GANs, Optimal Transport, and Implicit Distribution Estimation

Tengyuan Liang

Econometrics and Statistics

CHICAGO BOOTH
The University of Chicago Booth School of Business

Cowles Foundation for Research in Economics

Home People About Us Research Programs Publications News Events Resources Contact

HOME > FROM THE ARCHIVES > PEOPLE > TJALLING C. KOOPMANS (1910-1985)

From the Archives

From the Archives

Events

People

Photo Gallery

Research Reports

Tjalling C. Koopmans (1910-1985)

Director: July 1948–1955; 1961–1964; 1965–1967

Ph.D., University of Leiden, 1936

Tjalling C. Koopmans lectured at the Rotterdam School of Economics and served on the staff of the Netherlands Economic Institute, 1936–37. From 1938 to 1940 he was engaged in business-cycle research at the League of Nations in Geneva. In 1910–41 he was on the staff of the Local and State Government Section of the School for Public and International Affairs, Princeton University, and also taught statistics at New York University. In 1941–42, he was economist with the Penn Mutual Life Insurance Company, and in 1942–44 he was statistician to the Combined Shipping Adjustment Board at Washington. Koopmans joined the staff of the Cowles Commission in July 1944, as a research associate. In 1946 he also became an associate professor in the Department of Economics at the University of Chicago. In 1948 he was appointed director of research of the Commission and professor of economics at the University of Chicago. He was elected a Fellow of the Econometric Society in 1940, of the Institute of Mathematical Statistics in 1941, of the American Statistical Association in 1949, and a member of the International

Outline

Implicit Distribution Estimation

Given i.i.d. $Y_1, \ldots, Y_n \sim \nu$. Use transformation $T : \mathbb{R}^d \to \mathbb{R}^d$ to represent and learn unknown dist. $Y \sim \nu$ via simple $Z \sim \mu$ (say Uniform or Gaussian).

$$T(Z) \overset{\text{close in dist.?}}{\approx} Y$$

OUTLINE

Implicit Distribution Estimation

Given i.i.d. $Y_1, \ldots, Y_n \sim \nu$. Use transformation $T : \mathbb{R}^d \to \mathbb{R}^d$ to represent and learn unknown dist. $Y \sim \nu$ via simple $Z \sim \mu$ (say Uniform or Gaussian).

$$T(Z) \overset{\text{close in dist.?}}{\approx} Y$$

equivalently

$$T_{\#}\mu \overset{?}{\approx} \nu$$

OUTLINE

Implicit Distribution Estimation

Generative Adversarial Networks

- statistical rates
- pair regularization
- optimization

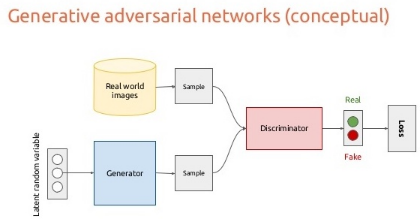
Optimal Transport

- estimate the Wasserstein metric

vs.

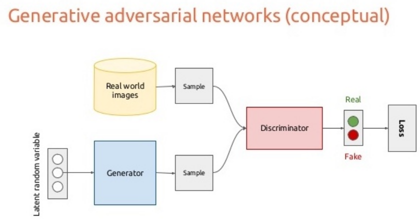
- estimate under the Wasserstein metric

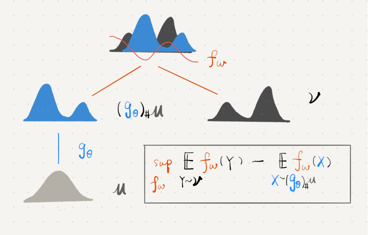
Generative adversarial networks



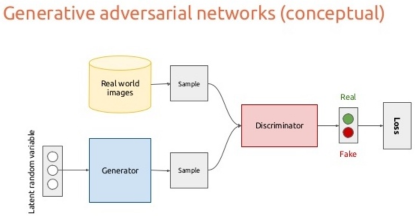
- GAN Goodfellow et al. (2014)
- WGAN Arjovsky et al. (2017); Arjovsky and Bottou (2017)
- MMD GAN Li, Swersky, and Zemel (2015); Dziugaite, Roy, and Ghahramani (2015); Arbel, Sutherland, Bińkowski, and Gretton (2018)
- f-GAN Nowozin, Cseke, and Tomioka (2016)
- Sobolev GAN Mroueh et al. (2017)
- many others... Liu, Bousquet, and Chaudhuri (2017); Tolstikhin, Gelly, Bousquet, Simon-Gabriel, and Schölkopf (2017)

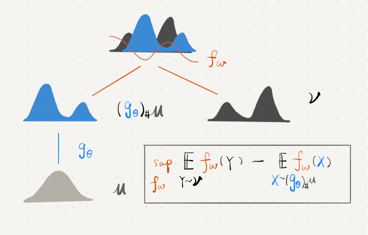
Generative adversarial networks





Generative adversarial networks





Generator g_θ, Discriminator f_ω

$$U(\theta, \omega) = \underbrace{\mathbb{E}_{Y\sim\nu}}_{\text{target}} [f_\omega(Y)] - \underbrace{\mathbb{E}_{Z\sim\mu}}_{\text{input}} [f_\omega(g_\theta(Z))]$$

$$\min_\theta \max_\omega U(\theta, \omega)$$

GANs are widely used in practice, however

Much needs to be understood, in theory

- Approximation:

  what dist. can be approximated by the generator $(g_\theta)_\#(\mu)$?

- **Statistical**:

  given n samples, what is the statistical/generalization error rate?

- Computational:

  local convergence for practical optimization, how to stablize?

- Landscape:

  are local saddle points good globally?

FORMULATION

$\mathcal{T}_G$ class of generator transformations, $\mathcal{F}_D$ class of discriminator functions
ν target dist.

population $$g^* \in \operatorname*{arg\,min}_{g \in \mathcal{T}_G} \max_{f \in \mathcal{F}_D} \left\{ \mathop{\mathbb{E}}_{X \sim g_{\#}\mu} [f(X)] - \mathop{\mathbb{E}}_{Y \sim \nu} [f(Y)] \right\}$$

Formulation

$\mathcal{T}_G$ class of generator transformations, $\mathcal{F}_D$ class of discriminator functions
ν target dist.

population $$g^* \in \operatorname*{arg\,min}_{g \in \mathcal{T}_G} \max_{f \in \mathcal{F}_D} \left\{ \mathop{\mathbb{E}}_{X \sim g_\# \mu} [f(X)] - \mathop{\mathbb{E}}_{Y \sim \nu} [f(Y)] \right\}$$

$\widehat{\nu}^n$ empirical dist.

empirical $$\widehat{g} \in \operatorname*{arg\,min}_{g \in \mathcal{T}_G} \max_{f \in \mathcal{F}_D} \left\{ \mathop{\mathbb{E}}_{X \sim g_\# \mu} [f(X)] - \mathop{\mathbb{E}}_{Y \sim \widehat{\nu}^n} [f(Y)] \right\}$$

