Statistical Analysis of Wasserstein GANs with Applications to Time-Series Forecasting

(arXiv: 2011.03074, with Stefan Richter)

Moritz Haas

IMPRS-IS 2021 Presentation

January 27^{th}, 2021

Learn to sample from $\mathbb{P}^Y$

Figure: StyleGAN 2 [Kar+19], www.thispersondoesnotexist.com

Unconditional Problem

Goal: Learn to sample from unknown $\mathbb{P}^Y$.

Given $Y_i \sim \mathbb{P}^Y$, $i = 1, \dots, n$ strictly stationary with values in $[0,1]^d$.

Sample i.i.d. latent noise $Z \in [0,1]^{d_Z}$ ($\mathbb{P}^Z$ known) independent of $Y_1, \dots, Y_n$.

Find a generator function $g : [0,1]^{d_Z} \to [0,1]^d$ such that

$$\mathbb{P}^{g(Z)} = \mathbb{P}^Y.$$

Conditional Problem

Goal: Learn to sample from unknown $\mathbb{P}^{Y|X=x}$ given conditional information $X = x$.

Given $(X_i, Y_i) \sim \mathbb{P}^{(X,Y)}$, $i = 1, \ldots, n$ strictly stationary with values in $[0,1]^{d_X+d}$.

Sample i.i.d. latent noise $Z \in [0,1]^{d_Z}$ ($\mathbb{P}^Z$ known) independent of $Y_1, \ldots, Y_n, X_1, \ldots, X_n$.

Find a generator function $g : [0,1]^{d_Z+d_X} \to [0,1]^d$ such that

$$\mathbb{P}^{X,\, g(Z,X)} = \mathbb{P}^{X,Y}.$$

$$\rightsquigarrow \qquad \mathbb{P}^{g(Z,x)} = \mathbb{P}^{g(Z,X)|X=x} = \mathbb{P}^{Y|X=x}.$$

Example: Temperature Data in German Cities

Learn conditional distribution of temperatures in 32 German cities given temperatures on previous day.

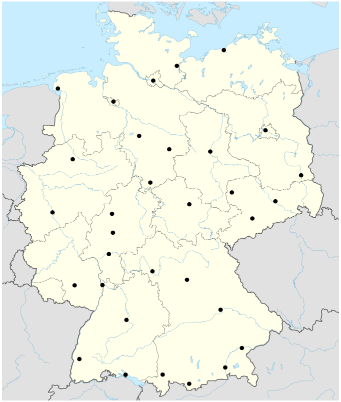
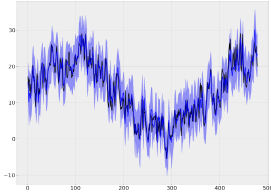
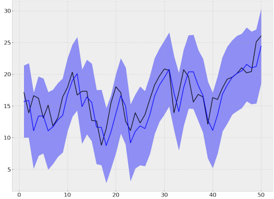


Figure: Dataset from Deutscher Wetterdienst [PR20].



(a) 478 days



(b) 50 days

1-Wasserstein Objective

Dual formulation [Vil08] of W_1-distance with critic functions f:

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

Approximation with critic networks f:

Modified network-based Wasserstein Distance

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

1-Wasserstein Objective

Dual formulation [Vil08] of W_1-distance with critic functions f:

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

Approximation with critic networks f:

Modified network-based Wasserstein Distance

$$W_{1,n}(g) := \sup_{f \in \mathcal{R}_D, \|f\|_L \leq 1} \left\{ \mathbb{E} f(Y) - \mathbb{E} f(g(Z)) \right\}.$$

1-Wasserstein Objective

Dual formulation [Vil08] of W_1-distance with critic functions f:

$$W_1(\mathbb{P}_1, \mathbb{P}_2) = \sup_{f:\mathbb{R}^{d+d_Y} \to \mathbb{R},\, ||f||_L \leq 1} \int_{\mathcal{X}} f \, \mathrm{d}\mathbb{P}_1 - \int_{\mathcal{X}} f \, \mathrm{d}\mathbb{P}_2.$$

Approximation with critic networks f:

Modified network-based Wasserstein Distance

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

The cWGAN Estimator

Empirical Risk Minimizer

$$\hat{g}_n := \arg\min_{g \in \mathcal{R}_G} \hat{W}_{1,n}(g)$$

with

$$\hat{W}_{1,n}(g) := \sup_{f \in \mathcal{R}_D, \|f\|_L \leq 1} \left\{ \frac{1}{n} \sum_{i=1}^{n} f(Y_i) - \sum_{j=1}^{n\mathcal{E}} f(g(Z_j)) \right\}$$

$\mathcal{E} \propto$ number of epochs (only for the unconditional case)

ReLU Networks

For $v, x \in R^p$, $p \in \mathbb{N}$, define ReLU function $\sigma : \mathbb{R}^d \to \mathbb{R}^d$,

$$\sigma_v(x) = \max(x - v, 0),$$

where max component-wise.
$L \in \mathbb{N}$ number of hidden layers, $\mathrm{p} = (p_0, ..., p_{L+1}) \in \mathbb{N}^{L+2}$ width vector.

