Statistical Analysis of Wasserstein GANs [1]

Moritz Haas[1], Stefan Richter[2]

[1]Ulrike von Luxburg, Theory of Machine Learning, Universität Tübingen.
[2]Universität Heidelberg

Unconditional Problem

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

Given $Y_i \sim \mathbb{P}^Y$, $i = 1, \ldots, n$ 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 $Y_1, \ldots, Y_n$.
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

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

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

useful for uncertainty quantification in prediction.

Network-based Wasserstein Objective

Dual formulation [2] of W_1-distance with critic functions f:

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

So find g minimizing

$$\begin{aligned} W_1(g) := \; & W_1(\mathbb{P}^{(X,Y)}, \mathbb{P}^{(g(Z,Y),Y)}) \\ = & \sup_{f:\mathbb{R}^{d+d_Y}\to\mathbb{R},\, ||f||_L \le 1} \mathbb{E}f(X, Y) - \mathbb{E}f(g(Z, Y), Y). \end{aligned}$$

W_1 not available in practice $\rightsquigarrow$ Approximation with critic networks f:

- Modified network-based Wasserstein Distance
$$W_{1,n}(g) := \sup_{f\in\mathcal{R}_D,\, ||f||_L\le 1} \big\{\mathbb{E}f(X, Y) - \mathbb{E}f(X, g(Z, X))\big\}.$$
- For empirical version replace $\mathbb{E}$ by $\frac{1}{n}\sum_{i=1}^n$.

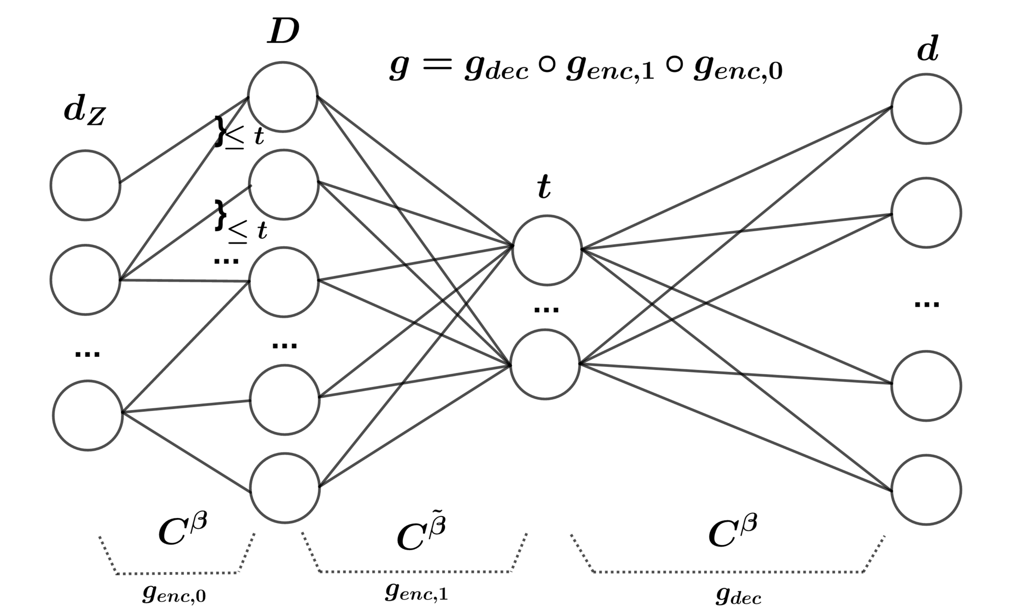
Assumptions

- Network class $\mathcal{R}(L, \mathbf{p}, s)$: bounded, sparse ReLU networks of depth L, width vector $\mathbf{p}$ and number of non-zero weight entries s (cf. [3]).
- Class of generator functions $\mathcal{G}$: Compositions of t-sparse, β-Hölder smooth functions (cf. [3]).

Assume

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

Structure of true generator function:



Network Growth Assumptions: With the rate

$$\phi_{n\mathcal{E}} := (n\mathcal{E})^{-\frac{2\beta}{2\beta+t}},$$

where $\mathcal{E} \propto$ number of epochs (if you can sample from $\mathbb{P}^X$),

(a) $L_g \asymp \log(n\mathcal{E})$,
(b) $\min_{i=1,\ldots,L_g} p_{g,i} \asymp (n\mathcal{E}) \cdot \phi_{n\mathcal{E}}$,
(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 ($L_g \lesssim L_f$, $s_g \lesssim s_f$).

Convergence Rates

Main Theorem (Excess Risk Bound):
Suppose assumptions (a)-(d) hold and (X, Y) β-mixing of order $O(k^{-\alpha})$ with $\alpha > 1$, then for the empirical risk minimizer $\hat{g}_n$,

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

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

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

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

$\rightsquigarrow$

- **approx. rate $\frac{1}{\sqrt{n}}$ for Hölder smoothness $\beta \to \infty$,**
- **remove influence of d and complexity of $\mathbb{P}^{X,Y}$ by training long enough!**

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

If the critic networks grow fast enough, $W_{1,n}$ **and** W_1 **are equivalent**.

