Electronic Journal of Statistics
ISSN: 1935-7524

Statistical analysis of Wasserstein GANs with applications to time series forecasting

Moritz Haas and Stefan Richter

Institute of Applied Mathematics
Heidelberg University
e-mail: Moritz.Haas@stud.uni-heidelberg.de; stefan.richter@iwr.uni-heidelberg.de

Abstract: We provide statistical theory for conditional and unconditional Wasserstein generative adversarial networks (WGANs) in the framework of dependent observations. We prove upper bounds for the excess Bayes risk of the WGAN estimators with respect to a modified Wasserstein-type distance. Furthermore, we formalize and derive statements on the weak convergence of the estimators and use them to develop confidence intervals for new observations. The theory is applied to the special case of high-dimensional time series forecasting. We analyze the behavior of the estimators in simulations based on synthetic data and investigate a real data example with temperature data. The dependency of the data is quantified with absolutely regular β-mixing coefficients.

MSC2020 subject classifications: Primary 62M45; secondary 62G05.
Keywords and phrases: Wasserstein GAN, excess Bayes risk, convergence rates, absolutely regular, high-dimensional, time series.

Contents

1 Introduction . . . 2
2 The Wasserstein GAN estimator . . . 5
2.1 Simplification of the WGAN objective and model assumption . . . 5
2.2 ReLU neural networks . . . 7
2.3 The unconditional WGAN estimator . . . 7
2.4 The conditional WGAN estimator . . . 8
3 Theoretical results for the unconditional WGAN . . . 10
3.1 Properties of the modified Wasserstein distance . . . 10
3.2 Excess Bayes risk . . . 11
3.3 Asymptotic confidence intervals . . . 14
4 Theoretical results for the conditional WGAN . . . 14
4.1 Results for the modified conditional Wasserstein distance . . . 14
4.2 Excess Bayes risk . . . 15
4.3 Asymptotic confidence intervals . . . 16
5 High-dimensional time series forecasting . . . 16
6 Simulation studies . . . 18
6.1 Synthetic data . . . 19
6.2 Real data application . . . 20

7 Conclusion 23
References 24
A Proofs of Section 3 26
B Error Decomposition 27
B.1 Unconditional WGAN: Basic inequality 27
B.2 Approximation error 29
B.3 Estimation error 30
B.4 Adaptation to the conditional case 34
C Entropy bound and large deviation bounds for absolutely regular sequences 37
C.1 Entropy bounds 37
C.2 Large deviation bounds 38

1. Introduction

Generative adversarial networks (GANs) are a class of algorithms in machine learning for learning distributions in high-dimensional feature spaces. After the training process, they are able to generate new random fake observations mimicking the observations already seen. In applications, they have shown to provide surprisingly good results in image and speech generation as well as in inpainting tasks.

The training process is designed as follows: Iteratively, two neural networks compete against each other. While the first network (the *generator*) produces new random observations which imitate the original training samples, the second network (the *critic* or *discriminator*) judges their quality and tries to discriminate between true and generated observations. The assessment is performed with a specific distance of probability distributions. The original GAN was defined with a Kullback-Leibler-type divergence (so called Vanilla GANs, cf. [15]). In practical applications, GANs using the Wasserstein distance (so called WGANs, cf. [4], [16]) have become popular due to their training stability and the high quality of the generated observations. In contrast to other divergence measures, such as the Kullback-Leibler divergence or the total variation divergence, the Wasserstein distance metrizes weak convergence, which makes it sensible to differences of distributions on lower-dimensional submanifolds and with disjoint supports (cf. [4]). This property stabilizes the training procedure of WGANs remarkably.

Let $d \in \mathbb{N}$ be the dimension of the feature space. WGANs learn a structured probability distribution $\mathbb{P}^X$ from potentially high-dimensional training samples $X_i \in \mathbb{R}^d$, $i \in \{1, ..., n\}$. To do so, a latent space $\mathbb{R}^{d_Z}$ with dimension $d_Z \in \mathbb{N}$ and latent random variables $Z_1, ..., Z_n \in \mathbb{R}^{d_Z}$ with a given "base distribution" $\mathbb{P}^Z$ are introduced. Then one tries to minimize the Wasserstein distance of the empirical measure of the training samples,

$$\hat{\mathbb{P}}_n^X := \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$$

(here, δ_{X_i} denotes the point measure on X_i) and the empirical measure of modified latent variables,

$$\hat{\mathbb{P}}_n^{g(Z)} := \frac{1}{n}\sum_{j=1}^{n} \delta_{g(Z_j)},$$

with respect to the *generator* $g : \mathbb{R}^{d_Z} \to \mathbb{R}^{d_X}$. After the learning process, an estimator $\hat{g}$ of g can produce new observations $\hat{g}(Z)$ which approximately follow $\mathbb{P}^X$ by sampling from the latent space $Z \sim \mathbb{P}^Z$.

The approach was generalized to conditional distributions $\mathbb{P}^{X|Y}$ in [21]. Let $d_Y \in \mathbb{N}$ be the dimension of the conditional feature space. If samples $(X_i, Y_i) \in \mathbb{R}^{d+d_Y}$, $i \in \{1, ..., n\}$ are observed, then the conditional WGAN approximately minimizes the Wasserstein distance between the empirical measure

$$\hat{\mathbb{P}}_n^{X,Y} := \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i,Y_i}$$

and

$$\hat{\mathbb{P}}_n^{g(Z,Y),Y} := \frac{1}{n}\sum_{i=1}^{n} \delta_{g(Z_i,Y_i),Y_i},$$

where here the *conditional generator* is a function $g : \mathbb{R}^{d_Z+d_Y} \to \mathbb{R}^d$ which also incorporates the values of Y during evaluation. In the same manner as before, approximate observations from $\mathbb{P}^{X|Y}$ can be obtained after the learning process (i.e. when an estimate $\hat{g}$ of g is available) by $\hat{g}(Z, Y)$ with samples $Z \sim \mathbb{P}^Z$ and given observations Y. In practice, conditional GANs (cGANs) introduce the information $Y = y$ to the generator in various stages of the architecture. cGANs are very popular for generating images given certain labels such as age, gender or glasses [2] and in image-to-image translation tasks (cf. [17, 20]), e.g. colorizing images or reconstructing higher resolution.

In both situations (unconditional and conditional), WGANs provide an approximation of the law of $\mathbb{P}^X$ or $\mathbb{P}^{X|Y}$ via $\hat{g}(Z)$ or $\hat{g}(Y, Z)$ and offer a powerful tool to obtain new samples even if the training data is high-dimensional. The reason is that under appropriate restrictions on the structure of g and the dimension d_Z of the latent variables, the data $g(Z)$ lies in a low-dimensional submanifold of $\mathbb{R}^d$.

The purpose of this paper is to provide a theoretical framework for conditional and unconditional WGANs and to prove convergence rates of the excess Bayes risk (with respect to a modified Wasserstein distance) in the context of time series X_i, $i = 1, ..., n$. We formalize in which sense the learned generator function can be used to provide asymptotic confidence sets for X. As an application, we will investigate conditional WGANs to provide confidence intervals for observations of high-dimensional time series. The use of WGANs and our corresponding theory is not limited to this example: For instance, one could think of new smoothed Bootstrap techniques.

Recent results from [6] and [7] already provided theoretical results for the excess Bayes risk of GANs and WGANs in the case of i.i.d. observations X_i.

They used network classes fixed in n for both discriminators and generators and therefore could not derive convergence rates for the whole excess Bayes risk. Furthermore, the minimized objective could not be used to derive (asymptotic) distributional properties of their corresponding estimators $\hat{g}$. With our results, we extend the theory of these publications in several ways:

1. We derive explicit statistical properties like characterization of weak convergence for the modified Wasserstein distance used in WGANs
2. We investigate the conditional WGAN, which is an important generalization for standard statistical applications as forecasting.
3. We allow the generator to be in a Hölder class and explicitly discuss upper bounds for the approximation error of $\hat{g}$. This yields explicit upper bounds on the whole excess Bayes risk and allows a discussion of the impact of structural assumptions on g and how the curse of dimension can be avoided in practice.
4. We allow the observations X_i, $i = 1, ..., n$ and $Y_i, i = 1, ..., n$ to be dependent.

From a technical point of view, we measure dependence with absolutely regular β-mixing coefficients. We use empirical process theory from [12] and [10] as well as refined Talagrand's inequalities from [8] to provide large deviation inequalities of the excess Bayes risk.

The paper is organized as follows. In Section 2, we introduce the Wasserstein metric as well as the conditional and unconditional WGAN estimator based on neural networks. Section 3 covers the unconditional case. We firstly relate the introduced modified Wasserstein distance (a network based integral probability metric, cf. [22]) to the 1-Wasserstein distance. Then we provide convergence rates for the excess Bayes risk with respect to this distance under structural assumptions on the underlying data generating process and the neural networks used for estimation. In Section 4 we establish equivalent results for the conditional case. In Section 5, we transfer our results from Section 4 to high-dimensional time series forecasting. In Section 6, we provide simulation results of the conditional WGAN algorithm both for simulated data and real-world temperature data. A short conclusion is drawn in Section 7. All proofs are deferred without further reference to the Appendix.

We now summarize some notation used in this paper. $(\Omega, \mathcal{A}, \mathbb{P})$ will denote a Borel probability space. For some vector $x \in \mathbb{R}^d$, let $|x| = (\sum_{j=1}^d |x_j|^2)^{1/2}$ denote its Euclidean norm, $|x|_\infty = \max_i |x_i|$ and $|x|_0 = \sum_i \mathbf{1}(x_i \neq 0)$. For measurable functions $f : T \to \mathbb{R}$, we write $\|f\|_\infty := \sup_{x \in T} |f(x)|$ whenever there is no ambiguity on the domain $T \subset \mathbb{R}^r$. For $f : T \to \mathbb{R}^{\tilde{d}}$, we further denote $\|f\|_\infty := \max_{j=1,...,\tilde{d}} \|f_j\|_\infty$ and the Lipschitz norm $\|f\|_L := \sup_{x \neq y} \frac{|f(x)-f(y)|}{|x-y|}$ we denote the Lipschitz norm w.r.t. the Euclidean norm $|\cdot|$. Finally, we use the following multi-index calculus: For differentiable functions $f : T \to \mathbb{R}$ and $\alpha = (\alpha_1, \ldots, \alpha_r) \in \mathbb{N}_0^r$, let $|\alpha| = \sum_{i=1}^r \alpha_i$ and let $\partial^\alpha f = \partial_1^{\alpha_1} \ldots \partial_r^{\alpha_r} f$ denote the r-th partial derivative. Finally, for real-valued random variables W and $q > 0$ we write $\|W\|_q := \mathbb{E}[|W|^q]^{1/q}$.

2. The Wasserstein GAN estimator

Throughout the paper, we consider X_i, $i = 1, ..., n$ to be a strictly stationary process taking values in $[0,1]^d$, where $d \in \mathbb{N}$ is an arbitrary dimension. Here, we restrict ourselves to the unit cube $[0,1]^d$ for convenience, our theory could easily be generalized to arbitrary compact Euclidean spaces. We start with an introduction of the typical approximations used in the Wasserstein GAN approach as well as the optimization problem we aim to discuss. Based on this notation, we then give a statistical formulation of the conditional Wasserstein GAN. To keep our assumptions concise, we define the set of functions $f : T \subset \mathbb{R}^r \to \mathbb{R}$ with Hölder coefficient $\beta \geq 1$ via

$$\begin{aligned} & C^\beta(T,K) \\ := \quad & \big\{ f : T \to \mathbb{R} \big| \sum_{\alpha:\, 0 \leq |\alpha| < \beta} \|\partial^\alpha f\|_\infty + \sum_{\alpha:\, |\alpha| = \beta - 1} \sup_{x \neq y} \frac{|\partial^\alpha f(x) - \partial^\alpha f(y)|}{|x-y|_\infty} \leq K \big\}, \end{aligned}$$

where $K > 0$.

2.1. Simplification of the WGAN objective and model assumption

Let $d_Z \in \mathbb{N}$ and $\mathbb{P}^Z$ a known distribution on $[0,1]^{d_Z}$. Let $\mathcal{G} \subset \{g : \mathbb{R}^{d_Z} \to \mathbb{R}^d \text{ measurable}\}$ be a space of generators. In this paper, we will assume that $\mathcal{G}$ consists of smooth functions with a special structure (this is made precise in Definition 2.1). The objective of a WGAN is to approximate the underlying probability distribution $\mathbb{P}^X$ by minimizing the 1-Wasserstein distance to $\mathbb{P}^{g(Z)}$ with respect to $g \in \mathcal{G}$. By the Kantorovich-Rubinstein duality (cf. [30]), the 1-Wasserstein distance of two probability distributions $\mathbb{P}_1, \mathbb{P}_2$ on $\mathbb{R}^d$ can be written as

$$W_1(\mathbb{P}_1, \mathbb{P}_2) = \sup_{f:\mathbb{R}^d \to \mathbb{R}, \|f\|_L \leq 1} \Big\{ \int f \, \mathrm{d}\mathbb{P}_1 - \int f \, \mathrm{d}\mathbb{P}_2 \Big\}.$$

The original objective of the WGAN is to find a suitable $g \in \mathcal{G}$ which minimizes

$$W_1(\mathbb{P}^X, \mathbb{P}^{g(Z)}) = \sup_{f:\mathbb{R}^d \to \mathbb{R}, \|f\|_L \leq 1} \big\{ \mathbb{E} f(X) - \mathbb{E} f(g(Z)) \big\}. \tag{2.1}$$

In practical applications, the set of critics $f : \mathbb{R}^d \to \mathbb{R}$ is replaced by a certain set of neural networks $\mathcal{R}_D \subset \{f : \mathbb{R}^d \to \mathbb{R}\}$. This replacement makes the objective more tractable and also allows the graphical interpretation of competing networks. Similarly, for estimation, the set of possible generators $\mathcal{G}$ is replaced by a set of neural networks $\mathcal{R}_G \subset \{f : \mathbb{R}^{d_Z} \to \mathbb{R}^d\}$.

In the following, we therefore replace the theoretical objective (2.1) by

$$W_{1,n}(g) := \sup_{f \in \mathcal{R}_D, \|f\|_L \leq 1} \big\{ \mathbb{E} f(X) - \mathbb{E} f(g(Z)) \big\}. \tag{2.2}$$

Note that $W_{1,n}$ may depend on n through the class $\mathcal{R}_D$. Due to the restriction on $f \in \mathcal{R}_D$, one can not expect that $W_{1,n}(g) = W_1(\mathbb{P}^X, \mathbb{P}^{g(Z)})$. This raises the

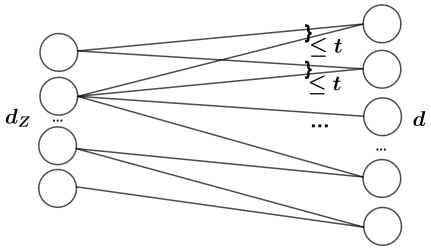


Fig 1: Structure of the generating functions g which are used to model the distribution of X with $g(Z)$.

question which properties $W_{1,n}(g)$ should preserve (and thus, how large and of which form $\mathcal{R}_D$ should be) to make it a meaningful distance of the measures $\mathbb{P}^X$ and $\mathbb{P}^{g(Z)}$. Our basic aim is to preserve the property that $W_{1,n}$ characterizes weak convergence in the sense that for any sequence $g_n \in \mathcal{R}_G$,

$$W_{1,n}(g_n) \to 0 \quad \text{implies} \quad g_n(Z) \xrightarrow{d} X.$$

The precise conditions on $\mathcal{R}_D$ and its connections to the space $\mathcal{R}_G$ of generators are given in Lemma 3.2 in Section 3.

In [24, Theorem 3] it was shown that the 1-Wasserstein distance between two measures in $\mathbb{R}^d$ can, in general, not be estimated with a better rate than $(n \log(n))^{-1/d}$. Even though $W_{1,n}(g)$ is smaller than $W_1(\mathbb{P}^X, \mathbb{P}^{g(Z)})$, one needs specific structural assumptions on the underlying distribution and the class of estimators to overcome the curse of dimension. Since we aim to approximate $\mathbb{P}^X$ by $\mathbb{P}^{g(Z)}$, it is clear that we expect some kind of "sparsity" of X if $d_Z < d$. If we expect $\mathbb{P}^X$ to lie (approximately) in a d_g-dimensional submanifold ($d_g \in \{1, ..., d\}$) of $\mathbb{R}^d$, it seems reasonable to choose $d_Z = d_g$. To allow for a more flexible choice of d_Z, we introduce the following function class.

Definition 2.1 (Generator function class)**.** Let $\mathcal{G}(d_Z, d_g, \beta, K)$ be the set of all measurable functions $g : \mathbb{R}^{d_Z} \to \mathbb{R}^d$ such that any component only depends on d_g arguments and lies in $C^\beta([0,1]^{d_Z}, K)$.

In principle, one can allow for much more complicated structures of the generator functions. Here we reduce ourselves to the above formulation for simplicity. An example of a more general class which has auto-encoder structure is introduced for the conditional case (cf. Definition 2.2) and could also be chosen here.

2.2. ReLU neural networks

We now specify the classes $\mathcal{R}_D$ and $\mathcal{R}_G$ of neural networks in more detail. To do so, we use a theoretical formulation from [28]. For $x \in \mathbb{R}$, let $\sigma(x) = \max\{x, 0\}$ denote the rectified linear unit (ReLU) activation function. For $v, x \in R^p$, $p \in \mathbb{N}$, define

$$\sigma_v(x) = \sigma(x - v),$$

where $\sigma(\cdot)$ is applied component-wise to the vector $x - v$. Let $L \in \mathbb{N}$ and $p = (p_0, ..., p_{L+1}) \in \mathbb{N}^{L+2}$. A neural network with network architecture $(L, \mathbf{p})$ is a function

$$h : \mathbb{R}^{p_0} \to \mathbb{R}^{p_{L+1}}, \quad h(x) = W^{(L)}\sigma_{v^{(L)}}W^{(L-1)} \dots W^{(1)}\sigma_{v^{(1)}}W^{(0)}x, \tag{2.3}$$

where $W^{(l)} \in \mathbb{R}^{p_l \times p_{l+1}}$, $l = 0, ..., L$ are the weight matrices and $v^{(l)} \in \mathbb{R}^{p_l}$, $l = 1, ..., L$ are the bias vectors associated to the network. Consequently, let

$$\mathcal{R}(L, \mathbf{p}) = \Big\{h : \mathbb{R}^{p_0} \to \mathbb{R}^{p_{L+1}} \mid g \text{ is of the form (2.3)}\Big\},$$

be the class of deep ReLU networks with network architecture $(L, \mathbf{p})$. Training of neural networks typically is done with a stochastic gradient descent method and a random initialization of the weight matrices. It is observed in practice that only few parameters of the resulting networks are "active" in the sense that they contribute to the final function value. Accordingly, we introduce the set of sparse networks bounded by $F > 0$ by

$$\begin{aligned}\mathcal{R}(L, \mathbf{p}, s, F) := \Big\{h \in \mathcal{R}(L, \mathbf{p}) \;\Big|\; &\max_{j=0,...,L} \|W_j\|_\infty \vee |v_j|_\infty \le 1, \\ &\sum_{j=0}^{L} \|W_j\|_0 + |v_j|_0 \le s \text{ and } \| \, |h|_\infty \|_{L^\infty([0,1]^{p_0})} \le F\Big\}.\end{aligned}$$

Since F is fixed, we will abbreviate $\mathcal{R}(L, \mathbf{p}, s) = \mathcal{R}(L, \mathbf{p}, s, F)$ in the following.

2.3. The unconditional WGAN estimator

We use the theoretical formulation in (2.2) but replace the expectation $\mathbb{E}f(X)$ by its empirical counterpart $\frac{1}{n}\sum_{i=1}^n f(X_i)$. Furthermore, $\mathbb{E}f(g(Z))$ is approximated by $\frac{1}{n\mathcal{E}}\sum_{j=1}^{n\mathcal{E}} f(g(Z_{ij}))$, where $Z_{i,j}$, $j = 1, ..., \mathcal{E}$, $i = 1, ..., n$ are i.i.d. realizations of $\mathbb{P}^Z$ (independent of X_i, $i = 1, ..., n$) and $\mathcal{E} \in \mathbb{N}$ is some parameter. We then obtain

$$\hat{g}_n := \operatorname*{arg\,min}_{g \in \mathcal{R}(L_g, \mathbf{p}_g, s_g)} \hat{W}_{1,n}(g) \tag{2.4}$$

with

$$\begin{aligned}\hat{W}_{1,n}(g) &:= \sup_{f \in \mathcal{R}(L_f, \mathbf{p}_f, s_f), \|f\|_L \le 1} \big\{\hat{\mathbb{P}}_n^X f - \hat{\mathbb{P}}_{n\mathcal{E}}^Z (f \circ g)\big\} \\ &= \sup_{f \in \mathcal{R}(L_f, \mathbf{p}_f, s_f), \|f\|_L \le 1} \frac{1}{n}\sum_{i=1}^n \big\{f(X_i) - \frac{1}{\mathcal{E}}\sum_{j=1}^{\mathcal{E}} f(g(Z_{i,j}))\big\}\end{aligned}$$

where $L_g, L_f \in \mathbb{N}$ are the layer sizes, $\mathbf{p}_g, \mathbf{p}_f$ the corresponding width vectors and s_g, s_f the sparsity parameters.

Note that an optimizer $\hat{g}_n$ exists (cf. [7]), since $\hat{W}_{1,n}$ is Lipschitz continuous with respect to $g \in \mathcal{R}(L_g, \mathbf{p}_g, s_g)$ and $g \in \mathcal{R}(L_g, \mathbf{p}_g, s_g)$ is Lipschitz continuous with respect to its parameters $W^{(l)}$, $v^{(l)}$, which in turn are defined on a compact set. Similarly there exists an optimal critic network in $\mathcal{R}(L_f, \mathbf{p}_f, s_f)$ for any function $g : [0,1]^{d_Z} \to [0,1]^d$.

The parameter $\mathcal{E}$ is motivated by algorithms which are used in practice to find approximations of (2.4), cf. Section 6. These algorithms work iteratively. Each iteration which uses all training data is called epoch. The random variables Z_i, $i = 1, ..., n$ are not sampled one time at the beginning but new samples are generated in each training epoch. Although in practice the generator only has access to a part of the data X_i, $i = 1, ..., n$ in each epoch, $\mathcal{E}$ roughly grows proportional to the number of epochs. Thus one can imitate the knowledge coming from the additional realizations of $\mathbb{P}^Z$ and study its implications.

Appropriate choices for these parameters to guarantee upper bounds for the excess Bayes risk are formulated in Section 3.

2.4. The conditional WGAN estimator

Conditional GANs (cGANs), firstly introduced in [21], extend the task of learning to sample from a given distribution $\mathbb{P}^X$ to learning to sample from conditional distributions $P^{X|Y=y}$, $y \in [0,1]^{d_Y}$, where Y is another random variable in a space $[0,1]^{d_Y}$ encoding some information about X. The idea is simply to learn the joint distribution $\mathbb{P}^{X,Y}$ with the same Wasserstein objective (2.1) with a generator that has access to Y. The formal legitimation is that if we find a function $g_c^* : [0,1]^{d_Z+d_Y} \to \mathbb{R}^d$ with $\mathbb{P}^{X,Y} = \mathbb{P}^{g_c^*(Z,Y),Y}$, then the independency of Y, Z implies $\mathbb{P}^{X|Y=y} = \mathbb{P}^{g_c^*(Z,y)}$.

We introduce a more complex class $\mathcal{G}^c(d_Z, d_Y, D, d_g, \beta, K)$ of generators with encoder-decoder structure, cf. Figure 2, so that generators can depend on all components of the conditional information given by Y, even for large dimensions d_Y.

Definition 2.2 (Encoder-decoder structure)**.** Let $\mathcal{G}^c(d_Z, d_Y, D, d_g, \beta, K)$ be the set of all measurable functions $g : \mathbb{R}^{d_Z+d_Y} \to \mathbb{R}^d$ which have the form

$$g = g_{dec} \circ g_{enc,1} \circ g_{enc,0},$$

where

- $g_{enc,0} : \mathbb{R}^{d_Z+d_Y} \to \mathbb{R}^D$ such that any component only depends on d_g arguments and lies in $C^\beta([0,1]^{d_g}, K)$,
- $g_{enc,1} : \mathbb{R}^D \to \mathbb{R}^{d_g}$ such that any component lies in $C^{\tilde{\beta}}([0,1]^D, K)$, with some $\tilde{\beta} \geq \frac{D}{d_g}\beta$.
- $g_{dec} : \mathbb{R}^{d_g} \to \mathbb{R}^d$ such that any component lies in $C^\beta([0,1]^{d_g}, K)$.

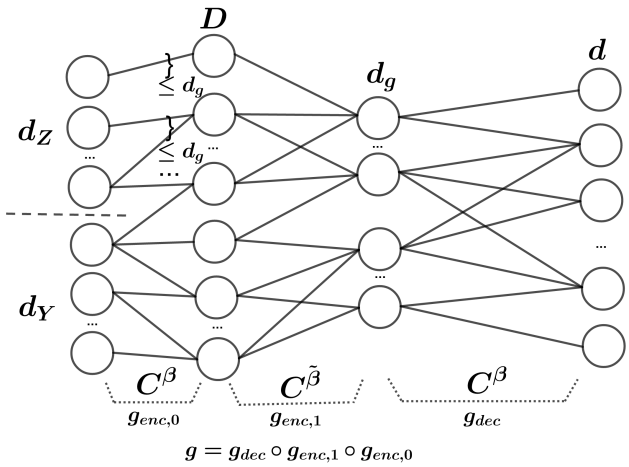


Fig 2: Structure of the generating functions g which are used to model the distribution of (X, Y) with $(g(Z, Y), Y)$.

The original objective now is to find a minimizer of

$$W^c_{1,n}(g) := \sup_{f\in\mathcal{R}(L_f,\mathbf{p}_f,s_f),\|f\|_L\leq 1} \left\{\mathbb{E}f(X,Y) - \mathbb{E}f\big(g(Z,Y),\, Y\big)\right\}.$$

If (X_i, Y_i), $i = 1, \ldots, n$ are strictly stationary realizations of $\mathbb{P}^{(X,Y)}$ and Z_i, $i = 1, ..., n$ are i.i.d. realizations of $\mathbb{P}^Z$ independent of (X_i, Y_i), $i = 1, ..., n$, we define

$$\hat{g}^c_n := \operatorname*{arg\,min}_{g\in\mathcal{R}(L_g,\mathbf{p}_g,s_g)} \hat{W}^c_{1,n}(g) \tag{2.5}$$

with

$$\hat{W}^c_{1,n}(g) \quad := \quad \sup_{f\in\mathcal{R}(L_f,\mathbf{p}_f,s_f),\|f\|_L\leq 1} \frac{1}{n}\sum_{i=1}^{n} \left\{f(X_i,Y_i) - f(g(Z_i,Y_i),Y_i)\right\}$$

where $L_g, L_f \in \mathbb{N}$ are the layer sizes, $\mathbf{p}_g, \mathbf{p}_f$ the corresponding width vectors and s_g, s_f the sparsity parameters. Appropriate choices for these parameters to guarantee upper bounds for the excess Bayes risk are formulated in Section 4. Note that in contrast to the unconditional WGAN, we do not implement the additional realizations of $\mathbb{P}^Z$ which may occur in practical algorithms. The reason is that for the conditional WGAN, the observations Y_i in the second summand in $\hat{W}^c_{1,n}(g)$ restrict the use of additional knowledge from Z without rather technical assumptions on the structure of g.

3. Theoretical results for the unconditional WGAN

The first part of this section is devoted to the properties of the modified distances $W_{1,n}(g)$. We show connections between $W_{1,n}(g)$ and distances which do not depend on n and prove that under certain assumptions on the set of networks $\mathcal{R}_D(L_f, \mathbf{p}_f, s_f)$, $W_{1,n}(g)$ characterizes weak convergence.

In the second part, we provide upper bounds and convergence rates for the excess Bayes risk

$$R_n(g) := W_{1,n}(g) - \inf_{g \in \mathcal{G}(d_Z, d_g, \beta, K)} W_{1,n}(g) \tag{3.1}$$

for the unconditional WGAN estimator $\hat{g}_n$ under assumptions on the network structure. In the third part, we summarize the results to provide asymptotic confidence intervals.

3.1. Properties of the modified Wasserstein distance

We first investigate the connection of $W_{1,n}(g)$ to

$$W_1^\gamma(g) := \sup_{f \in C^\gamma([0,1]^d, K), \|f\|_L \leq 1} \big\{ \mathbb{E} f(X) - \mathbb{E} f(g(Z)) \big\}.$$

In opposite to $W_{1,n}$, the quantity W_1^γ does not depend on n and therefore can be seen as a more "stable" distance measure for $\mathbb{P}^X$ towards $\mathbb{P}^{g(Z)}$. Note that

$$W_1^\gamma(g) \quad \leq \quad W_{1,n}(g) + 2 \sup_{f \in C^\gamma([0,1]^d, K)} \inf_{\tilde{f} \in \mathcal{R}(L_f, \mathbf{p}_f, s_f)} \|f - \tilde{f}\|_\infty.$$

Using approximation results for neural networks from [28] (cf. Theorem B.2 in the Appendix), one obtains the following result.

Lemma 3.1 (Lower bound on $W_{1,n}$)**.** *Let $a_n = n^{-\frac{2\gamma}{2\gamma+d}}$, and suppose that*

- $F \geq 1$,
- $L_f \geq \log_2(n) \log_2(4d \vee 4\gamma)$,
- $\min_{i=1,\dots,L} p_{f,i} \gtrsim n a_n$
- $s_f \gtrsim \log(n) n a_n$,

where the constants in the asymptotic expression above depend on γ, d. Then there exists some constants $C > 0, K \in (0,1)$ only depending on γ, d, F such that

$$\sup_{f \in C^\gamma([0,1]^d, K), \|f\|_L \leq 1} \inf_{\tilde{f} \in \mathcal{R}_D(L_f, \mathbf{p}_f, s_f)} \|f - \tilde{f}\|_\infty \leq C a_n^{1/2}.$$

Especially, for any measurable $g : \mathbb{R}^{d_Z} \to \mathbb{R}^d$,

$$W_1^\gamma(g) \leq W_{1,n}(g) + C a_n^{1/2}. \tag{3.2}$$

The lemma shows that the convergence rate of $W_{1,n}(\hat{g}_n)$ transfers to $W_1^\gamma(\hat{g}_n)$ as long as $a_n^{1/2} \leq W_{1,n}(\hat{g}_n)$. In fact, this imposes a *lower bound* on the Hölder exponent γ of functions considered with W_1^γ.

In the case that $\mathbb{P}^X = \mathbb{P}^{g^*(Z)}$ with some $g^* \in \mathcal{G}(d_Z, d_g, \beta, K)$, the results for the excess Bayes risk (3.1) presented in the following Section 3.2 can be used to derive weak convergence. The reasoning is as follows: If $R_n(\hat{g}_n) \to 0$, then $\mathbb{E}W_{1,n}(\hat{g}_n) \to 0$. Then the following lemma can be used.

Lemma 3.2 (Characterization of weak convergence)**.** *Suppose that* $\mathbb{P}^X = \mathbb{P}^{g^*(Z)}$ *for some* $g^* \in \mathcal{G}(d_Z, d_g, \beta, K)$ *and let the assumptions of Lemma 3.1 hold with some* $\gamma \geq 1$*. Let* $(\hat{g}_n)_{n\in\mathbb{N}}$ *be a sequence of random variables with* $\mathbb{E}W_{1,n}(\hat{g}_n) \to 0$*. Then*

$$\hat{g}_n(Z) \xrightarrow{d} g^*(Z) = X.$$

The lemma basically follows from Lemma 3.1 and the fact that $C^\gamma([0,1]^d)$ forms a convergence-determining class.

Remark 3.3. Lemma 3.2 implies weak convergence of $\mathbb{P}^{\hat{g}_n(Z)}$ towards $\mathbb{P}^X$, but it does not give any information about the speed of convergence. Nevertheless it seems reasonable that the speed depends on the upper bound in (3.2) which is given by the two summands $W_{1,n}(g)$ and $a_n^{1/2}$. Therefore, one should choose $\gamma \geq 1$ large enough such that $a_n^{1/2} \lesssim W_{1,n}(g)$. This is done in Remark 3.5 below. On the other hand, for larger γ, $W_1^\gamma(g)$ gives less information about the distance between $\mathbb{P}^{g(Z)}$ and $\mathbb{P}^X$. It is an open question how an optimal balance of γ should be chosen. A more detailed analysis of the approximation quality of the set of critic networks $\mathcal{R}_D(L_f, \mathbf{p}_f, s_f)$ could yield more insight. However, this would need sharp upper bounds on the Lipschitz constants of $\mathcal{R}(L_f, \mathbf{p}_f, s_f)$ and is a pure approximation problem, which is out of the scope of this paper.

3.2. Excess Bayes risk

To state the theoretical results on the excess Bayes risk $R_n(\hat{g}_n)$ in (3.1), we have to quantify the dependence structure of X_i, $i = 1, \ldots, n$ and Y_i, $i = 1, ..., n$. Basically, observations obtained at time steps which are far away from each other have to be "asymptotically independent". There exists a large variety of weak and strong mixing conditions. We refer to [9] for a detailed summary of conditions and basic properties. Here, we use absolutely regular β-mixing due to the well-established empirical process theory (cf. [12] and [10]).

The β-mixing coefficient between two σ-algebras $\mathcal{U}, \mathcal{V} \subseteq \mathcal{A}$ is defined by

$$\beta(\mathcal{U}, \mathcal{V}) := \frac{1}{2} \sup \sum_{(i,j)\in I\times J} \big|\mathbb{P}(U_i \cap V_j) - \mathbb{P}(U_i)\mathbb{P}(V_j)\big|,$$

where the supremum is taken over all finite partitions (A_i) and (B_j) $\mathcal{U}$- and $\mathcal{V}$-measurable respectively. For a time series X_i, $i = 1, ..., n$, one defines

$$\beta_X(0) = 1, \quad \beta_X(n) := \beta\big(\sigma(X_i;\, i \leq 0), \sigma(X_i;\, i \geq n)\big), \quad n \in \mathbb{N}.$$

Prominent examples of absolutely regular sequences are GARCH and ARMA as well as linear processes (cf. [14, 9, 11]).