$\widehat{g}_\# \mu$ as estimate for ν

Formulation

$\mathcal{T}_G$ class of generator transformations, $\mathcal{F}_D$ class of discriminator functions
ν target dist.

population $$g^* \in \arg\min_{g \in \mathcal{T}_G} \max_{f \in \mathcal{F}_D} \left\{ \mathop{\mathbb{E}}_{X \sim g_\# \mu} [f(X)] - \mathop{\mathbb{E}}_{Y \sim \nu} [f(Y)] \right\}$$

$\widehat{\nu}^n$ empirical dist.

empirical $$\widehat{g} \in \arg\min_{g \in \mathcal{T}_G} \max_{f \in \mathcal{F}_D} \left\{ \mathop{\mathbb{E}}_{X \sim g_\# \mu} [f(X)] - \mathop{\mathbb{E}}_{Y \sim \widehat{\nu}^n} [f(Y)] \right\}$$

$\widehat{g}_\# \mu$ as estimate for ν

- Density learning/estimation: long history nonparametric statistics
  model target density $\rho_\nu \in W^\alpha$ - Sobolev space with smoothness $\alpha \geq 0$
  Stone (1982); Nemirovski (2000); Tsybakov (2009); Wassermann (2006)
- GAN statistical theory is needed
  Arora and Zhang (2017); Arora et al. (2017a,b); Liu et al. (2017)

Discriminator metric

Define the critic metric (IPM)

$$d_{\mathcal{F}}(\mu, \nu) := \sup_{f \in \mathcal{F}} \left| \mathop{\mathbb{E}}_{X \sim \mu} f(X) - \mathop{\mathbb{E}}_{Y \sim \nu} f(Y) \right| .$$

DISCRIMINATOR METRIC

Define the critic metric (IPM)

$$d_{\mathcal{F}}(\mu, \nu) := \sup_{f \in \mathcal{F}} \left| \mathbb{E}_{X \sim \mu} f(X) - \mathbb{E}_{Y \sim \nu} f(Y) \right| .$$

- $\mathcal{F}$ Lip-1: Wasserstein metric d_W
- $\mathcal{F}$ bounded by 1: total variation/Radon metric d_{TV}
- RKHS $\mathcal{H}$, $\mathcal{F} = \{f \in \mathcal{H}, \|f\|_{\mathcal{H}} \leq 1\}$: MMD GAN
- $\mathcal{F}$ Sobolev smoothness β: Sobolev GAN

Statistical question: statistical error rate with n-i.i.d samples, $\mathbb{E}\, d_{\mathcal{F}}(\nu, \widehat{\mu}_n)$?

for a range of $\mathcal{F}$ and ν with certain regularity.

Summary of First Half of Talk

Goal	Evaluation Metric	Results		Generator Class $\mathcal{G}$	Discriminator Class $\mathcal{F}$	Property
Adversarial Framework (nonparametric)	$d_{\mathcal{F}}$	Sobolev GAN	minimax optimal	Sobolev W^{α}	Sobolev W^{β}	
		MMD GAN	upper bound	smooth subspace in RKHS	RKHS $\mathcal{H}$	
			oracle results	any	Sobolev W^{β}	$\mathcal{G}\dagger$
Generative Adversarial Networks (parametric)	d_{TV}	leaky-ReLU GANs	upper bound	leaky-ReLU	leaky-ReLU	$\mathcal{F}\ddagger$, $m*$
	d_{TV}, d_{KL}, d_H	any GANs	oracle results	neural networks	neural networks	$\mathcal{G}\dagger$, $\mathcal{F}\ddagger$, $m*$
	d_W	Lipschitz GANs	oracle results	Lipschitz neural networks	Lipschitz neural networks	$\mathcal{G}\dagger$, $\mathcal{F}\ddagger$, $m*$

SUMMARY OF FIRST HALF OF TALK

Goal	Evaluation Metric	Results		Generator Class $\mathcal{G}$	Discriminator Class $\mathcal{F}$	Property
Adversarial Framework (nonparametric)	$d_{\mathcal{F}}$	Sobolev GAN	minimax optimal	Sobolev W^{α}	Sobolev W^{β}	
		MMD GAN	upper bound	smooth subspace in RKHS	RKHS $\mathcal{H}$	
			oracle results	any	Sobolev W^{β}	$\mathcal{G}\dagger$
Generative Adversarial Networks (parametric)	d_{TV}	leaky-ReLU GANs	upper bound	leaky-ReLU	leaky-ReLU	$\mathcal{F}\ddagger$, $m*$
	d_{TV}, d_{KL}, d_H	any GANs	oracle results	neural networks	neural networks	$\mathcal{G}\dagger$, $\mathcal{F}\ddagger$, $m*$
	d_W	Lipschitz GANs	oracle results	Lipschitz neural networks	Lipschitz neural networks	$\mathcal{G}\dagger$, $\mathcal{F}\ddagger$, $m*$

The symbols: ($\mathcal{G}\dagger$) and ($\mathcal{F}\ddagger$) to denote the mis-specification for the generator class and the discriminator class respectively, and ($m*$) to indicate the dependence on the number of generator samples.

Implicit Distribution Estimator: GANs, Optimal Transport

vs.

Explicit Density Estimator: KDE, Projection/Series Estimator, …

Adversarial Framework
(nonparametric)

MINIMAX OPTIMAL RATES: SOBOLEV GAN

Consider the target $\mathcal{G} := \{\nu : \rho_\nu \in \mathsf{W}^{\alpha}\}$ Sobolev space with smoothness α, and the evaluation metric $\mathcal{F} = \mathsf{W}^{\beta}$ with smoothness β.

MINIMAX OPTIMAL RATES: SOBOLEV GAN

Consider the target $\mathcal{G} := \{\nu : \rho_\nu \in W^\alpha\}$ Sobolev space with smoothness α, and the evaluation metric $\mathcal{F} = W^\beta$ with smoothness β.

Theorem (L. '17 & L. '18, Sobolev)**.**

The **minimax optimal rate** is

$$\inf_{\widetilde{\nu}_n} \sup_{\nu \in \mathcal{G}} \mathbb{E}\, d_{\mathcal{F}}(\nu, \widetilde{\nu}_n) \asymp n^{-\frac{\alpha+\beta}{2\alpha+d}} \vee n^{-\frac{1}{2}} .$$

Here $\widetilde{\nu}_n$ any estimator based on n samples. d-dim.