Lemma 2 (Characterization of weak convergence):
If $L_f, \mathbf{p}_f, s_f$ satisfy assumptions (a)-(c) with $\phi_n = n^{-\frac{2\gamma}{2\gamma+d+d_X}}$ for some $\gamma \ge 1$. Then, for $n \to \infty$,

$$W_1(\mathbb{P}^{X_n}, \mathbb{P}^X) \to 0 \iff W_{1,n}(\mathbb{P}^{X_n}, \mathbb{P}^X) \to 0.$$

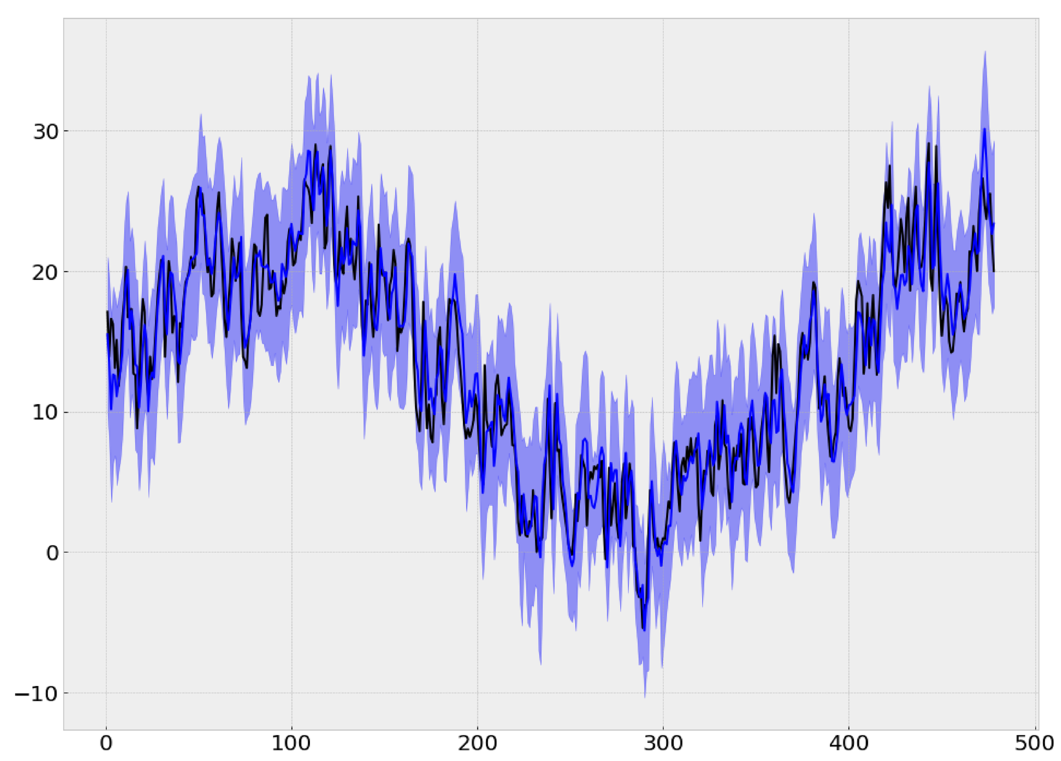
Lemma (estimated distribution converges):
Under the assumptions above, if $\mathbb{E}W_{1,n}(\tilde{g}_n) \to 0$, then

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

$\rightsquigarrow$ optimal growth rate of critic networks: To recover the convergence rate $\phi_{n\mathcal{E}}^{1/2}$, choose $\gamma = \frac{\beta d}{t}$.

Applications to prediction

Get **uncertainty estimates** from $\mathbb{P}^{\hat{g}_n(Z,x)} \approx \mathbb{P}^{Y|X=x}$:
Learn **conditional distribution** of temperatures in 32 German cities given temperatures on previous day.



Conclusions

- formalize Wasserstein GANs theoretically (with growing network architectures unlike [4]),
- $W_{1,n}$ 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. prediction,
  $\rightsquigarrow$ simulation studies show good empirical coverage,
- explains good performance under long training for large and complex generators and/or large dimension d.

References

[1] Moritz Haas and Stefan Richter. Statistical analysis of wasserstein gans with applications to time series forecasting, 2020.
[2] C. Villani. *Optimal transport – Old and new*, volume 338, pages 43–113. 01 2008. doi: 10.1007/978-3-540-71050-9.
[3] J. Schmidt-Hieber. Nonparametric regression using deep neural networks with relu activation function, 2017.
[4] Gérard Biau, Maxime Sangnier, and Ugo Tanielian. Some theoretical insights into wasserstein gans, 2020.

Contact information:
Moritz Haas
mo.haas@uni-tuebingen.de

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN