum_{i=1}^n p_t(z|Y_i)K^{\mathrm{x}}_{h_x}(x-X_i)} \\
&\le \max_{i\in\{1,\dots,n\}} |v_t(z|Y_i)| \\
&= \max_{i\in\{1,\dots,n\}} \frac{|Y_i-(1-\sigma_{\min})z|}{1-(1-\sigma_{\min})t} \\
&\le \max_{i\in\{1,\dots,n\}} \frac{|Y_i|+|z|}{1-(1-\sigma_{\min})t} \\
&\le \frac{\sqrt{d}(1+a)}{1-(1-\sigma_{\min})t}.
\end{aligned}$$

For the Lipschitz bound, note that

$$\begin{aligned}
\nabla_z \mathring{v}_t(z,x) &= \nabla_z \sum_{i=1}^n \Big(\frac{\frac{\partial\sigma_t}{\partial t}}{\sigma_t}z - \mu_t(Y_i) + \frac{\partial\mu_t}{\partial t}(Y_i)\Big)\frac{p_t(z|Y_i)K^{\mathrm{x}}_{h_x}(x-X_i)}{\sum_{i=1}^n p_t(z|Y_i)K^{\mathrm{x}}_{h_x}(x-X_i)} && (6.19)\\
&= \nabla_z \frac{-(1-\sigma_{\min})}{\sigma_t}z - \nabla_z \frac{-(1-\sigma_{\min})t}{\sigma_t}\frac{\sum_{i=1}^n Y_i p_t(z|Y_i)K^{\mathrm{x}}_{h_x}(x-X_i)}{\sum_{i=1}^n p_t(z|Y_i)K^{\mathrm{x}}_{h_x}(x-X_i)} && (6.20)\\
&\quad + \nabla_z \frac{\sum_{i=1}^n Y_i p_t(z|Y_i)K^{\mathrm{x}}_{h_x}(x-X_i)}{\sum_{i=1}^n p_t(z|Y_i)K^{\mathrm{x}}_{h_x}(x-X_i)}.
\end{aligned}$$

Now for the partial derivative of the ℓ-st coordinate function with respect to z_k, $\ell, k \in \{1,\dots,d\}$, we get

$$\begin{aligned}
&\frac{\partial}{\partial z_k}\frac{\sum_{i=1}^n Y_{i,l} p_t(z|Y_i)K^{\mathrm{x}}_{h_x}(x-X_i)}{\sum_{i=1}^n p_t(z|Y_i)K^{\mathrm{x}}_{h_x}(x-X_i)} \\
&= \frac{\big(\sum_{i=1}^n \big(-\frac{z_k-tY_{ik}}{\sigma_t^2}\big)Y_{i,\ell}p_t(z|Y_i)K^{\mathrm{x}}_{h_x}(x-X_i)\big)\big(\sum_{j=1}^n p_t(z|Y_j)K^{\mathrm{x}}_{h_x}(x-X_i)\big)}{\big(\sum_{j=1}^n p_t(z|Y_j)K^{\mathrm{x}}_{h_x}(x-X_i)\big)^2} \\
&\quad - \frac{\big(\sum_{i=1}^n Y_{i,\ell}p_t(z|Y_i)K^{\mathrm{x}}_{h_x}(x-X_i)\big)\big(\sum_{j=1}^n \big(-\frac{z_k-tY_{jk}}{\sigma_t^2}\big)p_t(z|Y_j)K^{\mathrm{x}}_{h_x}(x-X_i)\big)}{\big(\sum_{j=1}^n p_t(z|Y_j)K^{\mathrm{x}}_{h_x}(x-X_i)\big)^2} \\
&= \frac{t}{\sigma_t^2}\Big(\frac{\sum_{i=1}^n Y_{ik}Y_{i\ell}p_t(z|Y_i)K^{\mathrm{x}}_{h_x}(x-X_i)}{\sum_{j=1}^n p_t(z|Y_i)K^{\mathrm{x}}_{h_x}(x-X_i)} \\
&\qquad - \frac{\big(\sum_{i=1}^n Y_{ik}p_t(z|Y_i)K^{\mathrm{x}}_{h_x}(x-X_i)\big)\big(\sum_{i=1}^n Y_{i\ell}p_t(z|Y_i)K^{\mathrm{x}}_{h_x}(x-X_i)\big)}{\big(\sum_{j=1}^n p_t(z|Y_i)K^{\mathrm{x}}_{h_x}(x-X_i)\big)^2}\Big).
\end{aligned}$$

Since $\text{supp}(p^*) \subset [0,1]^{d_Y \times d_X}$, we can bound

$$\frac{\partial}{\partial z_k} \frac{\sum_{i=1}^n Y_{i,\ell} p_t(z|Y_i) K_{h_x}^{\mathrm{x}}(x - X_i)}{\sum_{j=1}^n p_t(z|Y_j) K_{h_x}^{\mathrm{x}}(x - X_i)} \leq \frac{2t}{\sigma_t^2}.$$

Using $t \in [0,1]$, $\sigma_{\min} \leq 1$, we get for (6.19)

$$|\nabla_z v(z,x)| \leq \frac{1}{\sigma_t} I_d + \frac{2}{\sigma_t^3} J_d,$$

where I_d denotes the $d \times d$ identity matrix, J_d denotes the $d \times d$ matrix consisting of ones and $\leq$ denotes entry wise inequality. Using the mean value theorem, we obtain for $z, w \in \mathbb{R}^d$ and $x \in \mathbb{R}^p$

$$|v(z,x) - v(w,x)| \leq \left\| \frac{1}{\sigma_t} I_d + \frac{2}{\sigma_t^3} J_d \right\| |z - w|.$$

As

$$\begin{aligned} \left| \frac{1}{\sigma_t} I_d + \frac{2}{\sigma_t^3} J_d \right|_2 &\leq \left\| \frac{1}{\sigma_t} I_d \right\| + \left\| \frac{2}{\sigma_t^3} J_d \right\| \\ &= \frac{1}{\sigma_t} + \frac{2d}{\sigma_t^3}, \end{aligned}$$

we get the desired bound. □

Chapter 7

Conclusion and Outlook

In this thesis, we studied different generative machine models in a distribution estimation setting. We started with GANs, showing the first rate of convergence that is also applicable in dimension reduction settings. As a byproduct, we showed that we can theoretically approximate a Lipschitz function using a Hölder network. Then we investigated the kernel density estimator in Wasserstein distance. For a suitable choice of kernels, we showed that this classical method achieves optimal rates. Our results specifically apply to the Gaussian kernel, which was not the case in previous results. Afterwards, we studied the recent model Flow Matching, where previous work is rather scarce. We connected the intrinsic smoothing of this model to the kernel density estimator. This connection allowed us to study the over-parameterized setting, where we showed that Flow Matching can achieve optimal rates. After a profound study of the Lipschitz constant of the underlying vector field for arbitrary variance functions, we also derived rates of convergence using smaller networks. Lastly, we investigated Flow Matching and a Nadaraya-Watson-type estimator in a conditional distribution estimation setting. We showed that the Nadaraya-Watson-type estimator achieves minimax optimal rates with respect to the Fourier score and transferred this result to Flow Matching in certain cases. Furthermore, an empirical study revealed the promising possibilities of Flow Matching in a forecasting setting.
There are several possible avenues for further research. These include specific questions about the models studied in this thesis, and overarching questions. The former will be presented, grouped by the model, followed by the latter.

Generative adversarial networks Our analysis demonstrates that GANs originally built on too sensitive distribution distances, such as the Jensen-Shannon distance, can be improved by a Lipschitz constraint in the discriminator class. This insight might also be applicable to other GANs, e.g. f-GANs of Nowozin et al. (2016), which rely on a divergence that cannot discriminate between different singular distributions and thus is not suitable for a dimension reduction setting. Overall, we conclude that the choice of the discriminator class is much more important for the data generation capabilities than the choice of the loss function, which is typically dictated by some distance. Moreover, our analysis of the discriminator approximation error is not limited to Vanilla GANs, but is also applicable to optimal transport based GANs as demonstrated for the Wasserstein-type GAN.
While our analysis was limited to feedforward ReLU networks, one advancement in neural

network research is the use of more sophisticated network architectures whose statistical analysis is not yet settled. In the context of Wasserstein GANs, see for example Radford et al. (2015).
Furthermore, the inclusion of a bound on the Lipschitz constant, and not only the Hölder constant, would enable a direct application of Theorem 3.10, thereby eliminating the need to include the parameter α and thus improving the rates. Additionally, it would be interesting whether there are conditions that allow for a faster rate of convergence for the Vanilla GAN, excluding scenarios as in Example 3.6.
The experiments also demonstrated that the GAN is capable of detecting data from a lower dimensional manifold if the latent space is of the same dimension as the ambient space. The proof of Theorem 3.9 is contingent upon the dimension of the latent space. If the dimension of the latent space is chosen to be too small, then $\inf_{G^* \in \mathrm{Lip}(M,\mathcal{Z})} \mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{G^*(Z)})$ will be large. If the dimension of the latent space is chosen too large, $\inf_{G^* \in \mathrm{Lip}(M,\mathcal{Z})} \mathsf{W}_1(\mathbb{P}^*, \mathbb{P}^{G^*(Z)})$ does not deteriorate, but the corresponding rate depends on the higher latent dimension. Therefore rates that are adaptive to the unknown intrinsic dimension, potentially benefiting from results like Berenfeld & Hoffmann (2021), would be interesting.

Flow Matching for unconditional distribution estimation While the networks used for Theorem 5.8 are very large and do not correspond to the networks used in practice, the rate in Theorem 5.21 is inferior. Although the results apply to different unknown distributions, a blending is desirable.
The inferior rate in Theorem 5.21 is partly due to the general approximation result used. While this leads to better comparability to other works, a more tailored approach could lead to optimal rates. In addition, the size of the network and the overall result of both settings could again benefit from the theoretical use of more sophisticated networks, like the U-Net (Ronneberger et al., 2015) construction in Lipman et al. (2023).
Section 5.5.1 paves the way for further research into broader classes of distributions whose vector fields have a bounded Lipschitz constant. The class of functions of form (5.20) served as a toy example and is an interesting starting point for generalization. A very natural extension is the extension to all distributions satisfying Assumption 5.16. Thus, characterizations of Assumption 5.16 would be of high interest. Conversely, the lower bound on the Lipschitz constant could be used to identify distributions that cannot be mimicked with a "good" rate of convergence and realistic networks. While Assumption 5.16 holds for a very broad range of variance functions, proofs based on concentration inequalities, such as Theorem 5.21, depend on the specific choice of the variance function. Given optimal pushforward mappings, Tsimpos et al. (2025) have investigated optimal noise schedules. Interestingly, their optimal schedule connects to the choice in Assumption 5.19. Exploring optimal choices of variance functions from a statistical perspective would thus be very interesting.
Another promising direction is the introduction and consequences of an artificial Lipschitz control of the vector field. In this setting, the network size in Theorem 5.8 could decline drastically. This influences the equivalence of gradients between (5.3) and (5.6) as well as their empirical counterparts. Controlling this effect could extend convergence results to larger classes of unknown

distributions.

Flow Matching for conditional distribution estimation As already indicated in Section 6.3.3, the use of smooth networks could extend our theoretical results for the Flow Matching estimator beyond the energy score and to higher dimensions.
Further, Gneiting & Raftery (2007) show that the energy score can be generalized to the kernel score, using positive definite kernels. Straightforward calculations reveal that for a fixed kernel, the corresponding divergence function recovers the squared maximum mean discrepancy (MMD) in form of Gretton et al. (2012, Lemma 6). Thus, moving from the Fourier score to the kernel score would further generalize the results obtained in this thesis.
In practice, guided diffusions (Dhariwal & Nichol, 2021; Ho & Salimans, 2021) and the adaptation to Flow Matching (Zheng et al., 2023) have recently gained attention. Guided generative modeling inter- and extrapolates between the conditional and the unconditional distribution to enhance the emphasis on the covariates. This leads to impressive results in conditional image generation. Our study of the conditional distribution estimator, which is a special instance of the guided model, could serve as a starting point.
Interestingly, Flow Matching behaves quite well in extrapolation too, as Figure 7.1 shows. The exact model configurations can be found in Appendix A. Extrapolation properties of other conditional generative models have been studied by Shen & Meinshausen (2025). It would be interesting to study the extrapolation properties of Flow Matching from a theoretical point of view.
On the simulation side, the next step is to study the conditional Flow Matching model in a comparable way on large scale datasets. Moving from a standard personal computer to a cluster computer would enable a reliable comparison to the results of Walz et al. (2024) and further state-of-the-art methods. Furthermore an interesting question is whether the conditional Flow Matching model can also work in a setting where the target space is high-dimensional.