$\mathcal{R}(L, p)$ ReLU networks with architecture (L, p)

$$h : \mathbb{R}^{p_0} \to \mathbb{R}^{p_{L+1}},$$

$$h(x) = W^{(L)}\sigma_{v^{(L)}}(W^{(L-1)}\sigma_{v^{(L-1)}}(\ldots W^{(1)}\sigma_{v^{(1)}}(W^{(0)}x)\ldots)),$$

where $W^{(l)} \in \mathbb{R}^{p_l \times p_{l+1}}$ weight matrices and $v^{(l)} \in \mathbb{R}^{p_l}$ bias vectors.

Sparse bounded ReLU Networks

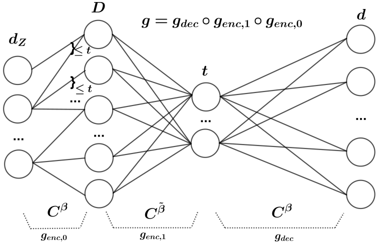
$$\mathcal{R}(L, \mathrm{p}, s) := \Big\{ h \in \mathcal{R}(L, \mathrm{p}) \;\Big|\; \max_{j=0,\dots,L} \|W_j\|_\infty \vee |v_j|_\infty \leq 1,$$
$$\sum_{j=0}^{L} \|W_j\|_0 + |v_j|_0 \leq s \text{ and } \|\, |h|_\infty \|_{L^\infty([0,1]^{p_0})} \leq F \Big\}.$$

J. Schmidt-Hieber. *Nonparametric regression using deep neural networks with ReLU activation function.* *Annals of Statistics, 2020.*

Main Result: Assumptions

Class of generator functions $\mathcal{G}$:
Compositions of t-sparse, β-Hölder smooth functions.
Assume

$$\exists g^* \in \mathcal{G} : \mathbb{P}^{X,g^*(Z,X)} = \mathbb{P}^{X,Y}.$$



Network Growth Assumptions: With the rate $\phi_{n\mathcal{E}} := (n\mathcal{E})^{-\frac{2\beta}{2\beta+t}}$,

(a) $L_g \asymp \log(n\mathcal{E})$,

(b) $\min_{i=1,\dots,L_g} p_{g,i} \asymp (n\mathcal{E}) \cdot \phi_{n\mathcal{E}}$,

(c) $s_g \asymp (n\mathcal{E}) \cdot \phi_{n\mathcal{E}} \log(n\mathcal{E})$,

(d) $(L_f \lesssim L_g,\ s_f \lesssim s_g)$ or $(L_g \lesssim L_f,\ s_g \lesssim s_f)$.

Main Result: Conditional Excess Risk Bound

Theorem 1 (Convergence rate for the conditional excess risk)

Suppose $F \geq K \vee 1$ and assumptions (a)-(d) hold.
If $\exists\, \kappa > 1, \alpha > 1 : \beta_X(k) \leq \kappa \cdot k^{-\alpha}$ for all $k \in \mathbb{N}$, then

$$\mathbb{E} W_{1,n}(\hat{g}_n) \lesssim \left(\frac{s_f L_f \log(s_f L_f)}{n} \right)^{1/2} + \sqrt{d}\, \phi_{n\mathcal{E}}^{1/2} \log(n\mathcal{E})^{3/2}.$$

Furthermore, with probability $\geq 1 - 3n^{-1} - (\frac{\log(n)}{n})^{\frac{\alpha-1}{2}}$,

$$W_{1,n}(\hat{g}_n) \lesssim \left(\frac{s_f L_f \log(s_f L_f)}{n} \right)^{1/2} + \sqrt{d}\, \phi_{n\mathcal{E}}^{1/2} \log(n\mathcal{E})^{3/2} + \left(\frac{\log(n)}{n} \right)^{1/2},$$

where $\lesssim$ dep. on characteristics of (X_1, Y_1), κ, α and hyperparameters of $\mathcal{G}$ but not on d.

Main Result: Conditional Excess Risk Bound

Theorem 1 (Convergence rate for the conditional excess risk)

Suppose $F \geq K \vee 1$ and assumptions (a)-(d) hold.
If $\exists\, \kappa > 1, \alpha > 1 : \beta_X(k) \leq \kappa \cdot k^{-\alpha}$ for all $k \in \mathbb{N}$, then

$$\mathbb{E} W_{1,n}(\hat{g}_n) \lesssim \left(\frac{s_f L_f \log(s_f L_f)}{n}\right)^{1/2} + \sqrt{d}\, \phi_{n\mathcal{E}}^{1/2} \log(n\mathcal{E})^{3/2}.$$

Furthermore, with probability $\geq 1 - 3n^{-1} - \left(\frac{\log(n)}{n}\right)^{\frac{\alpha-1}{2}}$,

$$W_{1,n}(\hat{g}_n) \lesssim \left(\frac{s_f L_f \log(s_f L_f)}{n}\right)^{1/2} + \sqrt{d}\, \phi_{n\mathcal{E}}^{1/2} \log(n\mathcal{E})^{3/2} + \left(\frac{\log(n)}{n}\right)^{1/2},$$

where $\lesssim$ dep. on characteristics of (X_1, Y_1), κ, α and hyperparameters of $\mathcal{G}$ but not on d.

Is $W_{1,n}$ a meaningful distance measure?

Main Theorem: $\mathbb{E} W_{1,n}(\hat{g}_n) \to 0, \quad n \to \infty.$

Lemma 2 (Characterization of weak convergence)

Let $a_n = n^{-\frac{2\gamma}{2\gamma+d+d_X}}$ for some $\gamma \geq 1$, and suppose that

(e) $F \geq 1$,

(f) $L_f \gtrsim \log_2(n)$,

(g) $\min_{i=1,\dots,L} p_{f,i} \gtrsim n a_n$,

(h) $s_f \gtrsim \log(n) n a_n$,

where $\gtrsim$ dep. on γ, d. Then, for r.v. $X_n, X \in \mathcal{P}([0,1]^d)$, $n \in \mathbb{N}$ the following convergence statements for $n \to \infty$ are equivalent:

(i) $X_n \xrightarrow{d} X$, (ii) $W_1(\mathbb{P}^{X_n}, \mathbb{P}^X) \to 0$, (iii) $W_{1,n}(\mathbb{P}^{X_n}, \mathbb{P}^X) \to 0$.

Lemma 3 (Convergence of the estimator)

Let assumptions (e)-(h) hold with some $\gamma \geq 1$. Let $(\tilde{g}_n)_{n \in \mathbb{N}}$ be a sequence of r.v. with $\mathbb{E} W_{1,n}(\tilde{g}_n) \to 0$. Then

$$(X, \tilde{g}_n(Z, X)) \xrightarrow{d} (X, Y).$$

Synthetic Data

$$U([0,1]^{10}) \sim (Z_i, X_i) \xrightarrow{h} \mathbb{R}^3 \xrightarrow{g^*} Y_i \in \mathbb{R}^{10}$$

Measured quantity	Number of samples n				
	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: Coverage prob. (in %) for $I_{n,N}$ with $\alpha = 0.05$, $T(x) = \sum_{j=1}^{10} x_j$ and $W_1(\hat{\mathbb{P}}_N^{X,Y}, \hat{\mathbb{P}}_N^{X,\hat{g}_n(Z,X)})$, where $N = 1000$. Train 5 models for 700 epochs.

Conclusion

- formalize Wasserstein GANs theoretically (with growing network architectures unlike [BST20]),
- $W_{1,n}$ characterizes weak convergence,
- first convergence rates for (conditional) WGANs,

  $\rightsquigarrow$ recommendations on network sizes,
- allow dependence (β- and ϕ-mixing),
- construct asymptotic confidence intervals for high-dim. time series forecasting,

  $\rightsquigarrow$ simulation studies show good empirical coverage,
- explains good performance under long training for large generators and/or large dimension d.

Research interests

Rather visionary:

- Learning meaningful representations, Disentanglement, Causality
- Transfer Learning, RL, Un-/Self-supervised Learning,
- Learning from Biology: Attention, Recurrence, Modularity, Graph structures, ...

Rather solid:

- Understanding implicit assumptions and biases,
- provable guarantees and failures,
- ...

References I

M. Arjovsky, S. Chintala, and L. Bottou. *Wasserstein GAN*. 2017. arXiv: 1701.07875.

C. Anil, J. Lucas, and R. Grosse. *Sorting out Lipschitz function approximation*. 2018. arXiv: 1811.05381.

H. C. P. Berbee. "Random Walks with Stationary Increments and Renewal Theory. Mathematisch Centrum, Amsterdam". In: *Mathematical Centre Tracts* 112 (1979). DOI: 10.1002/bimj.4710240208.

Gérard Biau, Maxime Sangnier, and Ugo Tanielian. *Some Theoretical Insights into Wasserstein GANs*. 2020. arXiv: 2006.02682.

References II

S. Dedecker and S. Louhichi. "Maximal Inequalities and Empirical Central Limit Theorems". In: *Empirical Process Techniques for Dependent Data* (2002). Ed. by Mikosch Dehling and Sorensen, pp. 137–159.

P. Doukhan, P. Massart, and E. Rio. "Invariance principles for absolutely regular empirical processes". In: *Annales de l'I.H.P. Probabilités et statistiques* 31.2 (1995), pp. 393–427.

J. L. Doob. "Regularity Properties of Certain Families of Chance Variables". In: *Transactions of the American Mathematical Society* 47.3 (1940), pp. 455–486. ISSN: 00029947. URL: http://www.jstor.org/stable/1989964.

I. J. Goodfellow et al. *Generative Adversarial Networks*. 2014. arXiv: 1406.2661.

References III

I. Gulrajani et al. *Improved Training of Wasserstein GANs*. 2017. arXiv: 1704.00028.

Moritz Haas and Stefan Richter. *Statistical analysis of Wasserstein GANs with applications to time series forecasting*. 2020. arXiv: 2011.03074.

Tero Karras et al. *Analyzing and Improving the Image Quality of StyleGAN*. 2019. arXiv: 1912.04958.

T. Klein and E. Rio. "Concentration around the mean for maxima of empirical processes". In: *Ann. Probab.* 33.3 (May 2005), pp. 1060–1077. DOI: 10.1214/009117905000000044.

X. Mao et al. *Least Squares Generative Adversarial Networks*. 2016. arXiv: 1611.04076.

References IV

Nathawut Phandoidaen and Stefan Richter. *Forecasting time series with encoder-decoder neural networks*. 2020. arXiv: 2009.08848.

Emmanuel Rio. "Inégalités de Hoeffding pour les fonctions lipschitziennes de suites dépendantes". In: *Comptes Rendus de l'Académie des Sciences - Series I - Mathematics* 330.10 (2000), pp. 905–908. ISSN: 0764-4442. DOI: https://doi.org/10.1016/S0764-4442(00)00290-1.

Alec Radford, Luke Metz, and Soumith Chintala. *Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks*. 2016. arXiv: 1511.06434.

J. Schmidt-Hieber. *Nonparametric regression using deep neural networks with ReLU activation function*. 2017. arXiv: 1708.06633.

References V

C. Villani. "Optimal transport – Old and new". In: vol. 338. Jan. 2008, pp. 43–113. DOI: 10.1007/978-3-540-71050-9.

Generator Class $\mathcal{G}$

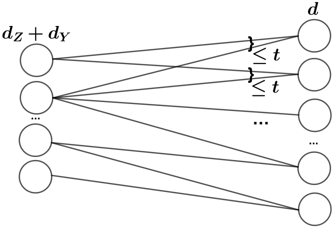
From now on assume $\exists\, g^* \in \mathcal{G}$: $\mathbb{P}^{g^*(Z)} = \mathbb{P}^Y$.

First define β-Hölder smooth $f : T \subset \mathbb{R}^t \to \mathbb{R}$ with $\beta \in \mathbb{N}$, $K > 0$:

$$C_t^\beta(T,K) := \Big\{ f : T \to \mathbb{R} \Big| \sum_{\alpha:\, 0 \le |\alpha| < \beta} \|\partial^\alpha f\|_\infty + \sum_{\alpha:\, |\alpha| = \beta - 1} \sup_{x \neq y} \frac{|\partial^\alpha f(x) - \partial^\alpha f(y)|}{|x-y|_\infty} \le K \Big\}.$$

Now

$$\mathcal{G} := \Big\{ g : [0,1]^{d_Z + d_X} \to [0,1]^d \Big| g_j \in C_t^\beta([0,1]^t, K) \quad \forall j \in \{1,\dots,d\} \Big\}$$



General Generator Class

Generator class $\mathcal{G}(q, \boldsymbol{d}, \boldsymbol{t}, \boldsymbol{\beta}, K)$:

$$g = g_q \circ \cdots \circ g_1 \circ g_0,$$

where $g_i : \mathbb{R}^{d_i} \to \mathbb{R}^{d_{i+1}}$ and $g_{ij} \in C_{t_i}^{\beta_i}([-K, K]^{t_i}, K)$ for all i, j.

$\to$ Compositions of sparse Hölder smooth functions

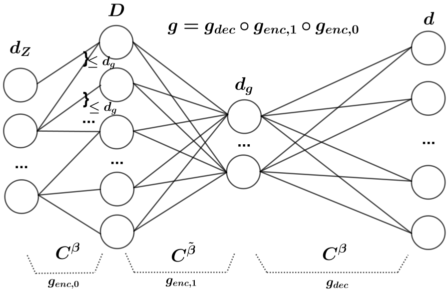


Figure: Possible Encoder-Decoder-Structure of g.

Bounding Constants in the Risk Bound

Size of $\mathcal{G}$	Error amplification	Constraints
$\exists i \in \{1,\ldots,q\} : d_i = d$	$(d \log d)^{1/2}$	-
$\exists i \in \{0,\ldots,q\} : t_i = d$	$(d^2 + \beta_i^2)6^d$	$n\mathcal{E} \gtrsim (\beta_i + 1)^{\frac{d}{2\frac{\beta_i}{d}+1}}$

Therefore:

- only applicable for low intrinsic dimensionalities t_i,
- mitigate through longer training $\mathcal{E} \to \infty$.

Future Work

- other function/network classes (e.g. Groupsort [ALG18]),
- local minima and estimators obtained by SGD,
- include gradient penalty in theory,
- refine approximation results from [Sch17] for more insight into properties of $W_{1,n}$ and good generator architectures,
- minimax rate for excess risk,
- rate of the weak convergence,
- understand double descent...

Proof: Bound on the Estimation Error

Use entropy [DL02; DMR95] and large deviation bounds [KR05] for β-mixing seq. on:

$$
\begin{aligned}
e_n \leq \cdots \leq 2 \sup_{g \in \mathcal{R}_G} \left| \hat{W}_{1,n}(g) - W_{1,n}(g) \right| \\
\leq 2 \sup_{f \in \mathcal{R}_D} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E} f(X) \right| \\
+ 2 \sup_{g \in \mathcal{R}_G, f \in \mathcal{R}_D} \left| \frac{1}{n\mathcal{E}} \sum_{j=1}^{n\mathcal{E}} f(g(Z_j)) - \mathbb{E} f(g(Z)) \right| .
\end{aligned}
$$

Proof: Empirical Process Theory

$N_{[]}(\delta, \mathcal{F}, \|\cdot\|_\infty)$ bracketing number.

Derived from [DL02]

Let $\mathcal{F} \subset \{f : \mathbb{R}^r \to \mathbb{R} \text{ measurable}\}$ with $\sup_{f\in\mathcal{F}} \|f\|_\infty \leq F$. Let

$$H := 1 \vee \log N_{[]}(2F \sum_{k=0}^{\infty} \beta_X(k), \mathcal{F}, \|\cdot\|_\infty).$$

If $\exists\, \kappa > 1, \alpha > 1 : \beta_X(k) \leq \kappa \cdot k^{-\alpha}$ for all $k \in \mathbb{N}$ and $H \leq n$, then there exist $C_1, C_2 > 0$ dep. on characteristics of $(X_i)_{i\in\mathbb{Z}}$ such that

$$\begin{aligned}\mathbb{E}^* \sup_{f\in\mathcal{F}} |(\hat{\mathbb{P}}_n^X - \mathbb{P}^X)f| \leq \quad & C_1 \cdot n^{-1/2} \cdot \int_0^F \sqrt{1 + \log N_{[]}(\varepsilon, \mathcal{F}, \|\cdot\|_\infty)}\, \mathrm{d}\varepsilon \\ & + C_2\, F \cdot \left(\frac{H}{n}\right)^{\frac{\alpha}{\alpha+1}}.\end{aligned}$$

Proof: Large Deviation Bounds

For i.i.d. and ϕ-mixing seq.: McDiarmid's inequality [Doo40; Rio00]

Coupling [Ber79; DL02]

Let $q \in \mathbb{N}$. Construct a sequence of r.v. $(X_i^0)_{i>0}$ such that:

(1) $U_i^0 := (X_{iq+1}^0, \dots, X_{iq+q}^0) \stackrel{d}{=} (X_{iq+1}, \dots, X_{iq+q}) =: U_i \quad \forall i \geq 0$.

(2) $(U_{2i}^0)_{i \geq 0}$ is i.i.d. and so is $(U_{2i+1}^0)_{i \geq 0}$.

(3) $\mathbb{P}(U_i \neq U_i^0) \leq \beta(q) \quad \forall i \geq 0$.

- Replace X_i by X_i^0, ((3) and Markov ineq.)
- Utilize averaging $\sum f(X_i^0) = \sum \tilde{f}(U_i^0)$ with $\tilde{f}(u) = \sum_{j=1}^q f(u_j)$
  $\rightsquigarrow$ Talagrand-type inequality [KR05] includes variance bound:

$$\mathbb{P}(Z \geq \mathbb{E}Z + \varepsilon_{n,\sigma^2}(x)) \leq \exp(-x) \stackrel{!}{=} n^{-b}$$

$$\overset{x = b \ln(n)}{\rightsquigarrow} \quad Z \leq \mathbb{E}Z + \varepsilon_{n,\sigma^2}(b \ln n) \text{ with prob. } \geq 1 - n^{-b}.$$

Proof: Large Deviation Bounds

For i.i.d. and ϕ-mixing seq.: McDiarmid's inequality [Doo40; Rio00]

Coupling [Ber79; DL02]

Let $q \in \mathbb{N}$. Construct a sequence of r.v. $(X_i^0)_{i>0}$ such that:

(1) $U_i^0 := (X_{iq+1}^0, \ldots, X_{iq+q}^0) \overset{d}{=} (X_{iq+1}, \ldots, X_{iq+q}) =: U_i \quad \forall i \geq 0.$

(2) $(U_{2i}^0)_{i\geq 0}$ is i.i.d. and so is $(U_{2i+1}^0)_{i\geq 0}$.

(3) $\mathbb{P}(U_i \neq U_i^0) \leq \beta(q) \quad \forall i \geq 0.$

- Replace X_i by X_i^0, ((3) and Markov ineq.)
- Utilize averaging $\sum f(X_i^0) = \sum \tilde{f}(U_i^0)$ with $\tilde{f}(u) = \sum_{j=1}^q f(u_j)$ $\rightsquigarrow$ Talagrand-type inequality [KR05] includes variance bound:

$$\mathbb{P}(Z \geq \mathbb{E}Z + \varepsilon_{n,\sigma^2}(x)) \leq \exp(-x) \overset{!}{=} n^{-b}$$

$$\overset{x=b\ln(n)}{\rightsquigarrow} \quad Z \leq \mathbb{E}Z + \varepsilon_{n,\sigma^2}(b \ln n) \text{ with prob. } \geq 1 - n^{-b}.$$

Proof: Approximation Error

Theorem ([Sch17], Theorem 5)

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

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

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,

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

Approximation Error: Composition

Recall $g = g_q \circ \cdots \circ g_0$. Define

$$h_0 = \frac{g_0}{2F} + 1/2, \quad h_i = \frac{g_i(2F \cdot -F)}{2F} + 1/2 \quad \text{and} \quad h_q = g_q(2F \cdot -F).$$

Then $g = g_q \circ \cdots \circ g_0 = h_q \circ \cdots \circ h_0$.

Defining $H_i = h_i \circ \cdots \circ h_0$ and $\tilde{H}_i = \tilde{h}_i \circ \cdots \circ \tilde{h}_0$,

$$\begin{aligned} |H_i(x) - \tilde{H}_i(x)|_\infty &\leq |h_i \circ H_{i-1}(x) - h_i \circ \tilde{H}_{i-1}(x)|_\infty + \| \, |h_i - \tilde{h}_i|_\infty \|_{L^\infty([0,1]^{d_i})} \\ &\leq Q|H_{i-1} - \tilde{H}_{i-1}|_\infty + \| \, |h_i - \tilde{h}_i|_\infty \|_{L^\infty([0,1]^{d_i})}. \end{aligned}$$

Wasserstein Distance

$(\mathcal{X}, ||\cdot||)$ Polish metric space, here $\mathcal{X} = [0,1]^d$. $\mathcal{P}(\mathcal{X})$ set of Borel probability measures.

$$W_p(\mu,\nu) := \inf_{\pi \in \Pi(\mu,\nu)} \left(\int_{\mathcal{X}\times\mathcal{X}} ||x-y||^p d\pi(x,y) \right)^{1/p},$$

where $\Pi(\mu,\nu)$ is the set of joint distributions on $\mathcal{X}\times\mathcal{X}$ with marginals μ and ν.

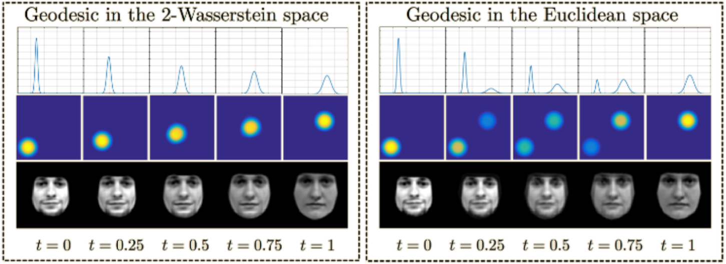


Figure: Interpolation in the optimal transport framework (left) and Euclidean space (right). Source: `www.math.cmu.edu/~mthorpe/OTNotes`

Conditional Encoder-Decoder-Structure

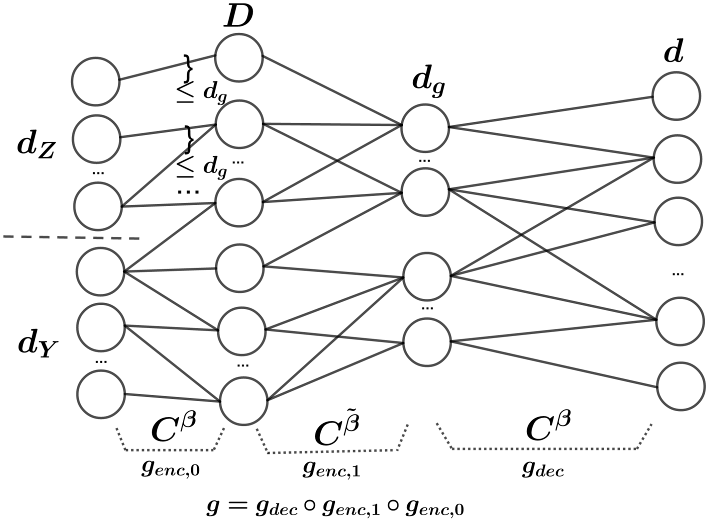


Figure: Possible Encoder-Decoder-Structure of g. [HR20]

Risk Bound in Terms of W_1^γ

Assume $\exists\, g^* \in \mathcal{G}: \quad \mathbb{P}^{g^*(Z)} = \mathbb{P}^X$.

Lemma 1: $$W_1^\gamma(g) \le W_{1,n}(g) + Cn^{-\frac{\gamma}{2\gamma+d}}.$$

Main Theorem: $$\mathbb{E} W_{1,n}(\hat{g}_n) \lesssim \left(\frac{s_f L_f \log(s_f L_f)}{n}\right)^{1/2} + \sqrt{d}\, \phi_{n\mathcal{E}}^{1/2} \log(n\mathcal{E})^{3/2}.$$

$$\overset{(i)-(iv)}{\Longrightarrow} \quad \mathbb{E} W_1^\gamma(\hat{g}_n) \lesssim n^{-\frac{\gamma}{2\gamma+d}} + \sqrt{d}\, \phi_{n\mathcal{E}}^{1/2} \log(n\mathcal{E})^{3/2}.$$

Choose minimal $\gamma \ge 1$, which recovers rate. E.g. for $\mathcal{E} = 1$:

$$\frac{\beta_{i^*}}{2\beta_{i^*} + t_{i^*}} = \min_{i=0,\dots,q} \frac{\beta_i}{2\beta_i + t_i} = \frac{\gamma}{2\gamma + d},$$

$$\overset{\gamma = \frac{\beta_{i^*} d}{t_{i^*}}}{\rightsquigarrow} \qquad \mathbb{E} W_1^\gamma(\hat{g}_n) \lesssim \phi_n^{1/2} \log(n)^{3/2}.$$

Risk Bound in Terms of W_1^γ

Assume $\exists\, g^* \in \mathcal{G}: \quad \mathbb{P}^{g^*(Z)} = \mathbb{P}^X$.

Lemma 1: $\quad W_1^\gamma(g) \leq W_{1,n}(g) + Cn^{-\frac{\gamma}{2\gamma+d}}$.

Main Theorem: $\quad \mathbb{E}W_{1,n}(\hat{g}_n) \lesssim \left(\frac{s_f L_f \log(s_f L_f)}{n}\right)^{1/2} + \sqrt{d}\,\phi_{n\mathcal{E}}^{1/2} \log(n\mathcal{E})^{3/2}$.

$$\overset{(i)-(iv)}{\Rightarrow} \quad \mathbb{E}W_1^\gamma(\hat{g}_n) \lesssim n^{-\frac{\gamma}{2\gamma+d}} + \sqrt{d}\,\phi_{n\mathcal{E}}^{1/2} \log(n\mathcal{E})^{3/2}.$$

Choose minimal $\gamma \geq 1$, which recovers rate. E.g. for $\mathcal{E} = 1$:

$$\frac{\beta_{i^*}}{2\beta_{i^*} + t_{i^*}} = \min_{i=0,\dots,q} \frac{\beta_i}{2\beta_i + t_i} = \frac{\gamma}{2\gamma + d},$$

$$\overset{\gamma = \frac{\beta_{i^*} d}{t_{i^*}}}{\rightsquigarrow} \quad \mathbb{E}W_1^\gamma(\hat{g}_n) \lesssim \phi_n^{1/2} \log(n)^{3/2}.$$

Risk Bound in Terms of W_1^γ

Assume $\exists\, g^* \in \mathcal{G}: \quad \mathbb{P}^{g^*(Z)} = \mathbb{P}^X$.

Lemma 1: $$W_1^\gamma(g) \leq W_{1,n}(g) + C n^{-\frac{\gamma}{2\gamma+d}}.$$

Main Theorem: $$\mathbb{E} W_{1,n}(\hat{g}_n) \lesssim \left(\frac{s_f L_f \log(s_f L_f)}{n}\right)^{1/2} + \sqrt{d}\, \phi_{n\mathcal{E}}^{1/2} \log(n\mathcal{E})^{3/2}.$$

$$\overset{(i)-(iv)}{\Rightarrow} \quad \mathbb{E} W_1^\gamma(\hat{g}_n) \lesssim n^{-\frac{\gamma}{2\gamma+d}} + \sqrt{d}\, \phi_{n\mathcal{E}}^{1/2} \log(n\mathcal{E})^{3/2}.$$

Choose minimal $\gamma \geq 1$, which recovers rate. E.g. for $\mathcal{E} = 1$:

$$\frac{\beta_{i^*}}{2\beta_{i^*} + t_{i^*}} = \min_{i=0,\dots,q} \frac{\beta_i}{2\beta_i + t_i} = \frac{\gamma}{2\gamma + d},$$

$$\overset{\gamma = \frac{\beta_{i^*} d}{t_{i^*}}}{\rightsquigarrow} \qquad \mathbb{E} W_1^\gamma(\hat{g}_n) \lesssim \phi_n^{1/2} \log(n)^{3/2}.$$

Is $W_{1,n}$ a meaningful distance measure?

$$W_{1,n}(g) := \sup_{f \in \mathcal{R}(L_f, p_f, s_f), \|f\|_L \le 1} \left\{ \mathbb{E} f(X,Y) - \mathbb{E} f(X, g(Z,X)) \right\}.$$

γ-Hölder smooth integral probability metric, $\gamma \ge 1$

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

Lemma 1 (Lower bound on $W_{1,n}$)

Let $a_n = n^{-\frac{2\gamma}{2\gamma+d+d_X}}$, and suppose that

(e) $F \ge 1$,

(f) $L_f \gtrsim \log_2(n)$,

(g) $\min_{i=1,\ldots,L} p_{f,i} \gtrsim n a_n$,

(h) $s_f \gtrsim \log(n) n a_n$,

where $\gtrsim$ dep. on γ, d. Then there exists a $C > 0$ only dep. on γ, d, F such that for any measurable $g : [0,1]^{d_Z + d_X} \to [0,1]^d$,

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

Asymptotic Confidence Intervals

Predict 1-dimensional *continuous* statistic $T(Y)$ given $X = x$ using $\left(X, \hat{g}_n(Z^*, X)\right) \overset{d}{\approx} \left(X, Y\right)$.

Sample N i.i.d. points $Z_j^* \sim \mathbb{P}^Z$ indep. of $X_i, Y_i, Z_i,\ i = 1\dots, n$.
Compute

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

yields asymptotic $(1-\alpha)$-confidence intervals for $T(Y)$ given $X = x$,

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

Temperature in German Cities

Predict mean temperature in Berlin given mean temperatures of $d_X = 32$ (or $d_X = 3$) German cities on the previous day.

Training set (4300 days): 2006/07/01 - 2018/04/09.

Test set (478 days): 2018/04/10 - 2019/07/31.

Generator: $p_g = (4 + d_X, 10, 10, 10, d)$.

Critic: $p_f = (d + d_X, 32, 32, 32, 32, 32, 1)$.

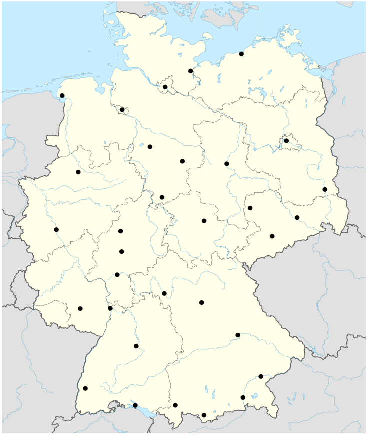
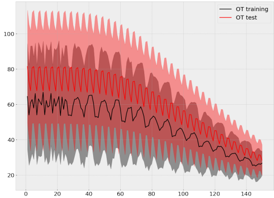


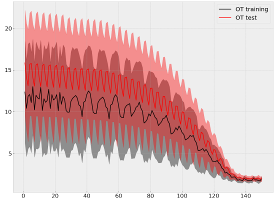
Figure: Data from Deutscher Wetterdienst, map from the authors of [PR20].