Liang (2017); Singh et al. (2018); Weed and Berthet (2019)

MINIMAX OPTIMAL RATES: MMD GAN

Consider a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$

- integral operator $\mathcal{T}$ with eigenvalue decay $t_i \asymp i^{-\kappa}, 0 < \kappa < \infty$
- evaluation metric $\mathcal{F} = \{f \in \mathcal{H} \mid \|f\|_{\mathcal{H}} \leq 1\}$
- target density ρ_ν in $\mathcal{G} = \{\nu \mid \|\mathcal{T}^{-\frac{\alpha-1}{2}} \rho_\nu\|_{\mathcal{H}} \leq 1\}$ with smoothness α

Minimax optimal rates: MMD GAN

Consider a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$

- integral operator $\mathcal{T}$ with eigenvalue decay $t_i \asymp i^{-\kappa}, 0 < \kappa < \infty$
- evaluation metric $\mathcal{F} = \{f \in \mathcal{H} \mid \|f\|_{\mathcal{H}} \leq 1\}$
- target density ρ_ν in $\mathcal{G} = \{\nu \mid \|\mathcal{T}^{-\frac{\alpha-1}{2}} \rho_\nu\|_{\mathcal{H}} \leq 1\}$ with smoothness α

Theorem (L. '18, RKHS)**.**

The **minimax optimal rate** is

$$\inf_{\widetilde{\nu}_n} \sup_{\nu \in \mathcal{G}} \mathbb{E}\, d_{\mathcal{F}}(\nu, \widetilde{\nu}_n) \precsim n^{-\frac{(\alpha+1)\kappa}{2\alpha\kappa+2}} \vee n^{-\frac{1}{2}} .$$

MINIMAX OPTIMAL RATES: MMD GAN

Consider a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$

- integral operator $\mathcal{T}$ with eigenvalue decay $t_i \asymp i^{-\kappa}, 0 < \kappa < \infty$
- evaluation metric $\mathcal{F} = \{f \in \mathcal{H} \mid \|f\|_{\mathcal{H}} \leq 1\}$
- target density ρ_ν in $\mathcal{G} = \{\nu \mid \|\mathcal{T}^{-\frac{\alpha-1}{2}} \rho_\nu\|_{\mathcal{H}} \leq 1\}$ with smoothness α

Theorem (L. '18, RKHS).

The **minimax optimal rate** is

$$\inf_{\widetilde{\nu}_n} \sup_{\nu \in \mathcal{G}} \mathbb{E} d_{\mathcal{F}}(\nu, \widetilde{\nu}_n) \precsim n^{-\frac{(\alpha+1)\kappa}{2\alpha\kappa+2}} \vee n^{-\frac{1}{2}} .$$

$\kappa > 1$: intrinsic dim. $\sum_{i \geq 1} t_i = \sum_{i \geq 1} i^{-\kappa} \leq C$, parametric rate $n^{-\frac{(\alpha+1)\kappa}{2\alpha\kappa+2}} \vee n^{-\frac{1}{2}} = n^{-1/2}$.

$\kappa < 1$: sample complexity scales $n = \epsilon^{2+\frac{2}{\alpha+1}\left(\frac{1}{\kappa}-1\right)}$, effective dim. $\frac{1}{\kappa}$.

Oracle inequality for GANs

Generator class $\mathcal{G}$ may not contain the **target** ν: oracle approach.

Oracle inequality for GANs

Generator class $\mathcal{G}$ may not contain the **target** ν: oracle approach.

Let $\mathcal{T}_G$ be **any** generator transformation. The discriminator metric $\mathcal{F}_D = \mathsf{W}^{\beta}$, target density $\rho_\nu \in \mathsf{W}^{\alpha}$.

$\widehat{g}_{\#}\mu$, $\widetilde{g}_{\#}\mu$ are Implicit Density Estimators!

Corollary (L. '17).

With empirical $\widehat{\nu}^n$ as plug-in, GAN

$$\widehat{g} \in \arg\min_{g \in \mathcal{T}_G} \max_{f \in \mathcal{F}_D} \left\{ \mathbb{E}_{X \sim g_{\#}\mu}[f(X)] - \mathbb{E}_{Y \sim \widehat{\nu}^n}[f(Y)] \right\},$$

attains a **sub-optimal rate**

$$\mathbb{E}\, d_{\mathcal{F}_D}(\widehat{g}_{\#}\mu, \nu) \le \min_{g \in \mathcal{T}_G} d_{\mathcal{F}_D}(g_{\#}\mu, \nu) + \boxed{n^{-\frac{\beta}{d}} \vee \frac{\log n}{\sqrt{n}}}.$$

Canas and Rosasco (2012): $\beta = 1$

Oracle inequality for GANs

Generator class $\mathcal{G}$ may not contain the **target** ν: oracle approach.

Let $\mathcal{T}_G$ be **any** generator transformation. The discriminator metric $\mathcal{F}_D = W^{\beta}$, target density $\rho_\nu \in W^{\alpha}$.

$\widehat{g}_{\#}\mu$, $\widetilde{g}_{\#}\mu$ are Implicit Dens[...] mators!

Corollary (L. '17).

With empirical $\widehat{\nu}^n$ as plug-in

$$\widehat{g} \in \arg\min_{g \in \mathcal{T}_G} \max_{f \in \mathcal{F}_D} \left\{ \mathop{\mathbb{E}}_{X \sim g_{\#}\mu} [f(X) \ldots \right.$$

attains a **sub-optimal rate**

$$\mathbb{E}\, d_{\mathcal{F}_D}(\widehat{g}_{\#}\mu, \nu) \le \min_{g \in \mathcal{T}_G} d_{\mathcal{F}_D}(g_{\#}\mu, \nu) + \boxed{n^{-\frac{\beta}{d}} \vee \frac{\log n}{\sqrt{n}}}.$$

Corollary (L. '17).

In contrast, a regularized empirical $\widetilde{\nu}^n$ as plug-in

$$\widetilde{g} \in \arg\min_{g \in \mathcal{T}_G} \max_{f \in \mathcal{F}_D} \left\{ \mathop{\mathbb{E}}_{X \sim g_{\#}\mu} [f(X)] - \mathop{\mathbb{E}}_{Y \sim \widetilde{\nu}^n} [f(Y)] \right\},$$

a **faster rate** is attainable

$$\mathbb{E}\, d_{\mathcal{F}_D}(\widetilde{g}_{\#}\mu, \nu) \le \min_{g \in \mathcal{T}_G} d_{\mathcal{F}_D}(g_{\#}\mu, \nu) + \boxed{n^{-\frac{\alpha+\beta}{2\alpha+d}} \vee \frac{1}{\sqrt{n}}}.$$

Canas and Rosasco (2012): $\beta = 1$

Sub-optimality and Regularization

Regularization helps achieve faster rate

Use $\widetilde{\nu}^n$ "smoothed" empirical estimate, that serves as regularization

For example, kernel smoothing: $\widetilde{\nu}^n(x) = \frac{1}{nh_n} K\left(\frac{x - x_i}{h_n}\right)$, SGD works

Turns out, this is used in practice, called "instance noise" or "data augmentation"

Sønderby et al. (2016); Liang et al. (2017); Arjovsky and Bottou (2017); Mescheder et al. (2018)

Generative Adversarial Networks and Pair Regularization
(parametric)

Consider the parametrized GAN estimator

$$\widehat{\theta}_{m,n} \in \operatorname*{arg\,min}_{\theta: g_\theta \in \mathcal{G}} \max_{\omega: f_\omega \in \mathcal{F}} \left\{ \widehat{\mathbb{E}}_m f_\omega(g_\theta(Z)) - \widehat{\mathbb{E}}_n f_\omega(Y) \right\},$$

with m generator samples and n **target samples**.