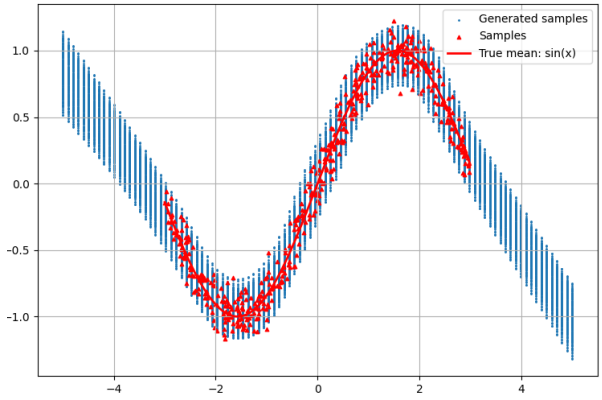


Figure 7.1.: Extrapolation capability of Flow Matching. Model trained on observations (red triangles) from $X \sim U[-3, 3]$, $Y \sim \mathcal{N}(\sin(X), 0.1^2)$, 200 latent samples are chosen once and then put through the model for different values of $x \in [-5, 5]$.

Overall conclusion and outlook An overall conclusion is that in order to profit from smoothness in data, a model that intrinsically smooths is beneficial. This way, even nonsmooth

networks can profit from larger smoothness in the unknown distribution. While in GANs, all attempts in the literature use either smooth networks (Puchkin et al., 2024; Stéphanovitch et al., 2024) or a smoothed version of the empirical measure (Liang, 2021), smoothing is natural to Flow Matching. This enabled the use of a different reference function, which leads to error bounds independent of complexity bounds on the network class.

Throughout this thesis, the objective of was to examine statistical perspectives, and thus, optimization problems were not addressed. In the proofs, we frequently employed global minimizers and maximizers. Since we often faced non-convex optimization problems, gradient based methods may suffer from a considerable optimization error, especially for high-dimensional parameter spaces. Incorporating this optimization error would be more consistent with real-world scenarios.

Additionally, all rates of convergence presented in this thesis are upper or lower bounds that quantify how close an estimator is to the true target distribution. Since the model is trained on observations, overfitting could cause it to mimic the empirical distribution rather than the true distribution. In statistical terms, this results in a convergence rate that deteriorates to that of the empirical distribution, at worst. While in a nonsmooth model, this is sometimes all that can be expected, the reproduction of the training data is undesirable in practice (Li et al., 2024). This has already sparked theoretical interest: For Wasserstein GANs, Vardanyan et al. (2024) are using a L^p type penalty to maximize the deviation form the empirical distribution. First of all, extensions to Flow Matching would be interesting. Secondly, penalties adaptive to human cognition would be very interesting to study. This could be essential when considering questions about intellectual property regarding outputs of generative models.

In this thesis, we have always studied one model in a given setting. In practice, concatenation of models, for example, latent diffusions (Rombach et al., 2022), which concatenate an autoencoder to a diffusion model, can reduce computational effort by enforcing dimension reduction. The extension of theoretical results to such concatenated models would therefore be of high interest. Such an analysis requires handling the dependencies induced by training with data that has been processed through another learned model and is thus an interesting but far from trivial possible avenue for further research.

Appendix A

Appendix

Model of Figure 1.1 The model uses the Vanilla GAN construction as introduced in Chapter 3. The unknown distribution are the handwritten 3s from the MNIST dataset Lecun et al. (1998). The total dataset consists of 60000 observations, 6131 of which are classified as an image of the number 3.
The network architecture of the generator class consists of two hidden LeakyReLU layers with a negative slope of 0.2 and a width of 256 in the first hidden layer and a width of 512 in the second hidden layer. After the last linear transformation, we apply the tanh activation function. The network architecture of the generator class consists of three hidden LeakyReLU layers with a negative slope of 0.2 and a width of $512, 512$ and 256. The sigmoid function is implemented in the loss function `BCEWithLogitsLoss` from the torch library directly.
We do 30 training iterations, in each training iteration we train the model with a batch size of 128 samples. For the optimization, we employ the Adam optimizer (Kingma & Ba, 2014) for both the discriminator and the generator function with learning rate $lr = 0.0002$ and $\beta_1 = 0.5$ and $\beta_2 = 0.999$. As latent distribution, we employ the standard normal distribution in $\mathbb{R}^{784}$ for the training of G_1 and in $\mathbb{R}^{25}$ for the training of G_2.

Model of Figure 1.3 The model uses the Flow Matching construction as introduced in Chapter 5. The unknown distribution is a Camelback distribution based on $\mathcal{N}(-1, 0.1^2)$ and $\mathcal{N}(1, 0.1^2)$ with a mixing probability of $\frac{1}{2}$. We draw 50 samples from this distribution. As latent distribution we employ $\mathcal{N}(0, 1)$. We set $\sigma_{\min} = 0.01$.
The network architecture consists of 3 hidden SeLU layers with width 64. We do 5000 training iterations and use the full 50 samples in each training iteration. For the optimization, we employ the Adam optimizer (Kingma & Ba, 2014) with the standard parameters $lr = 0.001$, $\beta_1 = 0.9$ and $\beta_2 = 0.999$).
The implementation includes parts of the code of Tong et al. (2024), which is available on github. Specifically, the network architecture and the optimization scheme was maintained on purpose and no further parameter tuning was conducted. We also use Poli et al. (2025) to solve the neural ODE. More precisely, we use the solver `dopri5`, the sensitivity `adjoint` and set $\text{atol} = \text{rtol} = 10^{-5}$.

Model of Figure 1.4 This model uses the conditional Flow Matching model as introduced in Chapter 6 with the objective function (6.3). The covariate kernel is a gaussian kernel with bandwidth $h_x = 0.2$. The latent distribution is the distribution of the Epanechnikov kernel with $\sigma_{\min} = 0.0001$.
The unknown conditional distribution is the distribution of $X \sim \mathcal{N}(\sin(X), 0.5^2)$ where $W \sim U[-3, 3]$. We draw 500 samples from the true distribution. The network architecture consists of 3 hidden SeLU layers with width 64, which again corresponds to the architecture of Tong et al. (2024). We do 10000 training iterations, in each training iteration we train the model with a batch size of 100 samples. For the optimization, we employ the Adam optimizer (Kingma & Ba, 2014) with the standard parameters $lr = 0.001$, $\beta_1 = 0.9$ and $\beta_2 = 0.999$). The implementation is also based on the code of Tong et al. (2024) and we used the same settings of the neural ODE solver as in the model of Figure 1.3.
For the image, we draw 200 latent samples once and then put them through the model for different values of x.

Model of Figure 7.1 The architecture of the network and the specifications of the Adam optimizer is exactly the same as in the model of Figure 1.4. The covariate kernel is a gaussian kernel with bandwidth $h_x = 0.1$. The latent distribution is the distribution of the Epanechnikov kernel with $\sigma_{\min} = 0.0001$.

Bibliography

Aggarwal, C. (2018). *Neural Networks and Deep Learning: A Textbook*. Springer Cham.

Albergo, M. S. & Vanden-Eijnden, E. (2023). Building normalizing flows with stochastic interpolants. In *International Conference on Learning Representations*.

Alet, F., Price, I., El-Kadi, A., Masters, D., Markou, S., Andersson, T. R., Stott, J., Lam, R., Willson, M., Sanchez-Gonzalez, A., et al. (2025). Skillful joint probabilistic weather forecasting from marginals. *arXiv preprint arXiv:2506.10772*.

Anderson, B. D. (1982). Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3), 313–326.

Anil, C., Lucas, J., & Grosse, R. (2019). Sorting out Lipschitz function approximation. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *PMLR* (pp. 291–301).

Arjovsky, M. & Bottou, L. (2017). Towards principled methods for training generative adversarial networks. In *International Conference on Learning Representations*.

Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, volume 70 of *PMLR* (pp. 214–223).

Arsenyan, V., Vardanyan, E., & Dalalyan, A. (2025). Assessing the quality of denoising diffusion models in Wasserstein distance: Noisy score and optimal bounds. *arXiv preprint arXiv:2506.09681*.

Asokan, S. & Seelamantula, C. S. (2023). Euler-Lagrange analysis of generative adversarial networks. *Journal of Machine Learning Research*, 24(126), 1–100.

Atanackovic, L., Zhang, X., Amos, B., Blanchette, M., Lee, L. J., Bengio, Y., Tong, A., & Neklyudov, K. (2025). Meta flow matching: Integrating vector fields on the Wasserstein manifold. In *International Conference on Learning Representations*.

Azangulov, I., Deligiannidis, G., & Rousseau, J. (2024). Convergence of diffusion models under the manifold hypothesis in high-dimensions. *arXiv preprint arXiv:2409.18804*.

Bakry, D. & Émery, M. (1985). Diffusions hypercontractives. *Séminaire de probabilités*, 19, 177–206.

Bakry, D., Gentil, I., & Ledoux, M. (2013). *Analysis and Geometry of Markov Diffusion Operators*. Grundlehren der mathematischen Wissenschaften. Cham, Switzerland: Springer International Publishing, 2014 edition.

Bayraktar, E. & Guo, G. (2021). Strong equivalence between metrics of Wasserstein type. *Electronic Communications in Probability*, 26, 1 – 13.

Bayraktar, E. & Guo, G. (2024). Errata to the paper strong equivalence between metrics of Wasserstein type. *arXiv preprint arXiv:1912.08247*.

Belomestny, D., Naumov, A., Puchkin, N., & Samsonov, S. (2023). Simultaneous approximation of a smooth function and its derivatives by deep neural networks with piecewise-polynomial activations. *Neural Networks*, 161, 242–253.

Benton, J., Deligiannidis, G., & Doucet, A. (2024). Error bounds for flow matching methods. *Transactions on Machine Learning Research*.

Berenfeld, C. & Hoffmann, M. (2021). Density estimation on an unknown submanifold. *Electronic Journal of Statistics*, 15(1), 2179 – 2223.

Berry, T. & Sauer, T. (2017). Density estimation on manifolds with boundary. *Computational Statistics & Data Analysis*, 107, 1–17.

Biau, G., Cadre, B., Sangnier, M., & Tanielian, U. (2020). Some theoretical properties of GANs. *The Annals of Statistics*, 48(3), 1539 – 1566.

Biau, G., Sangnier, M., & Tanielian, U. (2021). Some theoretical insights into Wasserstein GANs. *Journal of Machine Learning Research*, 22(119), 1–45.

Bieringer, S., Kasieczka, G., Kieseler, J., & Trabs, M. (2024). Classifier surrogates: sharing AI-based searches with the world. *The European Physical Journal C*, 84(9), 972.

Bobkov, S. G. (1999). Isoperimetric and analytic inequalities for log-concave probability measures. *The Annals of Probability*, 27(4), 1903 – 1921.

Boissard, E. & Gouic, T. L. (2014). On the mean speed of convergence of empirical and occupation measures in Wasserstein distance. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 50(2), 539 – 563.

Bonnotte, N. (2013). *Unidimensional and Evolution Methods for Optimal Transportation*. PhD thesis, Université Paris Sud - Paris XI ; Scuola normale superiore (Pise, Italie).

Bose, J., Akhound-Sadegh, T., Huguet, G., Fatras, K., Rector-Brooks, J., Liu, C.-H., Nica, A. C., Korablyov, M., Bronstein, M. M., & Tong, A. (2024). SE(3)-stochastic flow matching for protein backbone generation. In *International Conference on Learning Representations*.