We now present the theoretical result for the excess Bayes risk for the unconditional WGAN estimator $\hat{g}_n$.

Theorem 3.4. *Let* $\phi_n = (n\mathcal{E})^{-\frac{2\beta}{2\beta+d_g}}$. *Suppose that* $F \geq K \vee 1$, *and*

(i) $\log_2(n\mathcal{E})\log_2(4d_g \vee 4\beta) \leq L_g \lesssim \log(n\mathcal{E})$,
(ii) $min_{i=1,\dots,L_g} p_{g,i} \gtrsim n\mathcal{E}\phi_n$,
(iii) $s_g \asymp n\mathcal{E}\phi_n \log(n\mathcal{E})$
(iv) $L_f \leq L_g$, $s_f \leq s_g$.

Suppose that there exist constants $\kappa > 1, \alpha > 1$ *such that for all* $k \in \mathbb{N}$, $\beta_X(k) \leq \kappa \cdot k^{-\alpha}$. *Then*

$$\mathbb{E}R_n(\hat{g}_n) \lesssim \Big(\frac{s_f L_f \log(s_f L_f)}{n}\Big)^{1/2} + \phi_n^{1/2}\log(n\mathcal{E})^{3/2}, \tag{3.3}$$

and with probability at least $1 - 4n^{-1} - 2(\frac{\log(n)}{n})^{\frac{\alpha-1}{2}}$,

$$R_n(\hat{g}_n) \lesssim \Big(\frac{s_f L_f \log(s_f L_f)}{n}\Big)^{1/2} + \phi_n^{1/2}\log(n\mathcal{E})^{3/2} + \Big(\frac{\log(n)}{n}\Big)^{1/2},$$

where the bounding constants may depend on characteristics of X_1, κ, α *and* d, d_Z, d_g, β, K, F.

Remark on dependency. Let us discuss this result in more detail. First note that we basically only need that $\beta_X(k)$ is summable. The specific polynomial rate of the decay only enters the large deviation result, and is negligible if $\alpha \geq 3$. The statistical bounds are obtained by using empirical process theory for absolutely regular β-mixing sequences and refined versions of Talagrand's inequality for independent variables from [18].

Remark on critic networks. The rate in (3.3) decomposes into two terms, where the first term is determined by properties of the critic networks and the second term stems from generator networks. Regarding the first term, the only condition on the critic functions is given in (iv) which asks that the critic networks allow for less non-zero parameters and have less layers than the generator networks. The lack of conditions is clear since there is no a priori approximation task the critic networks have to fulfill in $W_{1,n}$. However, there is some interest in allowing for a large critic function class $\mathcal{R}(L_f, \mathbf{p}_f, s_f)$ due to the results from Lemma 3.1. This is discussed in more detail in Remark 3.5.

Remark on conditions. Assumptions (i)-(iii) stated in Theorem 3.4 are conditions on the network structure of the generator networks which are allowed (and needed) to grow with n. Basically, the lower bounds on $L_g, \mathbf{p}_g, s_g$ are used to bound the approximation error of finding an element $\tilde{g} \in \mathcal{R}_G(L_g, \mathbf{p}_g, s_g)$ which approximates $g \in \mathcal{G}(d_Z, d_g, \beta, K)$ well, while the upper bounds control the estimation error.

(i),(iii) ask the generator networks to have approximately $\log_2(n\mathcal{E})$ layers and allow for approximately

$$n\mathcal{E}\phi_n \log(n\mathcal{E}) = (n\mathcal{E})^{\frac{d_g}{2\beta+d_g}} \log(n\mathcal{E})$$

non-zero parameters (that is, entries in weight matrices and bias vectors). (ii) asks the layers to have a certain minimal width. Letting the minimal width of *all* layers grow polynomially in n seems rather unusual from a practical point of view. This is only due to the approximation technique adopted from [28] and can be improved.

Remark on convergence rate - dimensionality. In [19, Theorem 1] (cf. also [23]) it was shown that β-Hölder smooth densities in the space $\mathbb{R}^d$ can be estimated with a rate not faster than $n^{-\frac{\beta+1}{2\beta+d}}$ with respect to the Wasserstein distance. Note the additional $\beta+1$ in the nominator instead of β as it is the case, for instance, in standard nonparametric density estimation. Due to the additional structural assumptions, out method yields (up to a log factor) a convergence rate $\phi_n = n^{-\frac{\beta}{2\beta+d_g}}$ with respect to the modified Wasserstein-distance $W_{1,n}(\hat{g}_n)$. It does not depend on the underlying dimensionality d_Z of the generation space nor the dimension d of the observation space but only on the reduced dimension $d_g \leq d_Z$. Even if d_Z is chosen large (as it may occur in practice), the generator network estimator $\hat{g}_n$ can adapt to the unknown number d_g of relevant arguments without suffering from a curse of dimension.

Remark on convergence rate - generator size. In practice, along with each sampled batch of data one batch of generated data Z_{i1}, $i = 1, ..., n$ is produced, so that during the first epoch of training it holds that $\mathcal{E} = 1$. In subsequent epochs, the data set of fixed size n is reused, while the number of generated samples keeps growing. If we assume that the generator networks approximate the empirical optimizers at each step, the variable $\mathcal{E}$ introduced in the estimator $\hat{g}_n$ can be roughly seen as the number of training epochs and indicates that more and more realizations of $\mathbb{P}^Z$ are available to train $\hat{g}_n$. In principle, $\mathcal{E}$ can be chosen arbitrarily large, therefore one can use arbitrarily large generator architectures as long as one generates enough samples during training. Then the performance saturates due to the limited data samples n and the corresponding discriminator architecture (cf. (3.3)), but not due to the generator capacity. However, note that one cannot directly take $\mathcal{E}$ as the number of epochs. The reason is that in each epoch, the generator only sees a part of the data X_i, $i = 1, \ldots, n$ (see Table 1).

Remark 3.5 (Selection of critic and generator)**.** If there exists $g^* \in \mathcal{G}(d_Z, d_g, \beta, K)$ with $\mathbb{P}^{g^*(Z)} = \mathbb{P}^X$, then the results of Theorem 3.4, (3.3) and Lemma 3.1, (3.2) can be combined. In this case, one could ask for a suitable choice of $\gamma \geq 1$ such that the rates coincide, that is, $a_n = \phi_n$. This then also leads to more precise conditions on the discriminator network through Lemma 3.1. For simplicity, choose $\mathcal{E} = 1$. We see that equality is obtained with

$$\frac{\beta}{2\beta + d_g} = \frac{\gamma}{2\gamma + d},$$

which is fulfilled for $\gamma = \beta \frac{d}{d_g}$, and leads to

$$\mathbb{E}W_1^\gamma(\hat{g}_n) \lesssim \phi_n^{1/2} \log(n)^{3/2}.$$

3.3. Asymptotic confidence intervals

Based on the weak convergence, one can provide asymptotic confidence sets for X to a given level α. For simplicity, suppose that X is one-dimensional. For $N \in \mathbb{N}$, let $Z_j^*, j = 1, ..., N$ be i.i.d. samples of $\mathbb{P}^Z$, independent of X_i, Z_{ij} used to calculate the WGAN estimator $\hat{g}_n$ from (2.4). Define the empirical distribution function

$$\hat{F}_{N,n}(x) := \frac{1}{N} \sum_{j=1}^{N} \mathbb{1}_{\{\hat{g}_n(Z_j^*) \le x\}}$$

and let F_X denote the distribution function of X. Then the following result holds.

Lemma 3.6. *Suppose that $\mathbb{P}^X = \mathbb{P}^{g^*(Z)}$ for some $g^* \in \mathcal{G}(d_Z, d_g, \beta, K)$ and that $\mathbb{P}^X$ is continuous. Let the assumptions of Lemma 3.1 with some $\gamma \ge 1$ and Theorem 3.4 hold. Then for any $\rho > 0$,*

$$\limsup_{n\to\infty} \limsup_{N\to\infty} \mathbb{P}(|\hat{F}_{N,n}(X) - F_X(X)| \ge \rho) = 0.$$

By the probability integral transform, this shows that $\hat{F}_{N,n}(X)$ converges in probability to a uniform distribution on $[0, 1]$. For fixed $\alpha \in (0, 1)$, this justifies that the interval

$$I_{n,N} := \left\{ x \in \mathbb{R} : \hat{F}_{N,n}(x) \in \left(\frac{\alpha}{2}, 1 - \frac{\alpha}{2} \right] \right\} \tag{3.4}$$

which is built from the empirical $\frac{\alpha}{2}$ and $(1 - \frac{\alpha}{2})$ quantile curves of $\hat{g}_n(Z_j^*)$, $j = 1, ..., N$ is an asymptotic $(1 - \alpha)$-confidence set for X since

$$\mathbb{P}(X \in I_{n,N}) = \mathbb{P}\left(\hat{F}_{N,n}(X) \in \left(\frac{\alpha}{2}, 1 - \frac{\alpha}{2} \right] \right) \approx \mathbb{P}\left(F_X(X) \in \left(\frac{\alpha}{2}, 1 - \frac{\alpha}{2} \right] \right) = 1 - \alpha.$$

4. Theoretical results for the conditional WGAN

4.1. Results for the modified conditional Wasserstein distance

We now provide similar results as given in Lemma 3.1 and Lemma 3.2 for the conditional WGAN formulation.

In analogy, firstly define,

$$W_1^{c,\gamma}(g) := \sup_{f \in C^\gamma([0,1]^{d+d_Y}, K), \|f\|_L \le 1} \{ \mathbb{E} f(X, Y) - \mathbb{E} f(g(Z, Y), Y) \}.$$

Since we use the same approximation result [28], the connection of $W_{1,n}^c(g)$ to $W_1^{c,\gamma}(g)$ is essentially the same as in the unconditional case and we omit a proof.

Lemma 4.1. *Let $a_n = n^{-\frac{2\gamma}{2\gamma+d+d_Y}}$, and suppose that*

- $F \geq 1$,
- $L_f \geq \log_2(n)\log_2(4(d+d_Y) \vee 4\gamma)$,
- $\min_{i=1,\dots,L} p_{f,i} \gtrsim na_n$
- $s_f \gtrsim \log(n) n a_n$,

where the constants in the asymptotic expression above depend on γ, d, d_Y. *Then there exists some constants* $C > 0, K \in (0,1)$ *only depending on* γ, d, d_Y, F *such that*

$$\sup_{f \in C^\gamma([0,1]^{d+d_Y}, K), \|f\|_L \leq 1} \inf_{\tilde{f} \in \mathcal{R}_D(L_f, \mathbf{p}_f, s_f)} \|f - \tilde{f}\|_\infty \leq C a_n^{1/2}.$$

Especially, for any measurable $g : \mathbb{R}^{d_Z} \to \mathbb{R}^d$,

$$W_1^{c,\gamma}(g) \leq W_{1,n}^c(g) + C a_n^{1/2}. \tag{4.1}$$

In the case that $\mathbb{P}^{X,Y} = \mathbb{P}^{g^*(Z,Y),Y}$ with some $g^* \in \mathcal{G}^c(d_Z, d_Y, D, d_g, \beta, K)$, the results for the excess Bayes risk (4.3) presented in the following Section 4.2 can be used to derive weak convergence with the help of the following lemma.

Lemma 4.2. *Suppose that* $\mathbb{P}^{X,Y} = \mathbb{P}^{g^*(Z,Y),Y}$ *for some* $g^* \in \mathcal{G}^c(d_Z, d_Y, D, d_g, \beta, K)$ *and let the assumptions of Lemma 3.1 hold with some* $\gamma \geq 1$. *Let* $(\hat{g}_n^c)_{n\in\mathbb{N}}$ *be a sequence of random variables with* $\mathbb{E}W_{1,n}^c(\hat{g}_n^c) \to 0$. *Then*

$$\hat{g}_n^c(Z,Y) \xrightarrow{d} g^*(Z,Y) = X. \tag{4.2}$$

The lemma basically follows from Lemma 4.1 and the fact that $C^\gamma([0,1]^{d+d_Y})$ forms a convergence-determining class. Remark 3.3 applies here as well. Moreover, from (4.2) one directly obtains the convergence of $\mathbb{P}^{\hat{g}_n^c(Z,y)}$ towards the conditional distribution $\mathbb{P}^{X|Y=y}$.

4.2. Excess Bayes risk

We now provide a result for the excess Bayes risk of the conditional WGAN,

$$R_n^c(g) := W_{1,n}^c(g) - \inf_{g \in \mathcal{G}(d_Z, d_Y, D, d_g, \beta, K)} W_{1,n}^c(g). \tag{4.3}$$

Theorem 4.3. *Let* $\phi_n = n^{-\frac{2\beta}{2\beta+d_g}}$ *and* $\tilde{\beta} \geq \frac{D}{d_g}\beta$. *Suppose that* $F \geq K \vee 1$, *and*

(i) $\log_2(n)\big(2\log_2(4d_g \vee 4\beta) + \log_2(4D \vee 4\tilde{\beta})\big) \leq L_g \lesssim \log(n)$,
(ii) $min_{i=1,\dots,L_g} p_{g,i} \gtrsim n\phi_n$,
(iii) $s_g \asymp n\phi_n \log(n)$
(iv) $L_f \leq L_g$, $s_f \leq s_g$.

Suppose that there exist constants $\kappa > 1, \alpha > 1$ *such that for all* $k \in \mathbb{N}$, $\beta_{X,Y}(k) \leq \kappa \cdot k^{-\alpha}$. *Then*

$$\mathbb{E}R_n^c(\hat{g}_n^c) \lesssim \Big(\frac{s_f L_f \log(s_f L_f)}{n}\Big)^{1/2} + \phi_n^{1/2}\log(n)^{3/2}, \tag{4.4}$$

and with probability at least $1 - 4n^{-1} - 2(\frac{\log(n)}{n})^{\frac{\alpha-1}{2}}$,

$$R_n^c(\hat{g}_n^c) \lesssim \Big(\frac{s_f L_f \log(s_f L_f)}{n}\Big)^{1/2} + \phi_n^{1/2}\log(n)^{3/2} + \Big(\frac{\log(n)}{n}\Big)^{1/2},$$

where the bounding constants may depend on characteristics of (X_1, Y_1) *and* $\kappa, \alpha, d, d_Z, d_Y, D, d_g, \beta, \tilde{\beta}, K, F$.

All remarks for Theorem 3.4 apply here as well.

4.3. Asymptotic confidence intervals

For simplicity, suppose that X is one-dimensional. For $N \in \mathbb{N}$, let $Z_j^*, j = 1, ..., N$ be i.i.d. samples of $\mathbb{P}^Z$, independent of X_i, Y_i, Z_i used to calculate the WGAN estimator $\hat{g}_n^c$ from (2.5). Define the empirical distribution function

$$\hat{F}_{N,n}^c(x|y) := \frac{1}{N}\sum_{j=1}^{N} \mathbb{1}_{\{\hat{g}_n^c(Z_j^*, y) \le x\}}$$

and let $F_X(x|y) = \mathbb{P}(X \le x|Y = y)$ denote the distribution function of X conditional on $Y = y$. The proof of the following result is similar to Lemma 4.2 and therefore omitted.

Lemma 4.4. *Suppose that* $\mathbb{P}^{(X,Y)} = \mathbb{P}^{(g^*(Z,Y),Y)}$ *for some* $g^* \in \mathcal{G}(d_Z, d_Y, d_g, D, \beta, K)$ *and that* $F_X(x|y)$ *is continuous for* $\mathbb{P}^Y$*-a.e.* y. *Let the assumptions of Lemma 4.1 with some* $\gamma \ge 1$ *and Theorem 4.3 hold. Then for any* $\rho > 0$ *and for* $\mathbb{P}^Y$*-a.e.* y,

$$\limsup_{n\to\infty}\limsup_{N\to\infty} \mathbb{P}\big(\big|\hat{F}_{N,n}(X|y) - F_X(X|y)\big| \ge \rho \big| Y = y\big) = 0.$$

As before in the case of the unconditional WGAN, we can now construct an asymptotic $(1-\alpha)$ confidence set for X conditional on $Y = y$. For fixed $\alpha \in (0,1)$, let

$$I_{n,N}(y) := \Big\{x \in \mathbb{R} : \hat{F}_{N,n}(x|y) \in \Big(\frac{\alpha}{2}, 1 - \frac{\alpha}{2}\Big]\Big\}. \tag{4.5}$$

Then for large n, N, one has

$$\begin{aligned}\mathbb{P}(X \in I_{n,N}(y)|Y = y) &= \mathbb{P}(\hat{F}_{N,n}(X|y) \in \Big(\frac{\alpha}{2}, 1 - \frac{\alpha}{2}\Big]\,|Y = y) \\ &\approx \mathbb{P}(F_X(X|y) \in \Big(\frac{\alpha}{2}, 1 - \frac{\alpha}{2}\Big]\,|Y = y) = 1 - \alpha.\end{aligned}$$

5. High-dimensional time series forecasting

Earlier practical approaches of [26, 29] have shown that conditional WGANs can be used to determine distributional forecasts of time series. In this section we use our results to provide asymptotic confidence intervals.

Suppose that we have given a time series $A_i \in \mathbb{R}^p$, $i = -r+1, ..., n$ with *continuous* distribution which is absolutely regular β-mixing with coefficients $\beta_A(k)$, $k \geq 0$. We are interested in forecasting a statistic

$$T(A_i), \qquad T : \mathbb{R}^d \to \mathbb{R} \text{ continuous},$$

conditional on the finite past

$$\mathbb{A}_{i-1} := (A_{i-1}, ..., A_{i-r}) \in \mathbb{R}^{pr},$$

where $r \in \mathbb{N}$ denotes the number of lags considered. Let us furthermore assume that there exists some $\beta \geq 1, K > 0$, $d_g \in \mathbb{N}$ and $g^{*c} \in \mathcal{G}^c(d_Z, pr, d_g, \beta, K)$ (cf. Definition 2.2) such that

$$\mathbb{P}^{T(A_r), \mathbb{A}_{r-1}} = \mathbb{P}^{g^*(Z, \mathbb{A}_{r-1})},$$

that is, the distribution of $T(A_r)$ is obtained from $A_{r-1}, ..., A_1$ and some random noise Z. Let $\hat{g}_n^c$ denote the conditional WGAN estimator from (2.5), that is,

$$\hat{g}_n^c = \underset{g \in \mathcal{R}(L_g, \mathbf{p}_g, s_g)}{\arg\min} \hat{W}_{1,n}^c(g)$$

with

$$\hat{W}_{1,n}^c(g) := \sup_{f \in \mathcal{R}(L_f, \mathbf{p}_f, s_f), \|f\|_L \leq 1} \frac{1}{n} \sum_{i=1}^n \big\{ f(T(A_i), \mathbb{A}_{i-1}) - f(g(Z_i, \mathbb{A}_{i-1}), \mathbb{A}_{i-1}) \big\}.$$

With the above definitions, $(T(A_i), \mathbb{A}_{i-1})$ is absolutely regular β-mixing with coefficients $\beta(k) = \beta_A((k-r) \vee 0)$. Then Theorem 4.3 implies:

Corollary 5.1. *Under the conditions (i)-(iv) of Theorem 4.3,*

$$\mathbb{E} W_{1,n}^c(\hat{g}_n^c) \lesssim \Big(\frac{s_f L_f \log(s_f L_f)}{n} \Big)^{1/2} + n^{-\frac{\beta}{2\beta + d_g}} \log(n)^{3/2}.$$

We obtain confidence intervals for $T(A_i)$ given $\mathbb{A}_{i-1} = \mathbb{a}$ as follows based on the results for conditional WGANs from Section 4.3. For $N \in \mathbb{N}$, let $Z_1^*, ..., Z_N^*$ denote i.i.d. realizations of $\mathbb{P}^Z$. Define the empirical distribution function given $\mathbb{A}_{i-1} = \mathbb{a}$,

$$\hat{F}_{N,n}(t|\mathbb{a}) := \frac{1}{N} \sum_{j=1}^N \mathbb{1}_{\{\hat{g}_n^c(Z_j^*, \mathbb{a}) \leq t\}}.$$

Then for $\alpha \in (0,1)$, the interval

$$I_{N,n}(\mathbb{a}) := \Big\{ t \in \mathbb{R} : \hat{F}_{N,n}(t|\mathbb{a}) \in \Big(\frac{\alpha}{2}, 1 - \frac{\alpha}{2} \Big] \Big\}$$

is an asymptotic $(1-\alpha)$ confidence interval for $T(A_i)$ given $\mathbb{A}_{i-1} = \mathbb{a}$.

WGAN-GP.	
	Require: β_1, β_2, learning rate α, penalty weight λ, batch size m, number of critic iterations per generator iteration n_{critic}.
0:	Initialize critic parameters θ_{critic} and generator parameters θ_{gen}.
1:	**while** θ_{gen} has not converged:
2:	**for** $t = 0, \dots, n_{\text{critic}}$:
3:	Sample a batch $\{X^{(i)}\}_{i=1}^m \sim \mathbb{P}^X$ from the real data.
4:	Sample i.i.d. batches $\{Z^{(i)}\}_{i=1}^m \sim \mathbb{P}^Z$, $\{U^{(i)}\}_{i=1}^m \sim U[0,1]$.
5:	Compute $\tilde{X}^{(i)} = U^{(i)} X^{(i)} + (1 - U^{(i)}) g_{\theta_{\text{gen}}}(Z^{(i)})$.
6:	$G_{\text{c}} \leftarrow \nabla_{\theta_{\text{critic}}} \big(\frac{1}{m} \sum_{i=1}^m f_{\theta_{\text{critic}}}(X^{(i)}) - \frac{1}{m} \sum_{i=1}^m f_{\theta_{\text{critic}}}(g_{\theta_{\text{gen}}}(Z^{(i)})) \big)$.
7:	$Pen_{\text{c}} \leftarrow \lambda \cdot \frac{1}{m} \sum_{i=1}^m \big(\|\nabla_{\tilde{X}_i} f_{\theta_{\text{critic}}}\|_2 - 1 \big)^2$.
8:	$\theta_{\text{critic}} \leftarrow \theta_{\text{critic}} + \alpha \cdot \text{ADAM}(G_c + Pen_{\text{c}}, \theta_{\text{critic}}, \beta_1, \beta_2)$.
9:	**end for**
10:	Sample an i.i.d. batch $\{Z^{(i)}\}_{i=1}^m \sim \mathbb{P}^Z$.
11:	$G_{\text{gen}} \leftarrow -\nabla_{\theta_{\text{gen}}} \frac{1}{m} \sum_{i=1}^m f_{\theta_{\text{critic}}}(g_{\theta_{\text{gen}}}(Z^{(i)}))$.
12:	$\theta_{\text{gen}} \leftarrow \theta_{\text{gen}} - \alpha \cdot \text{ADAM}(G_{\text{gen}}, \theta_{\text{gen}}, \beta_1, \beta_2)$.
13:	**end while**

TABLE 1

The gradient descent algorithm proposed in [16], with adapted default values $\alpha = 0.0001$, $\lambda = 0.1$, $m = 64$, $n_{critic} = 5$, $\beta_1 = 0.5$, $\beta_2 = 0.9$ and $\mathbb{P}^Z = U[0,1]$. The critic tries to maximize the empirical Wasserstein distance, while the generator has the contrary objective. The penalty term in line 7 softly enforces the Lipschitz constraint on the critic. For the first 25 and every 100^{th} generator iterations, we train the critic for 100 iterations for each generator iteration to ensure critic convergence and meaningful gradients. We additionally apply 0.01 L_2-weight decay to both networks to softly enforce boundedness of the network parameters.

6. Simulation studies

In this section we study the behaviour of an approximation of the optimal estimators $\hat{g}_n$ from (2.4) and $\hat{g}_n^c$ from (2.5) obtained by gradient descent methods. In all our experiments we use the WGAN-GP [16] algorithm with the adapted default values given in Table 1, if not stated otherwise.

We now give some comments on the WGAN-GP algorithm. In opposite to the original WGAN algorithm from [4] (cf. [1, Theorem 1]), which uses crude weight clipping to guarantee a bounded Lipschitz constant of the critic networks, WGAN-GP realizes the Lipschitz constraint in the definition of $\hat{W}_{1,n}$ via a penalty term. Furthermore, the critic is learned from $\hat{W}_{1,n}$ using a gradient descent method. Although the critic network may not obey $\|f\|_L \leq 1$ by using a penalty term, the equation

$$L \cdot \sup_{\|f\|_L \leq 1} \mathbb{E}f(X) - \mathbb{E}f(g(Z)) = \sup_{\|f\|_L \leq L} \mathbb{E}f(X) - \mathbb{E}f(g(Z)),$$

shows that it is enough to bound the Lipschitz constant of the critic by some (unknown) constant. WGAN-GP therefore recovers the distributional stability of the W_1-distance, induced by metrizing weak convergence (cf. also [3]). However, it should be noted that the new latent variables Z_{ij} generated in each training epoch may slightly change the bound on the Lipschitz constant introduced by the penalty term.

For Vanilla GANs generator and discriminator training have to be carefully balanced, because a discriminator that classifies too well does not yield informative gradients (the so-called saturation phenomenon). For WGANs, in contrast, better critics yield better gradients. Hence one only has to train the critic "long enough", which is a huge practical advantage. We can confirm stable training behaviour in all our experiments.

For the conditional setting, we simply replace all $f_\theta(X^{(i)})$ by $f_\theta(X^{(i)}, Y^{(i)})$ and all $g_\theta(Z^{(i)})$ by $g_\theta(Z^{(i)}, Y^{(i)})$ in Table 1.

In our simulation studies, we are particularly interested in how well $\hat{g}_n(Z)$ or $\hat{g}_n(Z, y)$ approximate the underlying true distribution of X or X given $Y = y$, respectively. We measure the approximation quality in our simulation studies with

- the empirical optimal transport distance, from now on OT, computed with the Python package POT [13], between an equal amount of real samples X_i and generated samples $\hat{g}_n(Z_j)$ as an estimator for $W_1(\mathbb{P}^X, \mathbb{P}^{\hat{g}_n(Z)})$ (X_i given $Y_i = y$ and $\hat{g}_n(Z_j, y)$ in the conditional case),
- empirical 95%-confidence intervals (we use the abbreviation CI95) as discussed in Section 3.3, Section 4.3 and Section 5, where we compute the empirical 2.5%- and 97.5%-quantiles of a statistic evaluated on N generated samples.

6.1. Synthetic data

Models: To analyze the performance of the WGAN estimator in the unconditional case, we use the following model: We generate data

$$X_i = g^*(Z_i), \quad i = 1, ..., n \tag{6.1}$$

via the transformation

$$g^*(z) = \big(\sin(z_1), \sin(z_2), \sin(z_3), \exp(z_1), z_2^2 + 2z_3^3, \cos(2\pi z_1 \cdot z_2 \cdot z_3), \\ z_1 \cdot z_2 \cdot z_3, (z_1 + z_2 + z_3)^2, z_1 + z_2 + z_3, 2x_1^4 - x_2^3\big),$$

where Z_i are i.i.d. uniformly distributed on $[0, 1]^3$. That is, $d = 10$ and $d_Z = 3$. Here, g is designed to contain a variety of smooth functions but also similarities between some of the coordinates.

For the conditional WGAN estimator, we consider the following model: With $d = 10, d_Z = 7$ and $d_Y = 3$, we simulate

$$X_i = g_c^*(Z_i, Y_i), \quad i = 1, ..., n, \tag{6.2}$$

where $g_c^* = g^* \circ h$ with

$$h(z_1, \ldots, z_7, y_1, y_2, y_3) := (z_1 + z_2^2 + z_3^3, z_4 \cdot z_5 + z_6 \cdot z_7, \sin(y_1) - y_2 \cdot y_3),$$

and (Z_i, Y_i), $i = 1, ..., n$ are i.i.d. uniformly distributed on $[0, 1]^{10}$. Here, g_c^* has a encoder-decoder structure according to Definition 2.2.

For simplicity, we consider independent observations Z_i, Y_i in both situations (that is, no serial correlation along $i = 1, ..., n$).

Results: We examine the convergence behaviour of WGANs for increasing sample size $n \to \infty$ in Table 2. In the unconditional case, we construct confidence intervals for the statistic $T(X_1)$, where $T(x) = \sum_{j=1}^{10} x_j$, using $N = 1000$ generated samples, by computing the empirical 2.5%- and 97.5%-quantiles of $\{T(\hat{g}(Z_j))\}_{j=1,...,N}$ (as in (3.4)). In the conditional case, we approximate the statistic $T(X|Y = y)$ for $y = (0.5, 0.5, 0.5)$, using $N = 1000$ generated samples, by computing the empirical 2.5%- and 97.5%-quantiles of $\{T(\hat{g}(Z_j, y))\}_{j=1,...,N}$ (as in (4.5)). Note that the coverage of the constructed confidence intervals approaches 95%, while the optimal transport distance decreases. According to our results, the network sizes should grow with n, but we use a fixed architecture that performs well for all given n to ensure comparability. Good coverage probabilities are already achieved for $n = 960$ or $n = 3200$, respectively. This highlights the fact that the WGAN is capable of detecting the sparse structure in the models (6.1) and (6.2) and realizes a faster convergence rate as announced in Theorem 3.4 and Theorem 4.3.

Measured quantity	Number of samples				
	64	320	960	3200	9600
CI95, unc.	47.92 (5.72)	52.26 (6.24)	96.16 (1.18)	94.50 (0.86)	**94.56** (0.84)
OT, unc.	1.634 (0.077)	1.630 (0.102)	0.970 (0.130)	0.412 (0.029)	**0.342** (0.026)
CI95, cond.	24.96 (3.13)	23.2 (1.67)	45.32 (7.27)	94.76 (1.93)	**94.78** (0.97)
OT, cond.	7.181 (0.187)	6.720 (0.392)	7.670 (0.307)	1.967 (0.562)	**1.297** (0.341)

TABLE 2

Shows the coverage probability (in %) of empirical 95%-confidence intervals for the sum of all components $T(x) = \sum_{j=1}^{10} x_j$ and the empirical optimal transport distance, each computed over $N = 1000$ new i.i.d. samples, after 700 epochs of training. We train each model 5 times with different i.i.d. data.The left number denotes the mean over all runs, while the number in parentheses denotes the empirical standard deviation. We use a discriminator with 5 hidden layers of size 128 and a generator with 3 hidden layers of size 32.

6.2. Real data application

Practical approaches for time series generation using conditional GANs have been conducted by [26] using conditional vanilla GANs and [29] using conditional WGANs. These works emphasize the potential of using generated data for data augmentation in other tasks, given small data sets (few shot learning).

For a real world simulation study we consider the mean temperatures $A_i \in \mathbb{R}^d$, $i = 1, ..., n = 4779$ of $d = 32$ German cities provided by the Deutscher Wetterdienst (German Metereological Service)[1]. Note that the chosen cities are spread throughout Germany, which can be seen in Figure 3.

In total we observe 4779 temperature values for each city over the period from 2006/07/01 to 2019/07/31. In the notation of Section 5, given the temperatures

[1] https://opendata.dwd.de/climate_environment/CDC/observations_germany/climate/daily/kl/historical

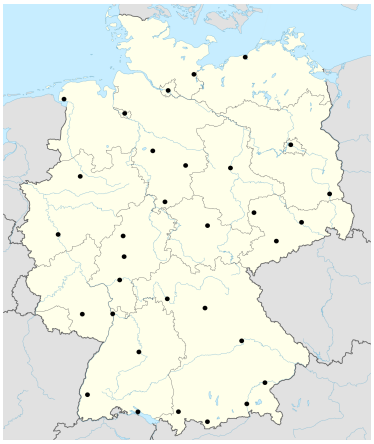


Fig 3: The authors of [25] collected weather data from the cities of Berlin, Braunschweig, Bremen, Chemnitz, Cottbus, Dresden, Erfurt, Frankfurt, Freiburg, Garmisch-Patenkirchen, Göttingen, Münster, Hamburg, Hannover, Kaiserslautern, Kempten, Köln, Konstanz, Leipzig, Lübeck, Magdeburg, Cölbe, Mühldorf, München, Nürnberg, Regensburg, Rosenheim, Rostock, Stuttgart, Würzburg, Emden and Mannheim.

of several cities of the previous day A_{i-1} (so using only one lag $r = 1$), we predict the temperature in Berlin $T(A_i) = A_{i1}$ of each day, i.e. $r = 1$ and $T(x) = x_1$. The first day is not predicted. We use the first $n_{train} = 4300$ days for training and the remaining $n_{test} = 478$ days from 2018/04/10 to 2019/07/31 for testing. We train cWGANs with 4-dimensional standard normal noise and use A_{i-1} as conditional information. We train 3 different models.

(M1) The first model only predicts the temperature in Berlin and A_{i-1} only consists of the temperatures in Berlin, Braunschweig and Bremen of the previous day.

(M2) The second model only predicts the temperature in Berlin but A_{i-1} consists of the temperatures in all 32 cities of the previous day.

(M3) The third model predicts the temperatures in all 32 cities and A_{i-1} consists of the temperatures in all 32 cities of the previous day. The quality of the confidence intervals is only assessed for 1 city, namely Berlin.

For all models we use the generators with 3 hidden layers with 10 neurons each and a discriminator with 5 hidden layers with 32 neurons each. Table 3 shows the progression over 1000 epochs of training. Table 4 complements these illustrations with OT and CI95 values after 1000 epochs of training. The confidence intervals $I_{N,n}(A_{i-1})$ (cf. (4.5)) for $T(A_i)$ are constructed from $N = 1000$ realizations of $\mathbb{P}^Z$ when i belongs to the training set and from $N = 10000$ realizations of $\mathbb{P}^Z$ when i belongs to the test set.

The optimal transport distance is computed jointly over the A_{i-1} and the

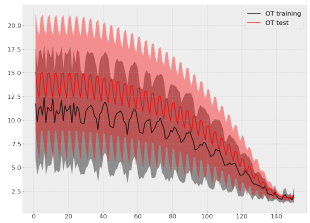
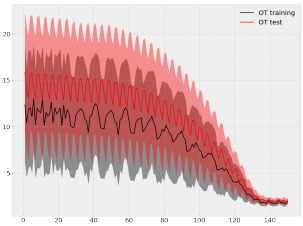
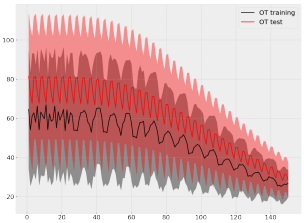
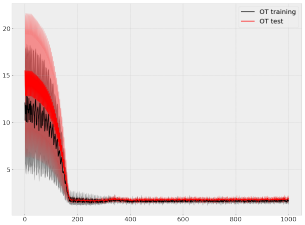
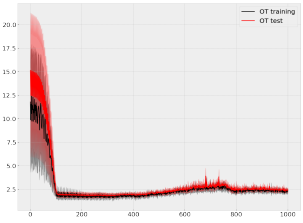
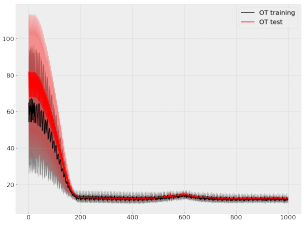
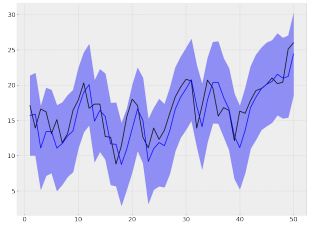
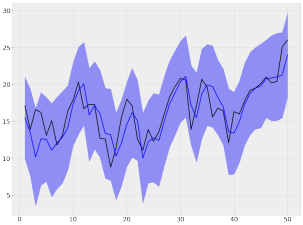
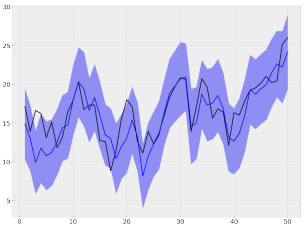
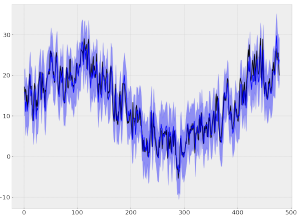
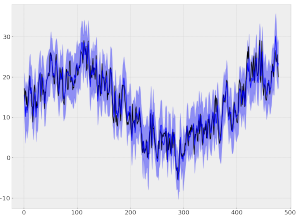
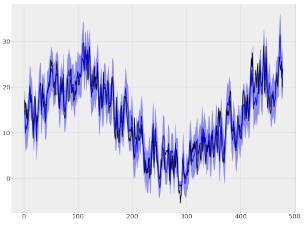
(M1)	(M2)	(M3)
		
		
		
		

TABLE 3

Training process of a cWGAN depicted after 1000 epochs over temperature data for the models (M1)-(M3).

The first two rows show the estimated optimal transport distance between real and generated data evaluated 10 times on 1 batch from the training set (black) and test set (red), over 150 and 1000 epochs respectively. Depicted is the mean and a $\pm\sigma$-confidence band. The other two rows show the actual mean temperatures in Berlin (black) and the average generated prediction (blue) with a $\pm 3\sigma$-confidence band over 1000 generated points per time step, for the first 50 days of the test set and the whole test set respectively.

predicted/real temperatures in Berlin and still has high variance as the dimension is 4 for (M1) and 33 for (M2) and (M3). Since in the first model A_{i-1} is only 3-dimensional, the optimal transport distance is only comparable between the second and third model.

	Three to Berlin	All to Berlin	All to All
Test OT, 1000 epochs	1.39	2.38	2.37
Empirical 95%-confidence, train	92.60%	91.14%	74.02%
Empirical 95%-confidence, test	89.96%	89.54%	70.71%

TABLE 4

The first row computes the empirical optimal transport distance (OT) on the whole test set. Rows 2 and 3 show the respective proportion of $T(A_i)$ lying in the intervals $I_{N,n}(A_{i-1})$ between the empirical 2.5%- and 97.5%-quantiles, for the training and test set, respectively.

Note that in all scenarios, the generator firstly learns the predictions with overconfidence. For (M1), the training data lies much more densely in the A_{i-1}-space. (M1) learns the distribution much faster, but eventually the (M2)-model achieves comparable performance. The most complex (M3)-model takes the longest to converge but already performs decently considering that it performs a 32-dimensional prediction with the same small generator architecture. To have comparable results, we used 1000 training epochs for the estimators in Table 4. However, the quality of the estimators may still increase for more training epochs. For instance, after 2000 epochs of training, we achieve 2.35 OT and 86.19% coverage on the test set for the temperature in the city Berlin in model (M3). Overall, the results are quite satisfying and motivate that the cWGAN estimator is able to find some sparse underlying structures in the data.

7. Conclusion

To our knowledge, this paper is the first where convergence rates for the excess Bayes risk of Wasserstein GANs and conditional Wasserstein GANs are derived under structural assumptions on the space of generators.

We have formalized the empirical WGAN objective with growing critic networks and have shown that this objective still metrizes weak convergence. Our results yield recommendations on the size of generator networks and unveil the potential use of conditional WGANs in high-dimensional time series forecasting, in particular the construction of confidence intervals. All our results hold for dependent data, where the dependence is measured with absolutely regular β-mixing. Both our synthetic and real world simulations demonstrate good empirical coverage for confidence intervals in multidimensional applications.

Additionally, we have included a first approach to formalize the availability of a growing number of observations when training is performed with multiple epochs. The corresponding result justifies the use of very large generator networks without suffering from slow convergence rates. Our attempt could be explored in other contexts and extended to the conditional case. In future work, one could also try to study the convergence behaviour of local instead of global

minimizers of the empirical WGAN objectives. Furthermore, it would be interesting to include the gradient penalty in the theoretical results and use other GAN losses or network architectures, such as the Groupsort activation function [1] for the critic. It would be interesting to refine the approximation results from [28] to gain more insight into the theoretical properties of our modified Wasserstein distance $W_{1,n}$.

References

[1] C. Anil, J. Lucas, and R. Grosse. Sorting out lipschitz function approximation, 2018.

[2] Grigory Antipov, Moez Baccouche, and Jean-Luc Dugelay. Face aging with conditional generative adversarial networks, 2017.

[3] M. Arjovsky and L. Bottou. Towards principled methods for training generative adversarial networks, 2017.

[4] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein gan, 2017.

[5] H. C. P. Berbee. Random walks with stationary increments and renewal theory. mathematisch centrum, amsterdam. *Mathematical Centre Tracts*, 112, 1979.

[6] G. Biau, B. Cadre, M. Sangnier, and U. Tanielian. Some theoretical properties of gans, 2018.

[7] Gérard Biau, Maxime Sangnier, and Ugo Tanielian. Some theoretical insights into wasserstein gans, 2020.

[8] Olivier Bousquet. A bennett concentration inequality and its application to suprema of empirical processes. *Comptes Rendus Mathematique*, 334(6):495 – 500, 2002.

[9] Richard C. Bradley. Basic properties of strong mixing conditions. a survey and some open questions. *Probability Surveys*, 2(0):107–144, 2005.

[10] S. Dedecker and S. Louhichi. Maximal inequalities and empirical central limit theorems. *Empirical Process Techniques for Dependent Data*, pages 137–159, 2002.

[11] P. Doukhan. *Mixing: Properties and Examples*. Lecture Notes in Statistics. Springer New York, 2012.

[12] P. Doukhan, P. Massart, and E. Rio. Invariance principles for absolutely regular empirical processes. *Annales de l'I.H.P. Probabilités et statistiques*, 31(2):393–427, 1995.

[13] R. Flamary and N. Courty. Pot python optimal transport library, 2017.

[14] Piotr Fryzlewicz and Suhasini Subba Rao. Mixing properties of arch and time-varying arch processes. *Bernoulli*, 17(1):320–346, Feb 2011.

[15] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks, 2014.

[16] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and Aaron Courville. Improved training of wasserstein gans, 2017.

[17] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation

with conditional adversarial networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017)*, 2017.

[18] T. Klein and E. Rio. Concentration around the mean for maxima of empirical processes. *Ann. Probab.*, 33(3):1060–1077, 05 2005.

[19] T. Liang. On the minimax optimality of estimating the wasserstein metric, 2019.

[20] I. Malkiel, S. Ahn, V. Taviani, A. Menini, L. Wolf, and C. J. Hardy. Conditional wgans with adaptive gradient balancing for sparse mri reconstruction, 2019.

[21] M. Mirza and S. Osindero. Conditional generative adversarial nets, 2014.

[22] A. Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997.

[23] J. Niles-Weed and Q. Berthet. Minimax estimation of smooth densities in wasserstein distance, 2019.

[24] J. Niles-Weed and P. Rigollet. Estimation of wasserstein distances in the spiked transport model, 2019.

[25] Nathawut Phandoidaen and Stefan Richter. Forecasting time series with encoder-decoder neural networks, 2020.

[26] G. Ramponi, P. Protopapas, M. Brambilla, and R. Janssen. T-cgan: Conditional generative adversarial network for data augmentation in noisy time series with irregular sampling, 2019.

[27] Emmanuel Rio. Inequalities and limit theorems for weakly dependent sequences. Lecture, September 2013.

[28] J. Schmidt-Hieber. Nonparametric regression using deep neural networks with relu activation function, 2017.

[29] Kaleb E Smith and Anthony O Smith. Conditional gan for timeseries generation, 2020.

[30] C. Villani. *Optimal transport – Old and new*, volume 338, pages 43–113. 01 2008.

Appendix A: Proofs of Section 3

Proof of Lemma 3.2. By Lemma 3.1,

$$\mathbb{E}W_1^\gamma(\hat{g}_n) \le \mathbb{E}W_{1,n}(\hat{g}_n) + C\phi_n \to 0.$$

Let $f \in C^\gamma([0,1]^d, 1)$ be arbitrary. By independence of Z and $\hat{g}_n(\cdot)$, it follows that $\mathbb{E}[\mathbb{E}[f(g(Z))]\big|_{g=\hat{g}_n}] = \mathbb{E}[\mathbb{E}[f(\hat{g}_n(Z))|\hat{g}_n]] = \mathbb{E}f(\hat{g}_n(Z))$. Thus

$$|\mathbb{E}f(X) - \mathbb{E}f(\hat{g}_n(Z))| \le \mathbb{E}W_1^\gamma(\hat{g}_n) \to 0. \tag{A.1}$$

We now show weak convergence $\hat{g}_n(Z) \xrightarrow{d} X$. Let $A \subset [0,1]^d$ be a closed set (and thus compact). Let $\varepsilon > 0$. Define $\rho_{\varepsilon,A}(x) := 1 - \varphi(\frac{d(x,A)}{\varepsilon})$, where $d(x,A) := \inf_{y\in A} \|x-y\|_\infty$ and $\varphi : \mathbb{R} \to [0,1]$ is an arbitrary infinitely differentiable function with $\varphi(x) = 0$ for $x \le 0$ and $\varphi(x) = 1$ for $x \ge 1$, for instance one may define $\varphi(x) = e^{-1/x} \cdot (e^{-1/x} + e^{-1/(1-x)})^{-1}$ for $x \in [0,1]$.

Note that $\mathbb{1}_A(x) \le \rho_{\varepsilon,A}(x)$. Furthermore, for $d(\varepsilon) > 0$ small enough, $d(\varepsilon) \cdot \rho_{\varepsilon,A} \in C^\gamma([0,1]^d, 1)$. By these arguments and (A.1),

$$\begin{aligned} \mathbb{P}(\hat{g}_n(Z) \in A) &\le \mathbb{E}\rho_{\varepsilon,A}(\hat{g}_n(Z)) = d(\varepsilon)^{-1}\mathbb{E}[d(\varepsilon)\rho_{\varepsilon,A}(\hat{g}_n(Z))] \\ &\to d(\varepsilon)^{-1}\mathbb{E}[d(\varepsilon)\rho_{\varepsilon,A}(X)] = \mathbb{E}\rho_{\varepsilon,A}(X). \end{aligned}$$

We conclude that $\limsup_{n\to\infty} \mathbb{P}(\hat{g}_n(Z) \in A) \le \mathbb{E}\rho_{\varepsilon,A}(X)$. Since $\rho_{\varepsilon,A}(x) \to \mathbb{1}_A(x)$ for $\varepsilon \to 0$, the dominated convergence theorem implies $\limsup_{n\to\infty} \mathbb{P}(\hat{g}_n(Z) \in A) \le \mathbb{P}(X \in A)$. The result now follows from the portmanteau lemma for weak convergence. □

Proof of Lemma 3.6. Fix $x \in \mathbb{R}$. By the law of large numbers (applied conditionally on $\hat{g}_n$), we have that almost surely, for $N \to \infty$,

$$\hat{F}_{N,n}(x) \to \mathbb{P}(\hat{g}_n(Z) \le x|\hat{g}_n). \tag{A.2}$$

We now conduct a similar argumentation as in the proof of Lemma 3.2 based on the stochastic convergence $W_1^\gamma(\hat{g}_n) \xrightarrow{p} 0$. Let $A \subset [0,1]$ be a closed subset. Then

$$\begin{aligned} &\mathbb{P}\big(\mathbb{P}(\hat{g}_n(Z) \in A|\hat{g}_n) - \mathbb{P}(X \in A) \ge \rho\big) \\ \le\ &\mathbb{P}\big(\mathbb{E}[d(\varepsilon)\rho_{\varepsilon,A}(\hat{g}_n(Z))|\hat{g}_n] - \mathbb{E}[d(\varepsilon)\rho_{\varepsilon,A}(X)] \ge \frac{\rho}{2}d(\varepsilon)\big) \\ &\quad + \mathbb{P}\big(\mathbb{E}\rho_{\varepsilon,A}(X) - \mathbb{P}(X \in A) \ge \frac{\rho}{2}\big). \end{aligned}$$

While the second summand is 0 for $\varepsilon > 0$ small enough, the first summand tends to zero by $W_1^\gamma(\hat{g}_n) \xrightarrow{p} 0$. This shows that

$$\lim_{n\to\infty} \mathbb{P}\big(\mathbb{P}(\hat{g}_n(Z) \in A|\hat{g}_n) - \mathbb{P}(X \in A) \ge \rho\big) = 0.$$

Using typical proof strategies from the portemanteau lemma, we first see that for any open subset $U \subset [0,1]$, and any $\rho > 0$,

$$\lim_{n\to\infty} \mathbb{P}\big(\mathbb{P}(\hat{g}_n(Z) \in U|\hat{g}_n) - \mathbb{P}(X \in U) \leq -\rho\big) = 0.$$

and for any $x \in [0,1]$ which is a continuity point of F_X and any $\rho > 0$,

$$\lim_{n\to\infty} \mathbb{P}\big(\big|\mathbb{P}(\hat{g}_n(Z) \leq x|\hat{g}_n) - F_X(x)\big| \geq \rho\big) = 0. \tag{A.3}$$

From (A.2) and (A.3) we obtain that for $\rho > 0$ and any $x \in \mathbb{R}$,

$$\begin{aligned} & \limsup_{n\to\infty} \limsup_{N\to\infty} \mathbb{P}\big(\big|\hat{F}_{N,n}(x) - F_X(x)\big| \geq \rho\big) \\ \leq\ & \limsup_{N\to\infty} \mathbb{P}\big(\big|\hat{F}_{N,n}(x) - \mathbb{P}(\hat{g}_n(Z) \leq x|\hat{g}_n)\big| \geq \frac{\rho}{2}\big) \\ & \quad + \limsup_{n\to\infty} \mathbb{P}\big(\big|\mathbb{P}(\hat{g}_n(Z) \leq x|\hat{g}_n) - F_X(x)\big| \geq \frac{\rho}{2}\big) = 0. \end{aligned}$$

By continuity of F_X, standard decomposition arguments from the Polya theorem about uniform convergence of distribution functions provide

$$\limsup_{n\to\infty} \limsup_{N\to\infty} \mathbb{P}\big(\sup_{x\in[0,1]} \big|\hat{F}_{N,n}(x) - F_X(x)\big| \geq \rho\big) = 0.$$

The result of the lemma now follows for plugging in $x = X$. □

Appendix B: Error Decomposition

B.1. Unconditional WGAN: Basic inequality

We abbreviate $\mathcal{G} = \mathcal{G}(d_Z, d_g, \beta, K)$ and $\mathcal{R}_G := \mathcal{R}(L_g, \mathbf{p}_g, s_g)$, $\mathcal{R}_D = \mathcal{R}(L_f, \mathbf{p}_f, s_f)$.