Temperature in German Cities

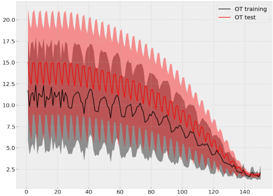
Empirical Wasserstein Loss



(a) 32 to 32

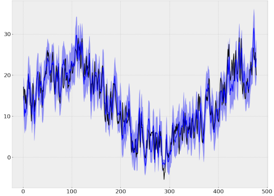


(b) 32 to 1

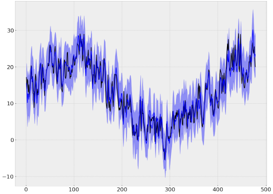


(c) 3 to 1

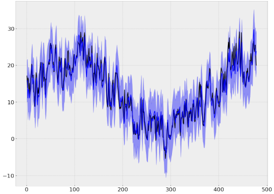
Predictions on Test Set



(a) 32 to 32



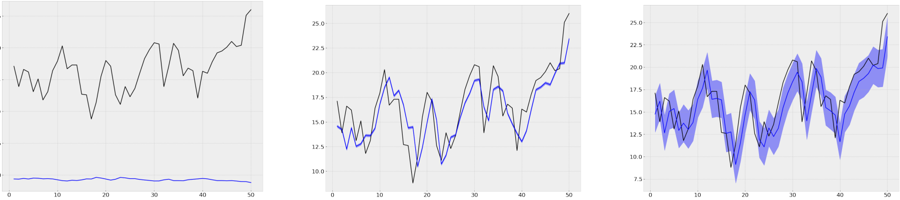
(b) 32 to 1



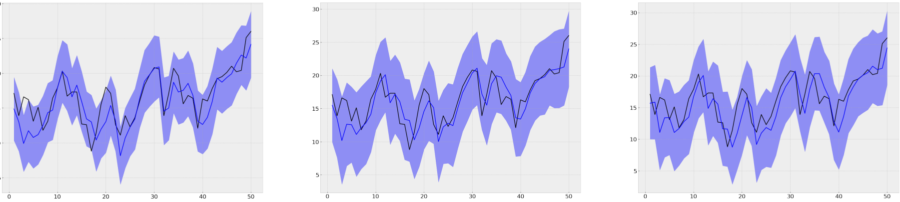
(c) 3 to 1

Temperature in German Cities

After 150 Epochs



After 1000 Epochs



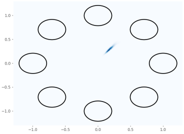
(a) 32 to 32: 70.71%
(88.70% after 2150 ep.)

(b) 32 to 1: 89.54%

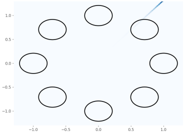
(c) 3 to 1: 89.96%

GAN Comparison

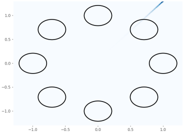
GAN [Goo+14]:



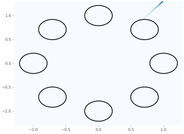
(a) 0 epochs



(b) 5 epochs

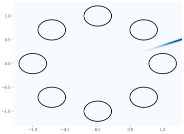


(c) 100 epochs

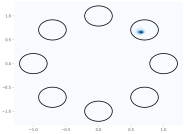


(d) 300 epochs
$\geq$ 35000 updates

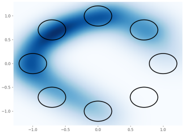
Least Squares GAN [Mao+16]:



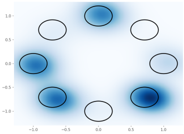
(a) 0 epochs



(b) 100 epochs



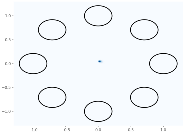
(c) 120 epochs



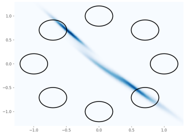
(d) 280 epochs
$\geq$ 34000 updates

GAN Comparison

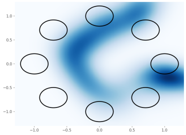
WGAN [ACB17]:



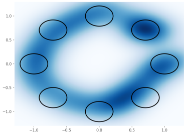
(a) 10 epochs



(b) 140 epochs

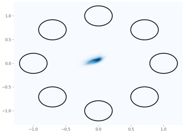


(c) 180 epochs

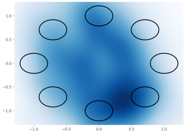


(d) 300 epochs
$\geq$ 6700 updates

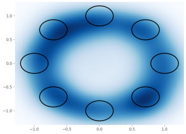
WGAN-GP [Gul+17]:



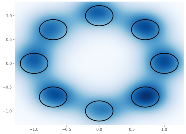
(a) 10 epochs



(b) 15 epochs



(c) 25 epochs



(d) 50 epochs
$\geq$ 1300 updates

Learn to generate samples from a probability distribution

Figure: WGAN-GP [Gul+17] with DC-GAN networks [RMC16] (2017)

StyleGAN2 [Kar+19] (2019): `www.thispersondoesnotexist.com`