Bott, A.-K. & Kohler, M. (2017). Nonparametric estimation of a conditional density. *Annals of the Institute of Statistical Mathematics*, 69(1), 189–214.

Boucheron, S., Lugosi, G., & Massart, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press.

Brascamp, H. J. & Lieb, E. H. (1976). On extensions of the Brunn-Minkowski and Prekopa-Leindler theorems, including inequalities for log concave functions, and with an application to the diffusion equation. *Journal of Functional Analysis*, 22(4), 366–389.

Brenner, S. C. & Scott, L. R. (2008). *The Mathematical Theory of Finite Element Methods*, volume 15 of *Texts in Applied Mathematics*. Springer.

Bruno, S., Zhang, Y., Lim, D., Akyildiz, O. D., & Sabanis, S. (2025). On diffusion-based generative models and their error bounds: The log-concave case with full convergence estimates. *Transactions on Machine Learning Research*.

Cao, H., Tan, C., Gao, Z., Xu, Y., Chen, G., Heng, P.-A., & Li, S. Z. (2024). A survey on generative diffusion models. *IEEE Transactions on Knowledge and Data Engineering*.

Chakraborty, S. & Bartlett, P. L. (2025). On the statistical properties of generative adversarial models for low intrinsic data dimension. *Journal of Machine Learning Research*, 26, 1–57.

Chen, H., Lee, H., & Lu, J. (2023a). Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions. In *International Conference on Machine Learning*, volume 202 of *PMLR* (pp. 4735–4763).

Chen, M., Huang, K., Zhao, T., & Wang, M. (2023b). Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data. In *International Conference on Machine Learning*, volume 202 of *PMLR* (pp. 4672–4712).

Chen, M., Liao, W., Zha, H., & Zhao, T. (2020). Distribution approximation and statistical estimation guarantees of generative adversarial networks. *arXiv preprint arXiv:2002.03938*.

Chen, R. T. Q. & Lipman, Y. (2024). Flow matching on general geometries. In *International Conference on Learning Representations*.

Chen, R. T. Q., Rubanova, Y., Bettencourt, J., & Duvenaud, D. K. (2018). Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, volume 31.

Chen, S., Chewi, S., Li, J., Li, Y., Salim, A., & Zhang, A. (2023c). Sampling is as easy as learning the score: Theory for diffusion models with minimal data assumptions. In *International Conference on Learning Representations*.

Chernodub, A. & Nowicki, D. (2016). Norm-preserving orthogonal permutation linear unit activation functions (OPLU). *arXiv preprint arXiv:1604.02313*.

Chernozhukov, V., Fernández-Val, I., & Galichon, A. (2010). Quantile and probability curves without crossing. *Econometrica*, 78(3), 1093–1125.

DeVore, R. A., Hanin, B., & Petrova, G. (2021). Neural network approximation. *Acta Numerica*, 30, 327 – 444.

Devroye, L. & Lugosi, G. (2012). *Combinatorial Methods in Density Estimation*. Springer Series in Statistics. Springer.

Dhariwal, P. & Nichol, A. (2021). Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34 (pp. 8780–8794).

Dinh, L., Krueger, D., & Bengio, Y. (2015). NICE: non-linear independent components estimation. In *International Conference on Learning Representations*.

Divol, V. (2022). Measure estimation on manifolds: an optimal transport approach. *Probability Theory and Related Fields*, 183(1), 581–647.

Dou, Z., Kotekal, S., Xu, Z., & Zhou, H. H. (2024). From optimal score matching to optimal sampling. *arXiv preprint arXiv:2409.07032*.

Dudley, R. M. (1969). The speed of mean Glivenko-Cantelli convergence. *The Annals of Mathematical Statistics*, 40(1), 40 – 50.

Dunn, I. & Koes, D. R. (2024). Mixed continuous and categorical flow matching for 3d de novo molecule generation. *arXiv preprint arXiv:2404.19739*.

Eckstein, S. (2020). Lipschitz neural networks are dense in the set of all Lipschitz functions. *arXiv preprint arXiv:2009.13881*.

Efromovich, S. (2007). Conditional density estimation in a regression setting. *The Annals of Statistics*, 35(6), 2504–2535.

Ehm, W. & Gneiting, T. (2009). Local proper scoring rules. *University of Washington, Department of Statistics, Technical Report no. 551*.

Elstrodt, J. (2018). *Maß- und Integrationstheorie*. Springer Berlin Heidelberg.

Endres, D. & Schindelin, J. (2003). A new metric for probability distributions. *IEEE Transactions on Information Theory*, 49, 1858 – 1860.

Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., Podell, D., Dockhorn, T., English, Z., & Rombach, R. (2024). Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *PMLR* (pp. 12606–12633).

Evans, L. C. (2010). *Partial Differential Equations*. Providence, R.I.: American Mathematical Society.

Farnia, F. & Tse, D. (2018). A convex duality framework for GANs. In *Advances in Neural Information Processing Systems*, volume 31.

Fedus, W., Rosca, M., Lakshminarayanan, B., Dai, A. M., Mohamed, S., & Goodfellow, I. (2017). Many paths to equilibrium: GANs do not need to decrease a divergence at every step. In *International Conference on Learning Representations*.

Fukumizu, K., Suzuki, T., Isobe, N., Oko, K., & Koyama, M. (2025). Flow matching achieves almost minimax optimal convergence. In *International Conference on Learning Representations*.

Gabcke, W. (1979). *Neue Herleitung und explizite Restabschätzung der Riemann-Siegel-Formel*. PhD thesis, University Goettingen Repository.

Gao, X., Nguyen, H. M., & Zhu, L. (2025). Wasserstein convergence guarantees for a general class of score-based generative models. *Journal of Machine Learning Research*, 26(43), 1–54.

Gao, X. & Zhu, L. (2025). Convergence analysis for general probability flow ODEs of diffusion models in Wasserstein distances. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *PMLR* (pp. 1009–1017).

Gao, Y., Huang, J., & Jiao, Y. (2024a). Gaussian interpolation flows. *Journal of Machine Learning Research*, 25(253), 1–52.

Gao, Y., Huang, J., Jiao, Y., & Zheng, S. (2024b). Convergence of continuous normalizing flows for learning probability distributions. *arXiv preprint arXiv:2404.00551*.

Gat, I., Remez, T., Shaul, N., Kreuk, F., Chen, R. T., Synnaeve, G., Adi, Y., & Lipman, Y. (2024). Discrete flow matching. In *Advances in Neural Information Processing Systems*, volume 37 (pp. 133345–133385).

Gerschgorin, S. (1931). Über die Abgrenzung der Eigenwerte einer Matrix. *Bulletin de l'Acad´emie des Sciences de l'URSS*, (pp. 749–754).

Gibbs, A. L. & Su, F. E. (2002). On choosing and bounding probability metrics. *International Statistical Review / Revue Internationale de Statistique*, 70(3), 419–435.

Gneiting, T., Balabdaoui, F., & Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 69(2), 243–268.

Gneiting, T. & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477), 359–378.

Goldfeld, Z., Greenewald, K., Niles-Weed, J., & Polyanskiy, Y. (2020). Convergence of smoothed empirical measures with applications to entropy estimation. *IEEE Transactions on Information Theory*, 66(7), 4368–4391.

Gong, C., Li, X., Liang, Y., Long, J., Shi, Z., Song, Z., & Tian, Y. (2025). Theoretical guarantees for high order trajectory refinement in generative flows. *arXiv preprint arXiv:2503.09069*.

Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 14(1), 107–114.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27.

Grathwohl, W., Chen, R. T. Q., Bettencourt, J., & Duvenaud, D. (2019). Scalable reversible generative models with free-form continuous dynamics. In *International Conference on Learning Representations*.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., & Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research*, 13(25), 723–773.

Gühring, I., Kutyniok, G., & Petersen, P. (2020). Error bounds for approximations with deep ReLU neural networks in $W^{s,p}$ norms. *Analysis and Applications*, 18(05), 803–859.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., & Courville, A. C. (2017). Improved training of Wasserstein GANs. In *Advances in Neural Information Processing Systems*, volume 30.

Guo, Y., Du, C., Ma, Z., Chen, X., & Yu, K. (2024). Voiceflow: Efficient text-to-speech with rectified flow matching. In *IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 11121–11125).

Györfi, L. (2002). *A Distribution-Free Theory of Nonparametric Regression*. Springer Series in Statistics. Springer.

Hall, P., Wolff, R. C., & Yao, Q. (1999). Methods for estimating a conditional distribution function. *Journal of the American Statistical Association*, 94(445), 154–163.

Harrison, D. & Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5(1), 81–102.

Henzi, A., Ziegel, J. F., & Gneiting, T. (2021). Isotonic distributional regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(5), 963–993.

Hersbach, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, 15(5), 559–570.

Ho, J. & Salimans, T. (2021). Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*.

Holley, R. & Stroock, D. (1987). Logarithmic Sobolev inequalities and stochastic Ising models. *Journal of Statistical Physics*, 46(5-6), 1159–1194.

Huang, J., Jiao, Y., Li, Z., Liu, S., Wang, Y., & Yang, Y. (2022). An error analysis of generative adversarial networks for learning distributions. *Journal of Machine Learning Research*, 23(116), 1–43.

Huster, T., Chiang, C.-Y. J., & Chadha, R. (2019). Limitations of the Lipschitz constant as a defense against adversarial examples. In *ECML PKDD 2018 Workshops, Proceedings 18*, volume 11329 of *Lecture Notes in Computer Science* (pp. 16–29).

Hyndman, R. J., Bashtannyk, D. M., & Grunwald, G. K. (1996). Estimating and visualizing conditional densities. *Journal of Computational and Graphical Statistics*, 5(4), 315.

Hyvärinen, A. (2005). Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(24), 695–709.

Kerrigan, G., Migliorini, G., & Smyth, P. (2024). Functional flow matching. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *PMLR* (pp. 3934–3942).

Khromov, G. & Singh, S. P. (2024). Some fundamental aspects about Lipschitz continuity of neural networks. In *International Conference on Learning Representations*.

Kingma, D. P. & Ba, J. (2014). Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

Klambauer, G., Unterthiner, T., Mayr, A., & Hochreiter, S. (2017). Self-normalizing neural networks. In *Advances in Neural Information Processing Systems*, volume 30.

Kobyzev, I., Prince, S. J., & Brubaker, M. A. (2020). Normalizing flows: An introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11), 3964–3979.

Kodali, N., Abernethy, J., Hays, J., & Kira, Z. (2017). On convergence and stability of GANs. *arXiv preprint arXiv:1705.07215*.

Koenker, R. & Bassett, G. (1978). Regression quantiles. *Econometrica*, 46(1), 33.

Kohler, M. (2015). Skript zur Vorlesung Kurvenschätzung.

Kohler, M. & Langer, S. (2021). On the rate of convergence of fully connected deep neural network regression estimates. *The Annals of Statistics*, 49(4), 2231 – 2249.

Kunkel, L. (2025). Distribution estimation via flow matching with Lipschitz guarantees. *arXiv preprint arXiv:2509.02337*.

Kunkel, L. & Trabs, M. (2025a). On the minimax optimality of flow matching through the connection to kernel density estimation. *arXiv preprint arXiv:2504.13336*.

Kunkel, L. & Trabs, M. (2025b). A Wasserstein perspective of vanilla GANs. *Neural Networks*, 181, 106770.

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W., Merose, A., Hoyer, S., Holland, G., Vinyals, O., Stott, J., Pritzel, A., Mohamed, S., & Battaglia, P. (2023). Learning skillful medium-range global weather forecasting. *Science*, 382(6677), 1416–1421.

Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278–2324.

Ledoux, M. (1999). Concentration of measure and logarithmic Sobolev inequalities. *Séminaire de probabilités*, 33, 120–216.

Ledoux, M. (2001). Logarithmic Sobolev inequalities for unbounded spin systems revisited. *Seminaire de probabilites (Strasbourg), tome 35*, (pp. 167–194).

Lee, J., Kwon, H. K., & Chae, M. (2025). Rates of convergence for nonparametric estimation of singular distributions using generative adversarial networks. *Journal of the Korean Statistical Society*, 54, 718–738.

Leonov, V. P. & Shiryaev, A. N. (1959). On a method of calculation of semi-invariants. *Theory of Probability & Its Applications*, 4(3), 319–329.

Li, M., Neykov, M., & Balakrishnan, S. (2022). Minimax optimal conditional density estimation under total variation smoothness. *Electronic Journal of Statistics*, 16(2), 3937–3972.

Li, S., Chen, S., & Li, Q. (2024). A good score does not lead to a good generative model. *arXiv preprint arXiv:2401.04856*.

Liang, T. (2017). How well can generative adversarial networks learn densities: A nonparametric view. *arXiv preprint arXiv:1712.08244*.

Liang, T. (2021). How well generative adversarial networks learn distributions. *Journal of Machine Learning Research*, 22(1), 10366–10406.

Lipman, Y., Chen, R. T. Q., Ben-Hamu, H., Nickel, M., & Le, M. (2023). Flow matching for generative modeling. In *International Conference on Learning Representations*.

Liu, X., Gong, C., & Liu, Q. (2023). Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*.

Lunardi, A. (2018). *Interpolation Theory*. Lecture Notes (Scuola Normale Superiore). Pisa, Italy: Scuola Normale Superiore, 3 edition.

Marzouk, Y., Ren, Z. R., Wang, S., & Zech, J. (2024). Distribution learning via neural differential equations: a nonparametric statistical perspective. *Journal of Machine Learning Research*, 25(232), 1–61.

Menz, G. (2014). A Brascamp-Lieb type covariance estimate. *Electronic Journal of Probability*, 19, 1–15.

Met Office (2010 - 2015). *Cartopy: a cartographic Python library with a Matplotlib interface*. Exeter, Devon.

Miyato, T., Kataoka, T., Koyama, M., & Yoshida, Y. (2018). Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*.

Müller, A. (1997). Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2), 429–443.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability & Its Applications*, 9(1), 141–142.

Nickel, S., Rebennack, S., Stein, O., & Waldmann, K.-H. (2022). *Operations Research*. Wiesbaden, Germany: Springer Gabler, 3 edition.

Niles-Weed, J. & Berthet, Q. (2022). Minimax estimation of smooth densities in Wasserstein distance. *The Annals of Statistics*, 50(3), 1519 – 1540.

Nowozin, S., Cseke, B., & Tomioka, R. (2016). f-GAN: Training generative neural samplers using variational divergence minimization. In *Advances in Neural Information Processing Systems*, volume 29.

Oko, K., Akiyama, S., & Suzuki, T. (2023). Diffusion models are minimax optimal distribution estimators. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *PMLR* (pp. 26517–26582).

Parzen, E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3), 1065 – 1076.

Petzka, H., Fischer, A., & Lukovnikov, D. (2018). On the regularization of Wasserstein GANs. In *International Conference on Learning Representations*.

Pic, R., Dombry, C., Naveau, P., & Taillardat, M. (2023). Distributional regression and its evaluation with the CRPS: Bounds and convergence of the minimax risk. *International Journal of Forecasting*, 39(4), 1564–1572.

Pick, L., Kufner, A., John, O., & Fucik, S. (2012). *Function Spaces, 1*. De Gruyter Series in Nonlinear Analysis & Applications. Berlin, Germany: De Gruyter, 2 edition.

Poli, M., Massaroli, S., Yamashita, A., Asama, H., Park, J., & Ermon, S. (2025). Torchdyn: Implicit models and neural numerical methods in pytorch.

Price, I., Sanchez-Gonzalez, A., Alet, F., Andersson, T. R., El-Kadi, A., Masters, D., Ewalds, T., Stott, J., Mohamed, S., Battaglia, P., Lam, R., & Willson, M. (2025). Probabilistic weather forecasting with machine learning. *Nature*, 637(8044), 84–90.

Puchkin, N., Samsonov, S., Belomestny, D., Moulines, E., & Naumov, A. (2024). Rates of convergence for density estimation with generative adversarial networks. *Journal of Machine Learning Research*, 25(29), 1–47.

Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. In *International Conference on Learning Representations*.

Rasp, S., Hoyer, S., Merose, A., Langmore, I., Battaglia, P., Russell, T., Sanchez-Gonzalez, A., Yang, V., Carver, R., Agrawal, S., Chantry, M., Ben Bouallegue, Z., Dueben, P., Bromberg, C., Sisk, J., Barrington, L., Bell, A., & Sha, F. (2024). Weatherbench 2: A benchmark for the next generation of data-driven global weather models. *Journal of Advances in Modeling Earth Systems*, 16(6).

Rasp, S. & Lerch, S. (2018). Neural networks for postprocessing ensemble weather forecasts. *Monthly Weather Review*, 146(11), 3885–3900.

Rezende, D. & Mohamed, S. (2015). Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *PMLR* (pp. 1530–1538).

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 10684–10695).

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer Assisted Intervention: 18th International Conference, Proceedings, Part III 18* (pp. 234–241).

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 27(3), 832 – 837.

Santambrogio, F. (2015). *Optimal Transport for Applied Mathematicians*. Progress in Nonlinear Differential Equations and Their Applications. Basel: Birkhauser Cham, 1 edition.

Schmidt-Hieber, J. (2020). Nonparametric regression using deep neural networks with ReLU activation function. *The Annals of Statistics*, 48(4), 1875 – 1897.

Schreuder, N. (2020). Bounding the expectation of the supremum of empirical processes indexed by Hölder classes. *Mathematical Methods of Statistics*, 29(1), 76–86.

Schreuder, N., Brunel, V.-E., & Dalalyan, A. (2021). Statistical guarantees for generative models without domination. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, volume 132 of *PMLR* (pp. 1051–1071).

Scott, D. W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley Series in Probability and Statistics. Wiley.

Shen, X. & Meinshausen, N. (2025). Engression: extrapolation through the lens of distributional regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(3), 653–677.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., & Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32st International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research* (pp. 2256–2265).

Solomon, J. (2018). Optimal transport on discrete domains. *AMS Short Course on Discrete Differential Geometry*.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., & Poole, B. (2021). Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*.

Stein, E. M. (1970). *Singular Integrals and Differentiability Properties of Functions*. Princeton University Press.

Stéphanovitch, A. (2024). Smooth transport map via diffusion process. *arXiv preprint arXiv:2411.10235*.

Stéphanovitch, A., Aamari, E., & Levrard, C. (2024). Wasserstein generative adversarial networks are minimax optimal distribution estimators. *The Annals of Statistics*, 52(5), 2167 – 2193.

Stéphanovitch, A., Aamari, E., & Levrard, C. (2025). Generalization bounds for score-based generative models: a synthetic proof. *arXiv preprint arXiv:2507.04794*.

Suh, N. & Cheng, G. (2024). A survey on statistical theory of deep learning: Approximation, training dynamics, and generative models. *Annual Review of Statistics and Its Application*, 12, 177–207.

Suzuki, T. (2019). Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: optimal rate and curse of dimensionality. In *International Conference on Learning Representations*.

Székely, G. J. (2003). E-statistics: The energy of statistical samples. *Bowling Green State University, Department of Mathematics and Statistics Technical Report*, 3(05), 1–18.

Székely, G. J., Rizzo, M. L., et al. (2005). Hierarchical clustering via joint between-within distances: Extending Ward's minimum variance method. *Journal of Classification*, 22(2), 151–184.

Tabak, E. G. & Turner, C. V. (2013). A family of nonparametric density estimation algorithms. *Communications on Pure and Applied Mathematics*, 66(2), 145–164.

Tabak, E. G. & Vanden-Eijnden, E. (2010). Density estimation by dual ascent of the log-likelihood. *Communications in Mathematical Sciences*, 8(1), 217 – 233.

Tang, R., Lin, L., & Yang, Y. (2025). Conditional diffusion models are minimax-optimal and manifold-adaptive for conditional distribution estimation. In *International Conference on Learning Representations*.

Tang, R. & Yang, Y. (2023). Minimax rate of distribution estimation on unknown submanifolds under adversarial losses. *The Annals of Statistics*, 51(3), 1282 – 1308.

Tang, R. & Yang, Y. (2024). Adaptivity of diffusion models to manifold structures. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *PMLR* (pp. 1648–1656).

Tang, W. & Zhao, H. (2024). Contractive diffusion probabilistic models. *arXiv preprint arXiv:2401.13115*.

Than, K. & Vu, N. (2021). Generalization of GANs and overparameterized models under Lipschitz continuity. *arXiv preprint arXiv:2104.02388*.

Tolstikhin, I., Bousquet, O., Gelly, S., & Schoelkopf, B. (2018). Wasserstein auto-encoders. In *International Conference on Learning Representations*.

Tong, A., Fatras, K., Malkin, N., Huguet, G., Zhang, Y., Rector-Brooks, J., Wolf, G., & Bengio, Y. (2024). Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*.

Tsanas, A. & Xifara, A. (2012). Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools. *Energy and Buildings*, 49, 560–567.

Tsimpos, P., Zhi, R., Zech, J., & Marzouk, Y. (2025). Optimal scheduling of dynamic transport. In *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *PMLR* (pp. 5441–5505).

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer, 1st edition.

Vardanyan, E., Hunanyan, S., Galstyan, T., Minasyan, A., & Dalalyan, A. S. (2024). Statistically optimal generative modeling with maximum deviation from the empirical distribution. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *PMLR* (pp. 49203–49225).

Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Villani, C. (2008). *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg.

Vincent, P. (2011). A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7), 1661–1674.

Walter, W. (1970). *Differential and Integral Inequalities*. Springer Berlin Heidelberg.

Walz, E.-M., Henzi, A., Ziegel, J., & Gneiting, T. (2024). Easy uncertainty quantification (EasyUQ): Generating predictive distributions from single-valued model output. *Siam Review*, 66(1), 91–122.

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, 26(4), 359–372.

Weed, J. (2018). Sharper rates for estimating differential entropy under Gaussian convolutions. *Massachusetts Institute of Technology (MIT), Technical report*.

Wei, X., Liu, Z., Wang, L., & Gong, B. (2018). Improving the improved training of Wasserstein GANs. In *International Conference on Learning Representations*.

Wu, H.-T. & Wu, N. (2022). Strong uniform consistency with rates for kernel density estimators with general kernels on manifolds. *Information and Inference: A Journal of the IMA*, 11(2), 781–799.

Yakovlev, K. & Puchkin, N. (2025). Generalization error bound for denoising score matching under relaxed manifold assumption. In *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *PMLR* (pp. 5824–5891).

Yang, X., Cheng, C., Yang, X., Liu, F., & Lin, G. (2025). Text-to-image rectified flow as plug-and-play priors. In *International Conference on Learning Representations*.

Yarotsky, D. (2017). Error bounds for approximations with deep ReLU networks. *Neural Networks*, 94, 103–114.

Yeh, I.-C. (1998). Modeling of strength of high-performance concrete using artificial neural networks. *Cement and Concrete Research*, 28, 1797–1808.

Yu, K. & Jones, M. C. (1998). Local linear quantile regression. *Journal of the American Statistical Association*, 93(441), 228–237.

Zhang, K., Yin, H., Liang, F., & Liu, J. (2024). Minimax optimality of score-based diffusion models: Beyond the density lower bound assumptions. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *PMLR* (pp. 60134–60178).

Zheng, Q., Le, M., Shaul, N., Lipman, Y., Grover, A., & Chen, R. T. (2023). Guided flows for generative modeling and decision making. *arXiv preprint arXiv:2311.13443*.

Zhou, Z., Liang, J., Song, Y., Yu, L., Wang, H., Zhang, W., Yu, Y., & Zhang, Z. (2019). Lipschitz generative adversarial nets. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *PMLR* (pp. 7584–7593).

Øksendal, B. (2003). *Stochastic Differential Equations*. Universitext. Berlin: Springer.