Recall from (2.4) that

$$\hat{g}_n = \operatorname*{arg\,min}_{g\in\mathcal{R}_G} \hat{W}_{1,n}(g).$$

Proposition B.1 (WGAN: Basic inequality)**.** *It holds that*

$$W_{1,n}(\hat{g}_n) - \inf_{g\in\mathcal{G}} W_{1,n}(g) \leq \sqrt{d} \cdot A_n + 2 \cdot E_n,$$

where

$$A_n := \sup_{g\in\mathcal{G}} \inf_{\tilde{g}\in\mathcal{R}_G} \|g - \tilde{g}\|_\infty, \tag{B.1}$$

$$E_n := \sup_{f\in\mathcal{R}_D} |(\hat{\mathbb{P}}_n^X - \mathbb{P}^X)f| + \sup_{g\in\mathcal{R}_G, f\in\mathcal{R}_D} |(\hat{\mathbb{P}}_{n\mathcal{E}}^Z - \mathbb{P}^Z)(f\circ g)|. \tag{B.2}$$

Proof of Proposition B.1. First, we have

$$W_{1,n}(\hat{g}_n) - \inf_{g\in\mathcal{G}} W_{1,n}(g) \le e_n + a_n,$$

where

$$e_n := W_{1,n}(\hat{g}_n) - \inf_{g\in\mathcal{R}_G} W_{1,n}(g)$$

is the estimation error and

$$a_n := \inf_{g\in\mathcal{R}_G} W_{1,n}(g) - \inf_{g\in\mathcal{G}} W_{1,n}(g)$$

is the approximation error. Note that A_n is upper bounded (not in absolute value!) as follows:

$$a_n \le \sup_{g\in\mathcal{G}} \inf_{\tilde{g}\in\mathcal{R}_G} |W_{1,n}(g) - W_{1,n}(\tilde{g})|. \tag{B.3}$$

Since all functions f in the supremum in $W_{1,n}$ satisfy $\|f\|_L \le 1$, we have

$$\begin{aligned}
& |W_{1,n}(g) - W_{1,n}(\tilde{g})| \\
\le\ & \Big| \sup_{f\in\mathcal{R}_D,\|f\|_L\le 1} \big\{\mathbb{E}f(X) - \mathbb{E}f(g(Z))\big\} - \sup_{f\in\mathcal{R}_D,\|f\|_L\le 1} \big\{\mathbb{E}f(X) - \mathbb{E}f(\tilde{g}(Z))\big\}\Big| \\
\le\ & \sup_{f\in\mathcal{R}_D,\|f\|_L\le 1} \mathbb{E}|f(g(Z)) - f(\tilde{g}(Z))| \\
\le\ & \sqrt{d}\|g - \tilde{g}\|_\infty.
\end{aligned}$$

We conclude from (B.3) that

$$a_n \le \sqrt{d} \inf_{g\in\mathcal{R}_G} \sup_{\tilde{g}\in\mathcal{G}} \|g - \tilde{g}\|_\infty.$$

We now investigate the estimation error E_n. Let $\varepsilon > 0$. Then there exists $g^* \in \mathcal{R}_G$ with $\inf_{g\in\mathcal{G}} W_{1,n}(g) \le W_{1,n}(g^*) + \varepsilon$. We obtain

$$e_n = W_{1,n}(\hat{g}_n) - \inf_{g\in\mathcal{G}} W_{1,n}(g) \le W_{1,n}(\hat{g}_n) - W_{1,n}(g^*) + -\varepsilon. \tag{B.4}$$

In order to bound $W_{1,n}(\hat{g}_n) - W_{1,n}(g^*)$, note that by the minimization property of $\hat{g}_n$,

$$\begin{aligned}
& W_{1,n}(\hat{g}_n) - W_{1,n}(g^*) \\
=\ & \hat{W}_{1,n}(\hat{g}_n) - \hat{W}_{1,n}(g^*) \\
& \quad - \Big(\{\hat{W}_{1,n}(\hat{g}_{m,n}) - W_{1,n}(\hat{g}_{m,n})\} - \{\hat{W}_{1,n}(g^*) - W_{1,n}(g^*)\}\Big) \\
\le\ & 2 \sup_{g\in\mathcal{R}_G} \big|\hat{W}_{1,n}(g) - W_{1,n}(g)\big|.
\end{aligned}$$

Letting $\varepsilon \downarrow 0$, we obtain from (B.4) that

$$e_n \le 2 \sup_{g\in\mathcal{R}_G} \big|\hat{W}_{1,n}(g) - W_{1,n}(g)\big|. \tag{B.5}$$

Note that

$$\begin{aligned} & \sup_{g\in\mathcal{R}_G} |\hat{W}_{1,n}(g) - W_{1,n}(g)| \\ \leq & \sup_{g\in\mathcal{R}_G} \Big| \sup_{f\in\mathcal{R}_D} \big\{(\hat{\mathbb{P}}_n^X - \mathbb{P}^X)f - (\hat{\mathbb{P}}_{n\mathcal{E}}^Z - \mathbb{P}^Z)(f\circ g)\big\}\Big| \\ \leq & \sup_{f\in\mathcal{R}_D} |(\hat{\mathbb{P}}_n^X - \mathbb{P}^X)f| + \sup_{g\in\mathcal{R}_G, f\in\mathcal{R}_D} |(\hat{\mathbb{P}}_{n\mathcal{E}}^Z - \mathbb{P}^Z)(f\circ g)| \end{aligned}$$

Insertion into (B.5) yields the assertion. □

B.2. Approximation error

To bound the approximation error A_n, we use the approximation theory from [28] and statements about the Lipschitz constant in [25].

Theorem B.2 ([28], Theorem 5 and [25], Theorem 9.14)**.** *For all*

$$h \in C^\beta([0,1]^r, K), \quad k \geq 1 \quad \text{and} \quad N \geq (\beta+1)^r \vee (K+1)e^r,$$

there exists a network

$$\tilde{h} \in \mathcal{R}\big(L, (r, 6(r+\lceil\beta\rceil)N, \dots, 6(r+\lceil\beta\rceil)N, 1), s, \infty\big)$$

with

$$L = 8 + (k+5)(1+\lceil\log_2(r\vee\beta)\rceil) \quad \text{and} \quad s \leq 141(r+\beta+1)^{3+r}N(k+6),$$

such that,

$$\|h - \tilde{h}\|_{L^\infty([0,1]^r)} \leq (2K+1)(1+r^2+\beta^2)6^r N 2^{-k} + K3^\beta N^{-\beta/r}.$$

Furthermore, $\tilde{h}$ satisfies for any $x, y \in [0,1]^r$ that

$$|\tilde{h}(x) - \tilde{h}(y)| \leq \mathrm{Lip}(N,k)\cdot|x-y|_\infty,$$

where

$$\mathrm{Lip}(N,k) := 2\beta F(K+1)e^r(24r^6 2^r N 2^{-k} + 3r).$$

Lemma B.3. *Let $\beta \geq 1$, $d_g \in \mathbb{N}, \mathcal{E} \in \mathbb{N}$. Let $N \geq (\beta+1)^{d_g} \vee (K+1)e^{d_g}$. If $\mathcal{R}_G = \mathcal{R}(L_g, \mathbf{p}_g, s_g)$ satisfies $F \geq K \vee 1$ and*

$$L_g \geq \log_2(n\mathcal{E})\log_2(4d_g \vee 4\beta), \quad \min_{i=1,\dots,L_g} p_i \gtrsim dN \quad \text{and} \quad s_g \gtrsim dN\log_2(n\mathcal{E}),$$

where the bounding constants only depend on β, d_g, then A_n from (B.1) satisfies that for n large enough,

$$A_n \lesssim \frac{N}{n\mathcal{E}} + N^{-\beta/d_g},$$

where the bounding constants only depend on β, d_g and K.

Proof of Lemma B.3. Given $g \in \mathcal{G}$, we can write $g_i \in C^\beta([0,1]^{d_g}, K)$, since each component function only depends on d_g arguments. Applying Theorem B.2 with $k = \lceil \log_2(n\mathcal{E}) \rceil$ to each component function yields that there exists a $\tilde{g}$ in the class

$$\mathcal{R}\left(L, (d_g, 6(d_g + \lceil\beta\rceil)N, \dots, 6(d_g + \lceil\beta\rceil)N, 1), s_g, \infty\right),$$

such that $||g_i - \tilde{g}_i||_\infty \lesssim N2^{-k} + N^{-\beta/d_g}$, where $L = k\lceil \log_2(4d_g \vee 4\beta)\rceil$, $s \lesssim Nk$ and the bounding constants only depend on K, d_g, β.
Thus a network computing all $\tilde{g} := (\tilde{g}_i)_{i=1,\dots,d}$ in parallel lies in the class

$$\mathcal{R}\left(L, (d_g, 6d(d_g + \lceil\beta\rceil)N, \dots, 6d(d_g + \lceil\beta\rceil)N, d), ds, \infty\right),$$

and it holds that

$$||g - \tilde{g}||_\infty \lesssim N2^{-k} + N^{-\beta/d_g}. \tag{B.6}$$

$\tilde{g}$ may not satisfy $\|\tilde{g}\|_\infty \leq F$. However, $\tilde{g}^\circ := (\frac{\|g\|_\infty}{\|\tilde{g}\|_\infty} \wedge 1)\tilde{g}$ still fulfills $\tilde{g}^\circ \in \mathcal{R}(L, p, s)$ and $\|\tilde{g}^\circ\|_\infty \leq \|g\|_\infty \leq K \leq F$. Due to $\|\tilde{g}^\circ - g\|_\infty \leq 2\|\tilde{g} - g\|_\infty$, (B.6) still holds for $\tilde{g}^\circ$ with changed constants. □

B.3. Estimation error

To upper bound the entropy bracketing numbers of the neural network sets $\mathcal{R}(L, \mathbf{p}, s)$, we use the following Lemma taken from [28].

For a class $\mathcal{F} \subset \{f : \mathbb{R}^r \to \mathbb{R} \text{ measurable}\}$ and some norm $\|\cdot\|$ on $\mathcal{F}$, we denote by $N_{[]}(\varepsilon, \mathcal{F}, \|\cdot\|)$ the number of ε-brackets which are needed to cover $\mathcal{F}$. Here, an ε-bracket $[l, u]$ is a set $[l, u] = \{f \in \mathcal{F} | \forall x \in \mathbb{R}^r : l(x) \leq f(x) \leq u(x)\}$ such that $\|u - l\| \leq \varepsilon$.

The bracketing entropy integral of $\mathcal{F}$ with respect to $\|\cdot\|$ is denoted by

$$J_{[]}(\delta, \mathcal{F}, \|\cdot\|) = \int_0^\delta \sqrt{1 + \log N_{[]}(\varepsilon, \mathcal{F}, \|\cdot\|)}\, \mathrm{d}\varepsilon.$$

The covering numbers $N(\varepsilon, \mathcal{F}, \|\cdot\|)$ denote the least number of elements $v_1, ..., v_m \in \mathcal{F}$ such that $\mathcal{F} \subset \bigcup_{j=1}^m \{y \in \mathcal{F} : \|y - v_j\| < \varepsilon\}$. Accordingly, we define the covering entropy integral $J(\delta, \mathcal{F}, \|\cdot\|) = \int_0^\delta \sqrt{1 + \log N(\varepsilon, \mathcal{F}, \|\cdot\|)}\, \mathrm{d}\varepsilon$. We need both bracketing and covering numbers since the approximation results in [28] were defined in terms of covering numbers while the empirical process results of [10] are in terms of bracketing numbers. However, there is a simple connection via

$$N_{[]}(\delta, \mathcal{F}, \|\cdot\|) \leq N(\frac{\delta}{2}, \mathcal{F}, \|\cdot\|). \tag{B.7}$$

For mixing coefficients $\beta_X(k)$, $k \in \mathbb{N}_0$, [12] defined the $\|f\|_{2,\beta}$-norm as follows: Let β_X^{-1} be the cadlag inverse of $\beta_X(t) = \beta(\lfloor t \rfloor)$ for $t \geq 1$ and $\beta_X(t) = 1$ otherwise. Let Q_f be the inverse of the tail function $t \mapsto \mathbb{P}(|f(X_1)| > t)$. Define

$$\|f\|_{2,\beta} = \Big(\int_0^1 \beta_X^{-1}(u) Q_f(u)^2 du \Big)^{1/2}.$$

In [12, Lemma 1] it is stated that for $B := \sum_{k=0}^{\infty} \beta_X(k)$, one has

$$\|f\|_{2,\beta} \leq B^{1/2} \cdot \|f\|_{\infty}. \tag{B.8}$$

Lemma B.4 ([28], Lemma 5)**.** *For all $\delta > 0$, it holds that*

$$\log\Big(N\big(\delta, \mathcal{R}(L, \mathbf{p}, s, \infty), \|\cdot\|_{\infty}\big)\Big) \leq (s+1)\log\Big(2\frac{L+1}{\delta}\big(\prod_{l=0}^{L+1}(p_l+1)\big)^2\Big).$$

In the following, we use the following abbreviation

$$\gamma(L, \mathbf{p}, s) := 2(s+1)\log\big(4(L+1)\prod_{l=0}^{L+1}(p_l+1)\big). \tag{B.9}$$

The following lemma is the basic result we use to bound the estimation error both in expectation and with high probability. It makes use of maximal inequalities and large deviation bounds derived in Section C for mixing sequences.

Lemma B.5. *Suppose that there exist constants $\kappa > 1, \alpha > 1$ such that for all $k \in \mathbb{N}$, $\beta_X(k) \leq \kappa \cdot k^{-\alpha}$. Suppose that*

$$\gamma(L, \mathbf{p}, s) \leq n.$$

Then there exists a constant $C > 0$ only depending on characteristics of (X_i) and B, F, κ, α such that

$$\mathbb{E}^* \sup_{f \in \mathcal{R}(L,\mathbf{p},s)} |(\hat{\mathbb{P}}_n^X - \mathbb{P}^X)f| \leq C \cdot \Big(\frac{\gamma(L, \mathbf{p}, s)}{n}\Big)^{1/2}. \tag{B.10}$$

Furthermore, with probability at least $1 - 2n^{-1} - (\frac{\log(n)}{n})^{\frac{\alpha-1}{2}}$ and a different constant $C > 0$ depending on the same quantities,

$$\sup_{f \in \mathcal{R}(L,\mathbf{p},s)} |(\hat{\mathbb{P}}_n^X - \mathbb{P}^X)f| \leq C \cdot \Big[\Big(\frac{\gamma(L, \mathbf{p}, s)}{n}\Big)^{1/2} + \Big(\frac{\log(n)}{n}\Big)^{1/2}\Big]. \tag{B.11}$$

Proof of Lemma B.5. We abbreviate $\mathcal{R} = \mathcal{R}(L, \mathbf{p}, s, F)$. Using Lemma B.4, we get for all $\delta > 0$,

$$\log N(\delta, \mathcal{R}, \|\cdot\|_{\infty}) \leq \gamma(L, \mathbf{p}, s) - (s+1)\log(\delta).$$

Using the simple bound $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, the bracketing integral is upper bounded by

$$\begin{aligned} J(\delta, \mathcal{R}, \|\cdot\|_{\infty}) &= \int_0^{\delta} \sqrt{1 + \gamma(L, \mathbf{p}, s) - (s+1)\log(\varepsilon)}\, d\varepsilon \\ &\leq \int_0^{\delta} \big(1 + \sqrt{\gamma(L, \mathbf{p}, s)}\big) d\varepsilon + \sqrt{s+1} \int_0^1 \sqrt{-\log(\varepsilon)}\, d\varepsilon \\ &= \delta + \sqrt{\gamma(L, \mathbf{p}, s)}\delta + \frac{\sqrt{(s+1)\pi}}{2} \leq c \cdot \gamma(L, \mathbf{p}, s)^{1/2}(1+\delta), \end{aligned}$$

where $c \geq 1$ is some universal constant.

By Lemma C.1, we have with some constants $K_1, K_2 > 0$ only depending on characteristics of X_1,

$$\mathbb{E}^* \sup_{f \in \mathcal{R}} \left|(\hat{\mathbb{P}}_n^X - \mathbb{P}^X)f\right| \leq r_n,$$

where

$$\begin{aligned} r_n &:= K_1 \cdot n^{-1/2} J_{[]}(F, \mathcal{R}, \|\cdot\|_\infty) + K_2 F \cdot \Big(\frac{1 \vee N_{[]}(2BF, \mathcal{R}, \|\cdot\|_\infty)}{n}\Big)^{\frac{\alpha}{\alpha+1}} \\ &\leq K_1 \cdot n^{-1/2} J(\frac{F}{2}, \mathcal{R}, \|\cdot\|_\infty) + K_2 F \cdot \Big(\frac{1 \vee N(BF, \mathcal{R}, \|\cdot\|_\infty)}{n}\Big)^{\frac{\alpha}{\alpha+1}} \\ &\leq C \cdot \Big((\frac{\gamma(L, \mathbf{p}, s)}{n})^{1/2} + (\frac{\gamma(L, \mathbf{p}, s)}{n})^{\frac{\alpha}{\alpha+1}}\Big), \end{aligned}$$

and $C > 0$ depends on F, B, K_1, K_2. Since $\gamma(L, \mathbf{p}, s) \leq n$ and $\alpha > 1$, the second summand is dominated by the first. This yields (B.10).

Note that $\mathcal{R}(L, \mathbf{p}, s, F)$ is separable in $\{f : [0,1]^d \to \mathbb{R} \text{ meas.}, \|f\|_\infty \leq F\}$, therefore $\sup_{f \in \mathcal{R}(L,\mathbf{p},s)} \left|(\hat{\mathbb{P}}_n^X - \mathbb{P}^X)f\right|$ is measurable and

$$\mathbb{P}\Big(\sup_{f \in \mathcal{R}(L,\mathbf{p},s)} \left|(\hat{\mathbb{P}}_n^X - \mathbb{P}^X)f\right| > x\Big) \leq \sup_{S \subset \mathcal{R}(L,\mathbf{p},s) \text{ countable}} \mathbb{P}\Big(\sup_{f \in S} \left|(\hat{\mathbb{P}}_n^X - \mathbb{P}^X)f\right| > x\Big).$$

By Lemma C.5, there exists some constant $C_2 > 0$ depending on F, B, κ, α such that

$$\mathbb{P}\Big(\sup_{f \in \mathcal{R}(L,\mathbf{p},s)} \left|(\hat{\mathbb{P}}_n^X - \mathbb{P}^X)f\right| \geq C_2 \cdot \big(r_n + (\frac{x}{n})^{1/2} + \frac{x}{n} \cdot z^{-\frac{1}{\alpha+1}}\big)\Big) \leq 2\exp(-x) + \frac{nz}{x}.$$

With $x = \log(n)$ and $z = \big(\frac{\log(n)}{n}\big)^{\frac{\alpha+1}{2}}$, we obtain

$$\mathbb{P}\Big(\sup_{f \in \mathcal{R}(L,\mathbf{p},s)} \left|(\hat{\mathbb{P}}_n^X - \mathbb{P}^X)f\right| \geq C_2 \cdot \big(r_n + 2(\frac{\log(n)}{n})^{1/2}\big)\Big) \leq \frac{2}{n} + \Big(\frac{\log(n)}{n}\Big)^{\frac{\alpha-1}{2}},$$

which yields (B.11). □

Lemma B.6 (Upper bound on the estimation error)**.** *Suppose that there exist constants $\kappa > 1, \alpha > 1$ such that for all $k \in \mathbb{N}$, $\beta_X(k) \leq \kappa \cdot k^{-\alpha}$. Suppose that $\gamma(L_f, \mathbf{p}_f, s_f) \leq n$ and $\gamma(L_g \vee L_f, \mathbf{p}_g \vee \mathbf{p}_f, s_g \vee s_f) \leq n\mathcal{E}$. Then there exists some constant $C > 0$ only depending on characteristics of X_1 and F, κ, α such that*

$$\mathbb{E}E_n \leq C \cdot \Big[\Big(\frac{\gamma(L_f, \mathbf{p}_f, s_f)}{n}\Big)^{1/2} + \Big(\frac{\gamma(L_g \vee L_f, \mathbf{p}_g \vee \mathbf{p}_f, s_g \vee s_f)}{n\mathcal{E}}\Big)^{1/2}\Big].$$

Furthermore, with probability at least $1 - 4n^{-1} - 2\big(\frac{\log(n)}{n}\big)^{\frac{\alpha-1}{2}}$,

$$E_n \leq C \cdot \Big[\Big(\frac{\gamma(L_f, \mathbf{p}_f, s_f)}{n}\Big)^{1/2} + \Big(\frac{\gamma(L_g \vee L_f, \mathbf{p}_g \vee \mathbf{p}_f, s_g \vee s_f)}{n\mathcal{E}}\Big)^{1/2} + \Big(\frac{\log(n)}{n}\Big)^{1/2}\Big].$$

Proof of Lemma B.6. With $g \in \mathcal{R}_G$, $f \in \mathcal{R}_D$, we have

$$f \circ g \in \mathcal{R} := \mathcal{R}(L_g + L_f + 1, (d_z, p_{g1}, \dots, p_{gL_g}, d, p_{f1}, \dots, p_{fL_f}, 1), s_g + s_f).$$

Note that there exists some universal constant $c > 0$ such that

$$\begin{aligned} & \gamma(L_g + L_f + 1, (d_z, p_{g1}, \dots, p_{gL_g}, d, p_{f1}, \dots, p_{fL_f}, 1), s_g + s_f) \\ \leq \; & c \cdot \gamma(L_f \vee L_g, \mathbf{p}_f \vee \mathbf{p}_g, s_f \vee s_g), \end{aligned}$$

where $x \vee y$ of vectors x, y is meant component-wise.

We now apply Lemma B.5 to both summands of E_n. The first summand reads

$$\sup_{f \in \mathcal{R}_D} \left| (\hat{\mathbb{P}}_n^X - \mathbb{P}^X) f \right|$$

with β-mixing X_i. The second summand of E_n is upper bounded by

$$\sup_{h \in \mathcal{R}} \left| (\hat{\mathbb{P}}_{n\mathcal{E}}^Z - \mathbb{P}^Z) h \right|$$

with i.i.d. Z_i, that is, β-mixing coefficients $\beta_Z(k) = \mathbb{1}_{\{k=0\}}$ $(k \geq 0)$. □

Proof of Theorem 3.4. By Proposition B.1,

$$R_n(\hat{g}_n) \leq \sqrt{d} \cdot A_n + 2 \cdot E_n.$$

Under the given assumptions on $L_f, \mathbf{p}_f, s_f$, we conclude from (B.9) (cf. also Remark 1 in [28]) that

$$\begin{aligned} & \gamma(L_f, \mathbf{p}_f, s_f) \\ \leq \; & 2(s_f + 1) \log \left(2^{L_f+3} (L_f + 1) p_0 p_{L+1} s_f^{L_f} \right) \lesssim s_f L_f \log(s_f L_f). \end{aligned}$$

Under the given assumptions on $L_g, \mathbf{p}_g, s_g$, we conclude by Lemma B.3 for N large enough that

$$A_n \lesssim \frac{N}{n\mathcal{E}} + N^{-\beta/d_g}.$$

Thus by Lemma B.6,

$$\begin{aligned} \mathbb{E} R_n(\hat{g}_n) \; \lesssim \; & \left(\frac{s_f L_f \log(s_f L_f)}{n} \right)^{1/2} + N^{-\beta/d_g} \\ & + \left(\frac{(s_f \vee s_g)(L_f \vee L_g) \log((s_f \vee s_g)(L_f \vee L_g))}{n\mathcal{E}} \right)^{1/2}. \end{aligned}$$

Choose $N = \lceil C_1 n\mathcal{E}\phi_n \rceil$, where C_1 is large enough such that $N \geq (\beta+1)^{d_g} \vee (K+1)e^{d_g}$, then

$$\mathbb{E} R_n(\hat{g}_n) \lesssim \left(\frac{s_f L_f \log(s_f L_f)}{n} \right)^{1/2} + \phi_n^{1/2} \log(n\mathcal{E})^{3/2}.$$

The large deviation statement is immediate from Lemma B.6. □

B.4. Adaptation to the conditional case

We follow the same procedure as in the unconditional case with slight adaptations. We abbreviate $\mathcal{G}^c := \mathcal{G}^c(d_Z, d_Y, d_g, \beta, K)$ and $\mathcal{R}_G := \mathcal{R}(L_g, \mathbf{p}_g, s_g)$, $\mathcal{R}_D = \mathcal{R}(L_f, \mathbf{p}_f, s_f)$ as before. Recall from (2.5) that

$$\hat{g}_n^c := \operatorname*{arg\,min}_{g \in \mathcal{R}_G} \hat{W}_{1,n}^c(g).$$

Proposition B.7 (cWGAN: Basic inequality)**.** *It holds that*

$$W_{1,n}^c(\hat{g}_n^c) - \inf_{g \in \mathcal{G}^c} W_{1,n}^c(g) \leq \sqrt{d} \cdot A_n^c + 2 \cdot E_n^c,$$

where

$$A_n^c := \sup_{g \in \mathcal{G}^c} \inf_{\tilde{g} \in \mathcal{R}_G} \|g - \tilde{g}\|_\infty, \tag{B.12}$$

$$E_n^c := \sup_{f \in \mathcal{R}_D} |(\hat{\mathbb{P}}_n^{X,Y} - \mathbb{P}^{X,Y}) f| + \sup_{g \in \mathcal{R}_G, f \in \mathcal{R}_D} \left| \frac{1}{n} \sum_{i=1}^n \{ f(g(Z_i, Y_i), Y_i) - \mathbb{E} f(g(Z_1, Y_1), Y_1) \} \right|. \tag{B.13}$$

Proof of Proposition B.7. Proceed as in the proof of proposition B.1. Note that

$$\begin{aligned} & |W_{1,n}^c(g) - W_{1,n}^c(\tilde{g})| \\ \leq & \sup_{f \in \mathcal{R}_D, \|f\|_L \leq 1} \mathbb{E} |f(g(Z,Y),Y) - f(\tilde{g}(Z,Y),Y)| \\ \leq & \sqrt{d} \|g - \tilde{g}\|_\infty, \end{aligned}$$

stays the same. □

Lemma B.8. *Let $\beta \geq 1$, $d_g \in \mathbb{N}$, $\tilde{\beta} \geq \frac{D}{d_g} \beta$. Suppose that for N large enough,*

- $L_g \geq \log_2(n) \left(2 \log_2(4 d_g \vee 4\beta) + \log_2(4D \vee 4\tilde{\beta}) \right)$,
- $\min_{i=1,\dots,L_g} p_i \gtrsim N$,
- $s_g \gtrsim N \log_2(n)$,

then A_n^c from (B.12) satisfies

$$A_n^c \lesssim \frac{N}{n} + N^{-\beta/d_g}, \tag{B.14}$$

Here, the bounding constants only depend on $\tilde{\beta}, \beta, d_g, D, d$ and K.

Proof of Lemma B.8. The proof basically follows from Theorem B.2 with $k = \lceil \log_2(n) \rceil$ along the same lines as in the proof of Theorem 1 in [28]. Let $g \in \mathcal{G}^c = \mathcal{G}^c(d_Z, d_Y, d_g, \beta, K)$. For ease of notation, let

$$\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2) := (\beta, \tilde{\beta}, \beta), \qquad \mathbf{t} = (t_0, t_1, t_2) := (d_g, D, d_g)$$

and $\mathbf{d} = (d_0, d_1, d_2, d_3) := (d_Z + d_Y, D, d_g, d)$. We furthermore abbreviate $g_0 := g_{enc,0}$, $g_1 := g_{enc,1}$ and $g_2 := g_{dec}$.

First transform the component functions g_0, g_1 as in [28, Proof of Theorem 1] to map to $[0,1]^{d_1}$ and $[0,1]^{d_2}$ respectively. Now by Theorem B.2, for $i = 0, 1, 2$ we find functions

$$\tilde{g}_i \quad \in \quad \mathcal{R}(L_i, (d_i, \mathbf{p}_i, d_{i+1}), d_{i+1}s_i),$$

where

$$\begin{aligned} L_i &= 8 + (k+5)(1 + \log_2(t_i \vee \beta_i)), \\ p_i &= (d_i, 6d_{i+1}(t_i + \lceil\beta_i\rceil)N, \ldots, 6d_{i+1}(t_i + \lceil\beta_i\rceil)N, d_{i+1}) \in \mathbb{R}^{L_i+2}, \\ s_i &\leq 141(t_i + \beta_i + 1)^{3+t_i} N(k+6) \end{aligned}$$

such that

$$\|g_i - \tilde{g}_i\|_\infty \quad \leq \quad (2K+1)(1 + t_i^2 + \beta_i^2)6^{t_i} N 2^{-k} + K 3^{\beta_i} N^{\frac{\beta_i}{t_i}}.$$

Apply $1 - (1 - \tilde{g}_{ij})_+$ for $i \in \{0, 1\}$ so that the network outputs lie in $[0,1]^{d_{i+1}}$. This does not increase the distance to g_i and adds 4 non-zero parameters per output dimension and 2 layers. The composed network $\tilde{g} := \tilde{g}_2 \circ \sigma(\tilde{g}_1) \circ \sigma(\tilde{g}_0)$ satisfies

$$\tilde{g} \in \mathcal{R}(\bar{L}, \bar{p}, \bar{s})$$

with

$$\begin{aligned} \bar{L} &:= \sum_{i=0}^{2} L_i + 6, \\ \bar{p} &:= (d_Z + d_Y, p_0, \ldots, p_0, D, p_1, \ldots, p_1, d_g, p_2, \ldots, p_2, d), \\ \bar{s} &:= \sum_{i=0}^{2} d_{i+1}(s_i + 4). \end{aligned}$$

and, in analogy to [28, Section 7.1, Lemma 3],

$$\|\tilde{g} - g\|_\infty \quad \leq \quad C \max_{i=0,1,2} \Big\{ \frac{N}{n} + N^{-\frac{\beta_i}{t_i}} \Big\} = C \Big\{ \frac{N}{n} + N^{-\frac{\beta}{d_g}} \Big\} \tag{B.15}$$

for a constant C that only depends on $\boldsymbol{\beta}, \mathbf{d}, K$. Up to now, $\tilde{g}$ may not satisfy $\|\tilde{g}\|_\infty \leq F$. However, $\tilde{g}^\circ := (\frac{\|g\|_\infty}{\|\tilde{g}\|_\infty} \wedge 1)\tilde{g}$ still fulfills $\tilde{g}^\circ \in \mathcal{R}(\bar{L}, \bar{p}, \bar{s})$ and $\|\tilde{g}^\circ\|_\infty \leq \|g\|_\infty \leq K \leq F$. Due to $\|\tilde{g}^\circ - g\|_\infty \leq 2\|\tilde{g} - g\|_\infty$, (B.15) still holds for $\tilde{g}^\circ$ with changed constants. □

We now provide an analogeous result for Lemma B.6 in the conditional case. If Z_i, $i \in \mathbb{Z}$ is a sequence of independent random variables and independent of (X_i, Y_i), $i \in \mathbb{Z}$, then β-mixing of (X_i, Y_i) implies β-mixing of (Y_i, Z_i) with the

same coefficients. The basic change is that the supremum in the second summand in E_n^c from (B.13) runs over a different class of neural networks, namely

$$\mathcal{R}_{fg} = \Big\{ h : [0,1]^{d_Z+d_Y} \to \mathbb{R},\ (z,y) \mapsto f(g(z,y),y) \Big|\ g \in \mathcal{R}_G,\ f \in \mathcal{R}_D,\ \|f\|_L \le 1 \Big\}.$$

By adding d_Y neurons in each layer of g, we can mimic the function $(g(z,y),y)$. Thus

$$\mathcal{R}_{fg} \subseteq \tilde{\mathcal{R}}^c = \mathcal{R}(L_{comp}, \mathbf{p}_{comp}, s_{comp}, \infty),$$

where

$$\begin{aligned} L_{comp} &= L_f + L_g + 1, \\ \mathbf{p}_{comp} &= (d_z + d_Y, p_{g,1} + d_Y, \dots, p_{g,L_g} + d_Y, d + d_Y, p_{f,1}, \dots, p_{f,L_f}, 1), \\ s_{comp} &:= s_g + s_f + (L_g + 1) d_Y. \end{aligned}$$

As long as $s_g \ge d_Y L_g$, $p_{g,i} \ge d_Y$ $(i = 1, ..., L_g)$, there exists a universal constant $c > 0$ such that

$$\gamma(L_{comp}, \mathbf{p}_{comp}, s_{comp}) \le c \cdot \gamma(L_f \vee L_g, \mathbf{p}_f \vee \mathbf{p}_g, s_f \vee s_g),$$

where $x \vee y$ for vectors x, y is meant component-wise. These remarks lead to the following result.

Lemma B.9 (Upper bound on the estimation error)**.** *Suppose that there exist constants $\kappa > 1, \alpha > 1$ such that for all $k \in \mathbb{N}$, $\beta_{X,Y}(k) \le \kappa \cdot k^{-\alpha}$. Suppose that $s_g \ge d_Y L_g$, $p_{g,i} \ge d_Y$ $(i = 1, ..., L_g)$ and $\gamma(L_g \vee L_f, \mathbf{p}_g \vee \mathbf{p}_f, s_g \vee s_f) \le n$. Then there exists some constant $C > 0$ only depending on characteristics of (X_1, Y_1) and F, κ, α such that*

$$\mathbb{E} E_n^c \le C \cdot \Big[\Big(\frac{\gamma(L_f, \mathbf{p}_f, s_f)}{n} \Big)^{1/2} + \Big(\frac{\gamma(L_g \vee L_f, \mathbf{p}_g \vee \mathbf{p}_f, s_g \vee s_f)}{n} \Big)^{1/2} \Big].$$

Furthermore, with probability at least $1 - 4n^{-1} - 2(\frac{\log(n)}{n})^{\frac{\alpha-1}{2}}$,

$$E_n \le C \cdot \Big[\Big(\frac{\gamma(L_f, \mathbf{p}_f, s_f)}{n} \Big)^{1/2} + \Big(\frac{\gamma(L_g \vee L_f, \mathbf{p}_g \vee \mathbf{p}_f, s_g \vee s_f)}{n} \Big)^{1/2} + \Big(\frac{\log(n)}{n} \Big)^{1/2} \Big].$$

Proof of Theorem 4.3. In order to bound the estimation error E_n^c proceed as in the proof of Theorem 3.4. By Proposition B.7,

$$R_n^c(\hat{g}_n^c) \le \sqrt{d} \cdot A_n^c + 2 \cdot E_n^c.$$

Under the given assumptions on $L_f, \mathbf{p}_f, s_f$, we conclude from (B.9) that

$$\begin{aligned} & \gamma(L_f, \mathbf{p}_f, s_f) \\ \le\ & (s_f + 1) \log \big(2^{2L_f+6} (L_f + 1) p_0^2 p_{L+1}^2 s_f^{2L_f} \big) \lesssim s_f L_f \log(s_f L_f). \end{aligned}$$

Under the given assumptions on $L_g, \mathbf{p}_g, s_g$, we conclude by Lemma B.8 for N large enough that

$$A_n \lesssim \frac{N}{n} + N^{-\beta/d_g}.$$

Thus by Lemma B.9,

$$\mathbb{E}R_n^c(\hat{g}_n^c) \lesssim \Big(\frac{s_f L_f \log(s_f L_f)}{n}\Big)^{1/2} + N^{-\beta/d_g} + \Big(\frac{(s_f \vee s_g)(L_f \vee L_g)\log((s_f \vee s_g)(L_f \vee L_g))}{n}\Big)^{1/2}.$$

Choose $N = \lceil C_1 n\phi_n \rceil$, where C_1 is large enough such that the conditions of Lemma B.8 are met. Then

$$\mathbb{E}R_n^c(\hat{g}_n^c) \lesssim \Big(\frac{s_f L_f \log(s_f L_f)}{n}\Big)^{1/2} + \phi_n^{1/2}\log(n)^{3/2}.$$

□

Appendix C: Entropy bound and large deviation bounds for absolutely regular sequences

C.1. Entropy bounds

In this section, we develop the entropy bound under absolutely regular β-mixing for deep sparse regularized ReLU networks $\mathcal{R} = \mathcal{R}(L, \mathbf{p}, s, F)$ in dependence on $L, \mathbf{p}$ and s. The basic theoretical ingredients consist of the empirical process theory invented in [12] and [10]. Recall the introduction of bracketing numbers and mixing coefficients from Section B.2.

Lemma C.1. *Let $\mathcal{F} \subset \{f : \mathbb{R}^r \to \mathbb{R}$ measurable$\}$ be any class of functions such that $\sup_{f\in\mathcal{F}} \|f\|_\infty \le F$. Let*

$$H := 1 \vee \log N_{[]}(2BF, \mathcal{F}, \|\cdot\|_\infty).$$

Suppose that there exist constants $\kappa > 1, \alpha > 1$ such that for all $k \in \mathbb{N}$, $\beta_X(k) \le \kappa \cdot k^{-\alpha}$ and $H \le n$.

Then there exist constants $K_1, K_2 > 0$ only depending on characteristics of $(X_i)_{i\in\mathbb{Z}}$ such that

$$\begin{aligned} &\mathbb{E}^* \sup_{f\in\mathcal{F}} |(\hat{\mathbb{P}}_n^X - \mathbb{P}^X)f| \\ \le\ & K_1 \cdot n^{-1/2} \cdot J_{[]}(F, \mathcal{F}, \|\cdot\|_\infty) + K_2\, F \cdot \Big(\frac{H}{n}\Big)^{\frac{\alpha}{\alpha+1}} =: r_n, \end{aligned} \tag{C.1}$$

where $\mathbb{E}^*$ denotes the outer expectation.

Proof. Let $\delta > 0$ arbitrary. From [10] (Remark 3.7 and the procedure in section 4.3 therein) yield that for any class $\mathcal{F}$ with $\sup_{f\in\mathcal{F}} \|f\|_{2,\beta} \le \delta$, $\sup_{f\in\mathcal{F}} \|f\|_\infty \le F$, we have with some constant $K' > 0$ only depending on characteristics of $(X_i)_{i\in\mathbb{Z}}$ that

$$\mathbb{E}^* \sup_{f\in\mathcal{F}} |(\hat{\mathbb{P}}_n^X - \mathbb{P}^X)f| \le K' \cdot n^{-1/2} \cdot J_{[]}(\delta, \mathcal{F}, \|\cdot\|_{2,\beta}) + 2F\mathbb{1}_{2F>M(n,\delta)} + 2\,R(n,\delta), \tag{C.2}$$

where with $q^*(x) := \min\{q \in \mathbb{N} : \ \beta_X(q) \leq qx\}$ and some universal constant $c > 0$,

$$\begin{aligned} R(n,\delta) &:= 2cF\frac{H(\delta)\cdot q^*\left(\frac{H}{n}\right)}{n}, \\ M(n,\delta) &:= \frac{16B^{1/2}\delta}{q^*(H(\delta)/n)}\sqrt{\frac{n}{H(\delta)}}, \\ H(\delta) &:= 1 \vee \log N_{[]}(2B^{1/2}\delta, \mathcal{F}, \|\cdot\|_{2,\beta}). \end{aligned}$$

By (B.8), choosing $\delta = B^{1/2}\cdot F$ yields $H(\delta) \leq H$ and $J_{[]}(\delta, \mathcal{F}, \|\cdot\|_{2,\beta}) \leq B^{1/2}J_{[]}(F, \mathcal{F}, \|\cdot\|_\infty)$.

For $x \leq 1$, we have

$$\begin{aligned} q^*(x) = \min\{q \in \mathbb{N} : \beta(q) \leq qx\} &\leq \min\{q \in \mathbb{N} : \kappa q^{-\alpha} \leq qx\} \\ &= \min\{q \in \mathbb{N} : \kappa x^{-1} \leq q^{\alpha+1}\} = \lceil \kappa^{\frac{1}{\alpha+1}} x^{-\frac{1}{\alpha+1}} \rceil \\ &\leq 2\kappa^{\frac{1}{\alpha+1}} x^{-\frac{1}{\alpha+1}}. \end{aligned} \tag{C.3}$$

Due to $H \leq n$, this implies

$$R(n,\delta) \leq 2cF\cdot\frac{H}{n}\cdot q^*\left(\frac{H}{n}\right) \leq 4cF\kappa^{\frac{1}{\alpha+1}}\left(\frac{H}{n}\right)^{\frac{\alpha}{\alpha+1}}. \tag{C.4}$$

In the same way we get that

$$M(n,\delta) = 16BF\left[\sqrt{\frac{H}{n}}\cdot q^*\left(\frac{H}{n}\right)\right]^{-1} \geq 8BF\delta\kappa^{-\frac{1}{\alpha+1}}\left(\frac{H}{n}\right)^{\frac{1}{\alpha+1}-\frac{1}{2}}.$$

For any $p > 0$, we have

$$2F\mathbb{1}_{2F>M(n,\delta)} \leq (2F)^{1+p}M(n,\delta)^{-p} \leq (2F)^{1+p}\cdot(8BF)^{-p}\kappa^{\frac{p}{\alpha+1}}\left(\frac{H}{n}\right)^{-\frac{p(\alpha-1)}{2(\alpha+1)}}.$$

Choosing $p = \frac{2\alpha}{\alpha-1}$ yields

$$2F\mathbb{1}_{2F>M(n,\delta)} \leq (2F)^{1+p}\cdot(8BF)^{-p}\kappa^{\frac{p}{\alpha+1}}\left(\frac{H}{n}\right)^{-\frac{\alpha}{\alpha+1}}. \tag{C.5}$$

Insertion of (C.4) and (C.5) into (C.2) yields the result.

□

C.2. Large deviation bounds

The essential techniques we use to derive large deviations bounds under absolutely regular β-mixing are coupling (cf. [5, 10]), a Talagrand-type concentration inequality by [18] and a covariance bound by Rio [27]. For completeness, we cite the results our derivations are based on.

Lemma C.2 (Coupling lemma, [10]). *Let X and Y be two random variables taking values in the Borel spaces $\mathcal{X}_1$ and $\mathcal{X}_2$ respectively, and let U be a random variable with uniform distribution on $[0,1]$, independent of (X,Y). There exists a random variable $Y^* = h(X,Y,U)$, where h is a measurable function from $\mathcal{X}_1 \times \mathcal{X}_2 \times [0,1]$ into $\mathcal{X}_2$, such that:*

(i) Y^* is independent of X and has the same distribution as Y.
(ii) $\mathbb{P}(Y \neq Y^*) = \beta(\sigma(X), \sigma(Y))$.

Theorem C.3 (Talagrand-type concentration inequality, [18], Theorem 1.1). *Assume the $W_i, i \in \mathbb{N}$ are independent random variables with values in $\mathbb{R}^r$. Let $\mathcal{F} \subset \{f : \mathbb{R}^r \to \mathbb{R}$ meausurable$\}$ be a countable set of functions with $\mathbb{E}f(W_1) = 0$, $\sup_{f\in\mathcal{F}} \|f(W_1)\|_2^2 < \infty$ and $\sup_{f\in\mathcal{F}} \|f\|_\infty \leq 1$. Define*

$$Z = \sup_{f\in\mathcal{F}} \Big| \sum_{i=1}^{m} f(W_i) \Big|.$$

Let σ be a positive real number such that $\sigma^2 \geq \sup_{f\in\mathcal{F}} \mathrm{Var}\|f(W_1)\|_2^2$. Then for all $x > 0$, it holds that

$$\mathbb{P}\left(Z \geq \mathbb{E}Z + \big(2x(m\sigma^2 + 2\mathbb{E}Z)\big)^{1/2} + \frac{x}{3} \right) \leq \exp(-x).$$

The following lemma is a direct consequence of Corollary 1.4 and Remark 1.6 in [27].

Lemma C.4 (Variance bound for β-mixing sequences). *Let $(X_i)_{i\in\mathbb{N}}$ be a strictly stationary sequence of random variables with values in a Polish space $\mathcal{X}$. Let $q \in \mathbb{N}$. Then for any $f : \mathcal{X} \to \mathbb{R}$ with $\|f\|_{2,\beta} < \infty$,*

$$\Big\| \sum_{i=1}^{q} f(X_i) \Big\|_2^2 \leq 4q \|f\|_{2,\beta}^2.$$

Lemma C.5. *Let $\mathcal{F} \subset \{f : \mathbb{R}^r \to \mathbb{R}$ measurable$\}$ be any countable class of functions such that $\mathbb{E}f(X_i) = 0$ for all $f \in \mathcal{F}$ and $\sup_{f\in\mathcal{F}} \|f\|_\infty \leq F$.*

Suppose that there exist constants $\kappa > 1, \alpha > 1$ such that for all $k \in \mathbb{N}$, $\beta_X(k) \leq \kappa \cdot k^{-\alpha}$. Define r_n as in (C.1).

Then for all $x > 0$ and $q \in \mathbb{N}$,

$$\mathbb{P}\Big(\sup_{f\in\mathcal{F}} \Big| \sum_{i=1}^{n} f(X_i) \Big| \geq 4nr_n + 8B^{1/2}Fn^{1/2}x^{1/2} + 4Fqx \Big) \leq 2\exp(-x) + \frac{n\beta_X(q)}{qx}. \tag{C.6}$$

Especially, for any $z \leq 1$,

$$\mathbb{P}\Big(\sup_{f\in\mathcal{F}} \Big| \sum_{i=1}^{n} f(X_i) \Big| \geq 4nr_n + 8B^{1/2}Fn^{1/2}x^{1/2} + 8F\kappa^{\frac{1}{\alpha+1}} z^{-\frac{1}{\alpha+1}} x \Big) \leq 2\exp(-x) + \frac{nz}{x}. \tag{C.7}$$

Proof of Lemma C.5. Let $q \in \mathbb{N}$. Starting from Lemma C.2 we construct by induction a sequence of random variables $(X_i^0)_{i>0}$ such that:

1. For any $i \geq 0$, the random variable $U_i^0 := (X_{iq+1}^0, \ldots, X_{iq+q}^0)$ has the same distribution as $U_i := (X_{iq+1}, \ldots, X_{iq+q})$.
2. The sequence $(U_{2i}^0)_{i\geq 0}$ is i.i.d. and so is $(U_{2i+1}^0)_{i\geq 0}$.
3. For any $i \geq 0$, $\mathbb{P}(U_i \neq U_i^0) \leq \beta(q)$.

Define $W_i(f) := \sum_{j=(i-1)q+1}^{iq\wedge n} f(X_i^0)$. From the above coupling we obtain the following decomposition:

$$\sum_{i=1}^{n} f(X_i) = \sum_{i=1, i \text{ odd}}^{\lceil \frac{n}{q} \rceil} W_i(f) + \sum_{i=1, i \text{ even}}^{\lceil \frac{n}{q} \rceil} W_i(f) + \sum_{i=1}^{n} \{f(X_i) - f(X_i^0)\}.$$

We conclude that

$$\sup_{f\in\mathcal{F}} \Big| \sum_{i=1}^{n} f(X_i) \Big| \leq Fq \cdot (Z_1 + Z_2) + A, \tag{C.8}$$

where

$$Z_1 := \sup_{f\in\mathcal{F}} \Big| \sum_{i=1, i \text{ odd}}^{\lceil \frac{n}{q} \rceil} \frac{W_i(f)}{Fq} \Big|, \qquad Z_2 := \sup_{f\in\mathcal{F}} \Big| \sum_{i=1, i \text{ even}}^{\lceil \frac{n}{q} \rceil} \frac{W_i(f)}{Fq} \Big|, \qquad A := F \sum_{i=1}^{n} \mathbb{1}_{X_i \neq X_i^0}.$$

Note that $W_{2k}(f)$, $W_{2k+1}(f)$, $k \geq 0$ are independent by construction. Furthermore, $\sup_{f\in\mathcal{F}} |\frac{Y_i(f)}{Fq}| \leq 1$ and by Lemma C.4 and (B.8),

$$\|W_i(f)\|_2^2 \leq 4q \|f\|_{2,\beta}^2 \leq 4qB\|f\|_\infty^2 \leq 4qBF^2$$

and thus $\|\frac{W_i(f)}{Fq}\|_2^2 \leq \frac{4B}{q}$. By Theorem C.3 applied with $\sigma^2 = \frac{4B}{q}$, we have

$$\mathbb{P}\Big(Z_1 \geq \mathbb{E}Z_1 + (2x(\frac{4Bn}{q^2} + 2\mathbb{E}Z_1))^{1/2} + \frac{x}{3} \Big) \leq \exp(-x).$$

Using the simple bound $2ab \leq a^2 + b^2$, we have

$$(2x(\frac{4Bn}{q^2} + 2\mathbb{E}Z_1))^{1/2} \leq \frac{(8Bxn)^{1/2}}{q} + 2(x\mathbb{E}Z_1)^{1/2} \leq \frac{(8Bxn)^{1/2}}{q} + x + \mathbb{E}Z_1.$$

This yields

$$\mathbb{P}\Big(Z_1 \geq 2\mathbb{E}Z_1 + (\frac{8Bxn}{q^2})^{1/2} + \frac{4x}{3} \Big) \leq \exp(-x). \tag{C.9}$$

Let $I_1 := \bigcup_{i=1, i \text{ odd}}^{\lceil \frac{n}{q} \rceil} \{(i-1)q+1, ..., iq\wedge n\}$. Then $Z_1 = \frac{1}{Fq} \sup_{f\in\mathcal{F}} \big| \sum_{i\in I_1} f(X_i^0) \big|$. Note that by construction, X_i^0 is still β-mixing with coefficients upper bounded by β_X, and I_1 has less than or equal n summands. It is therefore easily seen

that the upper bounds in [10] which were used in the proof of Lemma C.1 stay the same. We obtain from (C.1) that

$$Fq \cdot \mathbb{E}Z_1 \leq nr_n. \tag{C.10}$$

Similar results as given in (C.9) and (C.10) also hold for Z_2.

Finally, we have by Markov's inequality that

$$\mathbb{P}(A > x) \leq \frac{\|A\|_1}{x} \leq \frac{nF\mathbb{P}(X_i \neq X_i^0)}{x} \leq \frac{nF\beta_X(q)}{x}. \tag{C.11}$$

Insertion of (C.9), (C.10) and (C.11) into (C.8) yields

$$\begin{aligned}
&\mathbb{P}\Big(\sup_{f\in\mathcal{F}}\Big|\sum_{i=1}^{n} f(X_i)\Big| \geq 2\cdot\big(2nr_n + (8BF^2nx)^{1/2} + \frac{4Fqx}{3}\big) + Fqx\Big) \\
\leq\ & \mathbb{P}(Z_1 \geq 2\mathbb{E}Z_1 + (\frac{8Bxn}{q^2})^{1/2} + \frac{4x}{3}) + \mathbb{P}(Z_2 \geq 2\mathbb{E}Z_2 + (\frac{8Bxn}{q^2})^{1/2} + \frac{4x}{3}) \\
&\quad + \mathbb{P}(A \geq Fqx) \\
\leq\ & 2\exp(-x) + \frac{n\beta_X(q)}{qx},
\end{aligned}$$

which concludes the proof of (C.6). (C.7) follows from the upper bound in (C.3) and the choice $q = q^*(z)$. □