How well GANs learn the distribution, under **objective evaluation metric**, say

$$d_{TV}\left((g_{\widehat{\theta}_{m,n}})_{\#}\mu, \nu \right) \ ?$$

Generalized oracle inequality

$$\text{approx. err.} \quad \boxed{A_1(\mathcal{F},\mathcal{G},\nu) := \sup_{\theta}\inf_{\omega}\left\|\log\frac{\rho_\nu}{\rho_{\mu_\theta}} - f_\omega\right\|}, \quad A_2(\mathcal{G},\nu) := \inf_{\theta}\left\|\log\frac{\rho_{\mu_\theta}}{\rho_\nu}\right\|^{1/2},$$

$$\text{sto. err.} \quad S_{n,m}(\mathcal{F},\mathcal{G}) := \sqrt{\mathrm{Pdim}(\mathcal{F})\frac{\log(m\wedge n)}{m\wedge n}} \vee \sqrt{\mathrm{Pdim}(\mathcal{F}\circ\mathcal{G})\frac{\log(m)}{m}},$$

$\mathrm{Pdim}(\cdot)$ the pseudo-dimension of the neural network function.

Generalized oracle inequality

$$\text{approx. err.} \quad \boxed{A_1(\mathcal{F}, \mathcal{G}, \nu) := \sup_{\theta} \inf_{\omega} \left\| \log \frac{\rho_\nu}{\rho_{\mu_\theta}} - f_\omega \right\|}, \quad A_2(\mathcal{G}, \nu) := \inf_{\theta} \left\| \log \frac{\rho_{\mu_\theta}}{\rho_\nu} \right\|^{1/2},$$

$$\text{sto. err.} \quad S_{n,m}(\mathcal{F}, \mathcal{G}) := \sqrt{\mathrm{Pdim}(\mathcal{F}) \frac{\log(m \wedge n)}{m \wedge n}} \vee \sqrt{\mathrm{Pdim}(\mathcal{F} \circ \mathcal{G}) \frac{\log(m)}{m}},$$

$\mathrm{Pdim}(\cdot)$ the pseudo-dimension of the neural network function.

Theorem (L. '18, generalized oracle inequality)**.**

$$\mathbb{E}\, d_{TV}^2 \left(\nu, (g_{\widehat{\theta}_{m,n}})_\# \mu\right), \mathbb{E}\, d_W^2 \left(\nu, (g_{\widehat{\theta}_{m,n}})_\# \mu\right),$$
$$\mathbb{E}\, d_{KL} \left(\nu \| (g_{\widehat{\theta}_{m,n}})_\# \mu\right) + \mathbb{E}\, d_{KL} \left((g_{\widehat{\theta}_{m,n}})_\# \mu \| \nu\right)$$
$$\leq A_1(\mathcal{F}, \mathcal{G}, \nu) + A_2(\mathcal{G}, \nu) + S_{n,m}(\mathcal{F}, \mathcal{G}) \ .$$

GENERALIZED ORACLE INEQUALITY

approx. err. $\boxed{A_1(\mathcal{F},\mathcal{G},\nu) := \sup_{\theta} \inf_{\omega} \left\| \log \frac{\rho_\nu}{\rho_{\mu_\theta}} - f_\omega \right\|}$, $A_2(\mathcal{G},\nu) := \inf_{\theta} \left\| \log \frac{\rho_{\mu_\theta}}{\rho_\nu} \right\|^{1/2}$,

sto. err. $S_{n,m}(\mathcal{F},\mathcal{G}) := \sqrt{\mathrm{Pdim}(\mathcal{F}) \frac{\log(m \wedge n)}{m \wedge n}} \vee \sqrt{\mathrm{Pdim}(\mathcal{F} \circ \mathcal{G}) \frac{\log(m)}{m}}$,

$\mathrm{Pdim}(\cdot)$ the pseudo-dimension of the neural network function.

Theorem (L. '18, generalized oracle inequality)**.**

$$
\begin{aligned}
&\mathbb{E}\, d_{TV}^2\left(\nu, (g_{\widehat{\theta}_{m,n}})_{\#}\mu\right), \mathbb{E}\, d_W^2\left(\nu, (g_{\widehat{\theta}_{m,n}})_{\#}\mu\right), \\
&\mathbb{E}\, d_{KL}\left(\nu \| (g_{\widehat{\theta}_{m,n}})_{\#}\mu\right) + \mathbb{E}\, d_{KL}\left((g_{\widehat{\theta}_{m,n}})_{\#}\mu \| \nu\right) \\
&\leq A_1(\mathcal{F},\mathcal{G},\nu) + A_2(\mathcal{G},\nu) + S_{n,m}(\mathcal{F},\mathcal{G}) \ .
\end{aligned}
$$

We emphasize on the interplay between $(\mathcal{G},\mathcal{F})$
as a **pair** of tuning parameters for **regularization**.

approx. err. $$\boxed{A_1(\mathcal{F},\mathcal{G},\nu) := \sup_{\theta}\inf_{\omega}\left\| \frac{\sqrt{\rho_\nu}-\sqrt{\rho_{\mu_\theta}}}{\sqrt{\rho_\nu}+\sqrt{\rho_{\mu_\theta}}} - f_\omega \right\|},$$

$$A_2(\mathcal{G},\nu) := \inf_{\theta}\left\| \frac{\sqrt{\rho_\nu}-\sqrt{\rho_{\mu_\theta}}}{\sqrt{\rho_\nu}+\sqrt{\rho_{\mu_\theta}}} \right\|,$$

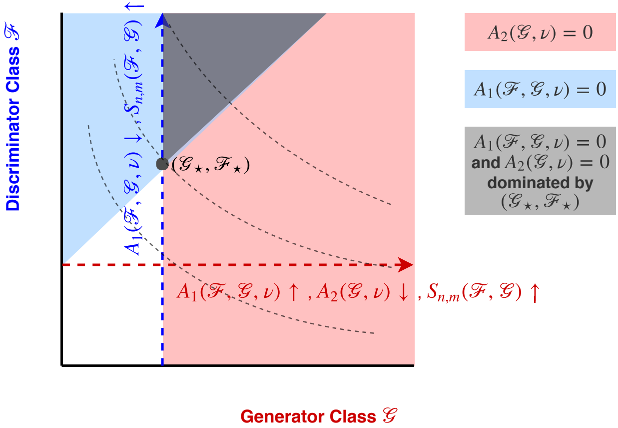
Theorem (L. '18, generalized oracle inequality)**.**

$$\mathbb{E}\, d_{TV}^2\left(\nu, (g_{\widehat{\theta}_{m,n}})_{\#}\mu\right), \mathbb{E}\, d_H^2\left(\nu, (g_{\widehat{\theta}_{m,n}})_{\#}\mu\right),$$
$$\le A_1(\mathcal{F},\mathcal{G},\nu) + A_2(\mathcal{G},\nu) + S_{n,m}(\mathcal{F},\mathcal{G}) \quad .$$

similar result for Hellinger d_H, for non-absolutely continuous $(g_\theta)_{\#}\mu$ and ν.

PAIR REGULARIZATION

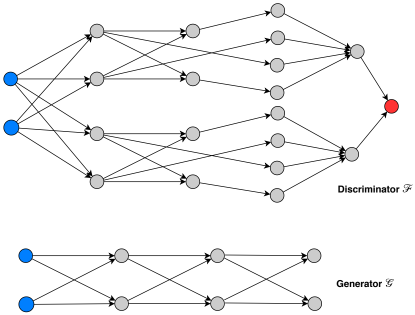
$$\text{fix } \mathcal{G}, \text{ as } \mathcal{F} \text{ increase}: \quad A_1(\mathcal{F}, \mathcal{G}, \nu) \text{ decrease}, A_2(\mathcal{G}, \nu) \text{ constant}, S_{n,m}(\mathcal{F}, \mathcal{G}) \text{ increase},$$
$$\text{fix } \mathcal{F}, \text{ as } \mathcal{G} \text{ increase}: \quad A_1(\mathcal{F}, \mathcal{G}, \nu) \text{ increase}, A_2(\mathcal{G}, \nu) \text{ decrease}, S_{n,m}(\mathcal{F}, \mathcal{G}) \text{ increase}.$$



Applications of pair regularization

APPLICATION I: PARAMETRIC RATES FOR LEAKY RELU NETWORKS

When the generator $\mathcal{G}$ and discriminator $\mathcal{F}$ are both leaky ReLU networks with depth **L** (width properly chosen depends on dimension).



When the target density is realizable by the generator.

$$\log \rho_{(g_\theta)_\# \mu}(x) = c_1 \sum_{l=1}^{\mathbf{L}-1} \sum_{i=1}^{d} \mathbf{1}_{m_{li}(x) \geq 0} + c_0,$$

Bai et al. (2018)

APPLICATION I: PARAMETRIC RATES FOR LEAKY RELU NETWORKS

When the generator $\mathcal{G}$ and discriminator $\mathcal{F}$ are both leaky ReLU networks with depth $\mathbf{L}$ (width properly chosen depends on dimension).

Theorem (L. '18, leaky ReLU).

$$\mathbb{E}\, d_{TV}^2\left(\nu, (g_{\widehat{\theta}_{m,n}})_{\#}\mu\right) \precsim \sqrt{d^2\mathbf{L}^2 \log(d\mathbf{L})\left(\frac{\log m}{m} \vee \frac{\log n}{n}\right)}.$$

The results hold for very deep networks with depth $\mathbf{L} = o(\sqrt{n/\log n})$.

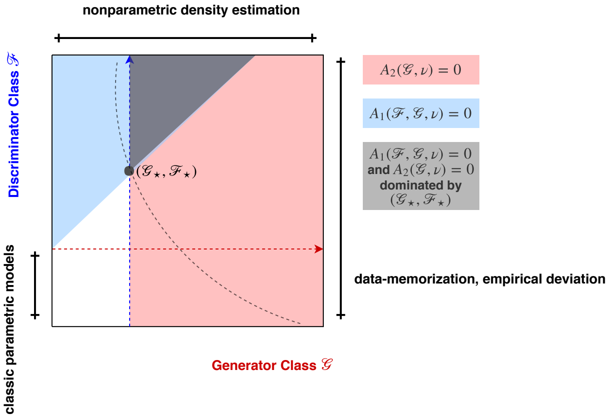
Application II: learning multivariate Gaussian

Corollary (L. '18, Gaussian)**.**

Consider $\nu \sim N(\mu, \Sigma)$. GANs enjoy near optimal sampling complexity (w.r.t. dim. d), with proper choices of the architecture and activation,

$$\mathbb{E}\, d_{TV}^2 \left(\nu, (g_{\widehat{\theta}_{m,n}})_{\#} \mu \right) \precsim \sqrt{\frac{d^2 \log d}{n \wedge m}}.$$

PAIR REGULARIZATION: WHY GANS MIGHT BE BETTER



Optimization
(local convergence)

FORMULATION

Generator g_θ, Discriminator f_ω

$$U(\theta, \omega) = \underbrace{\mathbb{E}_{Y \sim \nu}}_{\text{target}} [h_1 \circ f_\omega(Y)] - \underbrace{\mathbb{E}_{Z \sim \mu}}_{\text{input}} [h_2 \circ f_\omega(g_\theta(Z))]$$

$$\min_\theta \max_\omega U(\theta, \omega)$$

- global optimization for general $U(\theta, \omega)$ is hard Singh et al. (2000); Pfau and Vinyals (2016); Salimans et al. (2016)

FORMULATION

Generator g_θ, Discriminator f_ω

$$U(\theta,\omega) = \underbrace{\mathbb{E}_{Y\sim\nu}}_{\text{target}}[h_1 \circ f_\omega(Y)] - \underbrace{\mathbb{E}_{Z\sim\mu}}_{\text{input}}[h_2 \circ f_\omega(g_\theta(Z))]$$

$$\min_\theta \max_\omega \; U(\theta,\omega)$$

- global optimization for general $U(\theta,\omega)$ is hard Singh et al. (2000); Pfau and Vinyals (2016); Salimans et al. (2016)

Local saddle point (θ_*,ω_*) such that no incentive to deviate locally

$$U(\theta_*,\omega) \le U(\theta_*,\omega_*) \le U(\theta,\omega_*) \;\;,$$

for (θ,ω) in an open neighborhood of (θ_*,ω_*).

- also called local Nash Equilibrium (NE)
- modest goal: initialized properly, algorithm converges to a local NE

INTERACTION MATTERS: $\frac{\partial^2}{\partial \theta \partial \omega} U(\theta, \omega)$

Geometrically fast local convergence to **stable equilibrium**
However, "interaction term" matters, slows down the convergence $\Leftarrow$ curse

INTERACTION MATTERS: $\frac{\partial^2}{\partial\theta\partial\omega}U(\theta, \omega)$

Geometrically fast local convergence to **stable equilibrium**
However, "interaction term" matters, slows down the convergence $\Leftarrow$ curse

Unstable equilibrium? turns out "interaction term" matters, utilize it renders geometrically fast convergence $\Leftarrow$ blessing

Motivation for: *optimistic mirror descent*, *extra-gradients*, *negative-momentum* . . .

"However, no guarantees are known beyond the convex-concave setting and, more importantly for the paper, even in convex-concave games, no guarantees are known for the last-iterate pair."

— Daskalakis, Ilyas, Syrgkanis, and Zeng (2017)

Geometrically fast convergence to unstable equilibrium

OMD proposed in Daskalakis et al. (2017)

$$\theta_{t+1} = \theta_t - 2\eta\nabla_\theta U(\theta_t, \omega_t) + \boxed{\eta\nabla_\theta U(\theta_{t-1}, \omega_{t-1})}$$

$$\omega_{t+1} = \omega_t + 2\eta\nabla_\omega U(\theta_t, \omega_t) - \boxed{\eta\nabla_\omega U(\theta_{t-1}, \omega_{t-1})}$$

Rakhlin and Sridharan (2013)

For bi-linear game $U(\theta, \omega) = \theta^T C\omega$, to obtain ϵ-close solution

shown in Daskalakis et al. (2017) : $\quad T \gtrsim \boxed{\epsilon^{-4}\log\frac{1}{\epsilon}} \cdot \mathrm{Poly}\left(\frac{\lambda_{\max}(CC^T)}{\lambda_{\min}(CC^T)}\right)$

Geometrically fast convergence to unstable equilibrium

OMD proposed in Daskalakis et al. (2017)

$$\theta_{t+1} = \theta_t - 2\eta \nabla_\theta U(\theta_t, \omega_t) + \boxed{\eta \nabla_\theta U(\theta_{t-1}, \omega_{t-1})}$$

$$\omega_{t+1} = \omega_t + 2\eta \nabla_\omega U(\theta_t, \omega_t) - \boxed{\eta \nabla_\omega U(\theta_{t-1}, \omega_{t-1})}$$

Rakhlin and Sridharan (2013)

For bi-linear game $U(\theta, \omega) = \theta^T C \omega$, to obtain ϵ-close solution

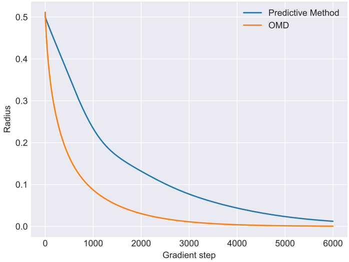
shown in Daskalakis et al. (2017) : $\quad T \gtrsim \boxed{\epsilon^{-4} \log \frac{1}{\epsilon}} \cdot \mathrm{Poly}\left(\frac{\lambda_{\max}(CC^T)}{\lambda_{\min}(CC^T)}\right)$

Theorem (L. & Stokes, '18)**.**

we proved : $\quad T \gtrsim \boxed{\log \frac{1}{\epsilon}} \cdot \frac{\lambda_{\max}(CC^T)}{\lambda_{\min}(CC^T)}$

further generalized beyond bi-linear game in Mokhtari et al. (2019).

Geometrically fast convergence to unstable equilibrium



Theorem (L. & Stokes, '18).

we proved : $$T \gtrsim \boxed{\log \frac{1}{\epsilon}} \cdot \frac{\lambda_{\mathsf{max}}(CC^T)}{\lambda_{\mathsf{min}}(CC^T)}$$

further generalized beyond bi-linear game in Mokhtari et al. (2019).

- **Statistical** $:-)$

  given n samples, what is the statistical/generalization error rate?

- $\boxed{\text{Approximation :-(}}$

  what dist. can be approximated by the generator $g_\theta(Z)$?

- $\boxed{\text{Computational :-O}}$

  local convergence for practical optimization, how to stablize?

- $\boxed{\text{Landscape :-(}}$

  are local saddle points good globally?

Other approach? theory of optimal transport $\Rightarrow$ GANs?

Optimal Transport

Wasserstein-p metric,

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

Optimal Transport

Wasserstein-p metric,

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

Theorem (Brenier, '87, $p = 2$)**.**

Let $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$. Let μ, ν absolutely continuous w.r.t. Lebesgue measure. There exists a unique **convex** $\psi_{\text{opt}} : \mathbb{R}^d \to \mathbb{R}$,

$$\frac{1}{2}W_2^2(\mu,\nu) = \inf_{\pi\in\Pi(\mu,\nu)} \int \frac{1}{2}\|x-y\|^2 d\pi$$

$$= \int \left(\frac{\|x\|^2}{2} - \psi_{\text{opt}}(x)\right)\mu(dx) + \int \left(\frac{\|y\|^2}{2} - \psi^\star_{\text{opt}}(y)\right)\nu(dy)$$

Here $\psi^\star(y) = \sup_y \{\langle y,x\rangle - \psi(x)\}$ is the Legendre-Fenchel conjugate of ψ.

Peyré et al. (2019)

Optimal Transport

$\boxed{\text{Approximation} \quad \text{:-)}}$ Consider $[0,1]^d$, $Z \sim \text{Unif}([0,1]^d)$, with a convex ψ
$(\nabla\psi)(Z)$ can represent distribution ν!

Theorem (Brenier, '87, $p = 2$)**.**

Let $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$. Let μ, ν absolutely continuous w.r.t. Lebesgue measure. There exists a unique **convex** $\psi_{\text{opt}} : \mathbb{R}^d \to \mathbb{R}$,

$$\frac{1}{2} W_2^2(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \int \frac{1}{2}\|x - y\|^2 d\pi$$

$$= \int \left(\frac{\|x\|^2}{2} - \psi_{\text{opt}}(x)\right)\mu(dx) + \int \left(\frac{\|y\|^2}{2} - \psi^\star_{\text{opt}}(y)\right)\nu(dy)$$

$$= \int \frac{1}{2}\|x - (\nabla\psi_{\text{opt}})(x)\|^2 \mu(dx), \qquad \boxed{\nu = (\nabla\psi_{\text{opt}})_\# \mu}$$

Peyré et al. (2019)

Optimal Transport

$\boxed{\text{Computation} \quad \text{:-)}}$

linear program, or smooth convex program simple $\boxed{\text{Landscape}}$

Recall input measure μ given, empirical target measure $\widehat{\nu}^n$

$$\frac{1}{2}W_2^2(\mu, \widehat{\nu}^n) = \sup_{\phi}\left\{\int \phi^c(x)\mu(dx) + \int \phi(y)\widehat{\nu}^n(dy)\right\}$$

where $\phi^c(x) := \inf_y\{\frac{1}{2}\|x-y\|^2 - \phi(y)\}$.

Genevay, Cuturi, Peyré, and Bach (2016)

Optimal Transport

Computation :-)

linear program, or smooth convex program simple Landscape

Add ϵ-entropic regularization

$$\frac{1}{2} W_{2,\epsilon}^2(\mu, \widehat{\nu}^n) = \sup_{\phi} \left\{ \int \phi_\epsilon^c(x)\mu(dx) + \int \phi(y)\widehat{\nu}^n(dy) \right\}$$

where $\phi_\epsilon^c(x) := -\epsilon \log \left[\int \exp\left(-\frac{\frac{1}{2}\|x-y\|^2 - \phi(y)}{\epsilon} \right) \widehat{\nu}^n(dy) \right]$.

On data $y_1, \ldots, y_n$
optimization reduces to SGD on $[\phi(y_1), \ldots, \phi(y_n)] \in \mathbb{R}^n$

Genevay, Cuturi, Peyré, and Bach (2016)

varying ϵ, solving $W_{2,\epsilon}^2(\mu, \widehat{\nu}^n)$ induced transportation map

$$(Id - \nabla\phi^c)(x) = \frac{\sum_{i=1}^n y_i \exp\left(-\frac{\frac{1}{2}\|x-y_i\|^2 - \phi(y_i)}{\epsilon}\right)}{\sum_{i=1}^n \exp\left(-\frac{\frac{1}{2}\|x-y_i\|^2 - \phi(y_i)}{\epsilon}\right)}$$

On data $y_1, \ldots, y_n$
optimization reduces to SGD on $[\phi(y_1), \ldots, \phi(y_n)] \in \mathbb{R}^n$

Optimal Transport and Pair Regularization

Recall input measure μ given, empirical target measure $\widehat{\nu}^n$

$$\frac{1}{2}W_2^2(\mu, \widehat{\nu}^n) = \sup_{\phi}\left\{\int \phi^c(x)\mu(dx) + \int \phi(y)\widehat{\nu}^n(dy)\right\}$$

where $\phi^c(x) := \inf_y\{\frac{1}{2}\|x-y\|^2 - \phi(y)\}$.

Analogy to GANs:

$$\phi : \mathbb{R}^d \to \mathbb{R} \text{ as discriminator function}$$

$$Id - \nabla\phi^c : \mathbb{R}^d \to \mathbb{R}^d \text{ as generator transformation}$$

Optimal Transport and Pair Regularization

Recall input measure μ given, empirical target measure $\widehat{\nu}^n$

$$\frac{1}{2}W_2^2(\mu, \widehat{\nu}^n) = \sup_{\phi}\left\{\int \phi^c(x)\mu(dx) + \int \phi(y)\widehat{\nu}^n(dy)\right\}$$

where $\phi^c(x) := \inf_y\{\frac{1}{2}\|x-y\|^2 - \phi(y)\}$.

Analogy to GANs:

$$\phi : \mathbb{R}^d \to \mathbb{R} \text{ as discriminator function}$$

$$Id - \nabla\phi^c : \mathbb{R}^d \to \mathbb{R}^d \text{ as generator transformation}$$

However, $(Id - \nabla\phi^c)_{\#}\mu = \widehat{\nu}^n$ data memorization

$$W_2\left((Id - \nabla\phi^c)_{\#}\mu, \nu\right) = W_2\left(\widehat{\nu}^n, \nu\right) \asymp n^{-\frac{1}{d}}$$

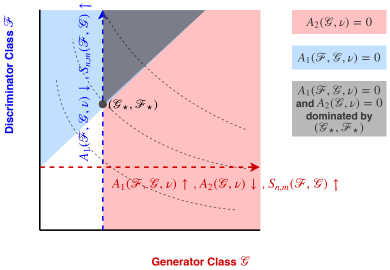
Pair Regularization, again

Analogy to GANs:

$$\phi : \mathbb{R}^d \to \mathbb{R} \text{ as discriminator function}$$

$$Id - \nabla \phi^c : \mathbb{R}^d \to \mathbb{R}^d \text{ as generator transformation}$$

Solution: **pair regularization** $\mathcal{F}_* = \{\phi, \text{ regular}\}$, $\mathcal{G}_* = \{Id - \nabla \phi^c, \text{ regular}\}$ for better **statistical rate**



Estimating Transportation Cost

Another Application of Pair Regularization

Regularity in OT Caffarelli (1992, 1991): $\mu, \nu \in \mathsf{C}^{\alpha}$ Hölder.

Statistical question: estimate "transportation cost" $W_2^2(\mu, \nu)$ based on n-i.i.d. samples $y_1, \ldots, y_n \sim \nu$. Suppose $\mu \sim \mathrm{Unif}([0,1]^d)$ known.

Lemma (L. & Sadhanala, '19).

$$\sup_{\nu \in \mathsf{C}^{\alpha}} \mathbb{E}\,|\widetilde{W}_n - W_2^2(\mu, \nu)| \precsim n^{-\frac{2\alpha+2}{2\alpha+d}} + n^{-\frac{1}{2}}$$

Another Application of Pair Regularization

Regularity in OT Caffarelli (1992, 1991): $\mu, \nu \in \mathsf{C}^{\alpha}$ Hölder.

Statistical question: estimate "transportation cost" $W_2^2(\mu, \nu)$ based on n-i.i.d. samples $y_1, \ldots, y_n \sim \nu$. Suppose $\mu \sim \mathrm{Unif}([0,1]^d)$ known.

Lemma (L. & Sadhanala, '19)**.**

$$\sup_{\nu \in \mathsf{C}^{\alpha}} \mathbb{E}\,|\widetilde{W}_n - W_2^2(\mu, \nu)| \precsim n^{-\frac{2\alpha+2}{2\alpha+d}} + n^{-\frac{1}{2}}$$

Elbow phenomenon: $\alpha \geq \frac{d}{2} - 2$, one gets parametric rate

Another Application of Pair Regularization

Regularity in OT Caffarelli (1992, 1991): $\mu, \nu \in \mathsf{C}^{\alpha}$ Hölder.

Statistical question: estimate "transportation cost" $W_2^2(\mu, \nu)$ based on n-i.i.d. samples $y_1, \ldots, y_n \sim \nu$. Suppose $\mu \sim \mathrm{Unif}([0,1]^d)$ known.

Lemma (L. & Sadhanala, '19)**.**

$$\sup_{\nu \in \mathsf{C}^{\alpha}} \mathbb{E}|\widetilde{W}_n - W_2^2(\mu, \nu)| \lesssim n^{-\frac{2\alpha+2}{2\alpha+d}} + n^{-\frac{1}{2}}$$

Pair regularization: $\phi \in \mathsf{C}^{\alpha+2}$, $Id - \nabla \phi^c \in \mathsf{C}^{\alpha+1}$, by Caffarelli (1992, 1991)

Another Application of Pair Regularization

Regularity in OT Caffarelli (1992, 1991): $\mu, \nu \in \mathsf{C}^{\alpha}$ Hölder.

Statistical question: estimate "transportation cost" $W_2^2(\mu,\nu)$ based on n-i.i.d. samples $y_1, \ldots, y_n \sim \nu$. Suppose $\mu \sim \mathrm{Unif}([0,1]^d)$ known.

Lemma (L. & Sadhanala, '19).

$$\sup_{\nu \in \mathsf{C}^{\alpha}} \mathbb{E}\,|\widetilde{W}_n - W_2^2(\mu,\nu)| \lesssim n^{-\frac{2\alpha+2}{2\alpha+d}} + n^{-\frac{1}{2}}$$

typically an easier problem than estimating measure under W_2, or estimating transportation map T under metric $\mathbb{E}_{X\sim\mu}\,\|\widehat{T}(X) - T(X)\|^2$

Hütter and Rigollet (2019)

Back to the adversarial framework

Two related problems

Estimate under the metric/loss

Theorem (L.,'17).

$$\inf_{\tilde{\nu}_n} \sup_{\nu \in \mathcal{G}} \mathbb{E}\, d^2_{\mathcal{F}}(\nu, \tilde{\nu}_n) \asymp n^{-\frac{2\alpha+2\beta}{2\alpha+d}} \vee n^{-1}$$

$\mathcal{G} = \mathrm{W}^{\alpha}$, $\mathcal{F} = \mathrm{W}^{\beta}$

No elbow phenomenon on α.

Liang (2017); Singh et al. (2018); Weed and Berthet (2019)

Back to the adversarial framework

Two related problems

Estimate under the metric/loss

Theorem (L.,'17).

$$\inf_{\widetilde{\nu}_n} \sup_{\nu \in \mathcal{G}} \mathbb{E}\, d^2_{\mathcal{F}}(\nu, \widetilde{\nu}_n) \asymp n^{-\frac{2\alpha+2\beta}{2\alpha+d}} \vee n^{-1}$$

$\mathcal{G} = \mathsf{W}^{\alpha}$, $\mathcal{F} = \mathsf{W}^{\beta}$

No elbow phenomenon on α.

Liang (2017); Singh et al. (2018); Weed and Berthet (2019)

Estimating the metric/loss itself

Theorem (L. & Sadhanala, '19).

$$\inf_{\widetilde{W}_n} \sup_{\nu \in \mathcal{G}} \mathbb{E}\, |\widetilde{W}_n - d^2_{\mathcal{F}}(\mu, \nu)|^2 \asymp n^{-\frac{8\alpha+8\beta}{4\alpha+d}} \vee n^{-1}$$

$\mathcal{G} = \mathsf{W}^{\alpha}$, $\mathcal{F} = \mathsf{W}^{\beta}$

Elbow phenomenon on $\alpha = d/4 - 2\beta$.

typically an easier problem.

However, for Wasserstein metric

Theorem (L., '19).

Consider $d \geq 2$ and the domain $\Omega = [0,1]^d$. Given n i.i.d. samples $y_1, \ldots, y_n$ from ν,

$$\inf_{\widetilde{W}_n} \sup_{\nu \in \mathsf{C}^{\alpha}} \mathbb{E}|\widetilde{W}_n - W_1(\mu, \nu)| \precsim n^{-\frac{\alpha+1}{2\alpha+d}} \ ,$$

as we know

$$\inf_{\widetilde{\nu}_n} \sup_{\nu \in \mathsf{C}^{\alpha}} \mathbb{E}\, W(\widetilde{\nu}_n, \nu) \asymp n^{-\frac{\alpha+1}{2\alpha+d}} .$$

However, for Wasserstein metric

Theorem (L., '19).

Consider $d \geq 2$ and the domain $\Omega = [0,1]^d$. Given n i.i.d. samples $y_1, \ldots, y_n$ from ν,

$$\frac{\log\log(n)}{\log(n)} \cdot n^{-\frac{\alpha+1}{2\alpha+d}} \precsim \inf_{\widetilde{W}_n} \sup_{\nu \in C^{\alpha}} \mathbb{E}|\widetilde{W}_n - W_1(\mu,\nu)| \precsim n^{-\frac{\alpha+1}{2\alpha+d}} \quad ,$$

as we know

$$\inf_{\widetilde{\nu}_n} \sup_{\nu \in C^{\alpha}} \mathbb{E}\, W(\widetilde{\nu}_n, \nu) \asymp n^{-\frac{\alpha+1}{2\alpha+d}}.$$

estimating the Wasserstein-1 metric itself
is **almost as hard** as
estimating under the Wasserstein-1 metric

HOWEVER, FOR WASSERSTEIN METRIC

Theorem (L., '19).

Consider $d \geq 2$ and the domain $\Omega = [0,1]^d$. Given n i.i.d. samples $y_1, \ldots, y_n$ from ν,

$$\frac{\log\log(n)}{\log(n)} \cdot n^{-\frac{\alpha+1}{2\alpha+d}} \lesssim \inf_{\widetilde{W}_n} \sup_{\nu \in \mathsf{C}^{\alpha}} \mathbb{E}\,|\widetilde{W}_n - W_1(\mu,\nu)| \lesssim n^{-\frac{\alpha+1}{2\alpha+d}} \quad ,$$

as we know

$$\inf_{\widetilde{\nu}_n} \sup_{\nu \in \mathsf{C}^{\alpha}} \mathbb{E}\, W(\widetilde{\nu}_n, \nu) \asymp n^{-\frac{\alpha+1}{2\alpha+d}} .$$

- the main technicality is in deriving the lower bound: wavelets
- construct two composite/fuzzy hypotheses using delicate priors with matching $\log n$ moments
- and the Wasserstein metric differs sufficiently
- calculate total variation metric directly on the posterior of data (sum-product form), via a telescoping trick

SUMMARY

- In this talk, we study **statistical rates** for $d(\widehat{T}_{\#}\mu, \nu)$ and $\widehat{d(\mu, \nu)}$, with $\nu = T^{\star}_{\#}\mu$.

  Implicit Distribution Estimation motivated from GANs, OT.

- Conceptually, to learn the distribution via transformation/transportation, vs., to estimate the transformation/transportation difficulty.

- Closely related problems in the lens of Optimal Transport.

$$\text{harder} \quad d(\widehat{T}_{\#}\mu, \nu) \quad \underset{\text{sometimes induces a transportation map}}{\overset{\text{induces plug-in estimate}}{\rightleftarrows}} \quad \widehat{d(\mu, \nu)} \quad \text{easier}$$

- Idea of pair regularization

  what GANs have over classical nonparametrics.

Summary

- In this talk, we study **statistical rates** for $d(\widehat{T}_{\#}\mu, \nu)$ and $\widehat{d(\mu, \nu)}$, with $\nu = T^{\star}_{\#}\mu$.

  Implicit Distribution Estimation motivated from GANs, OT.

- Conceptually, to learn the distribution via transformation/transportation, vs., to estimate the transformation/transportation difficulty.

- Closely related problems in the lens of Optimal Transport.

$$\text{harder} \quad d(\widehat{T}_{\#}\mu, \nu) \; \underset{\text{sometimes induces a transportation map}}{\overset{\text{induces plug-in estimate}}{\rightleftarrows}} \; \widehat{d(\mu, \nu)} \quad \text{easier}$$

- Idea of pair regularization

  what GANs have over classical nonparametrics.

Many interesting open problems both **statistically** and **computationally**, with new insights on **regularization** and **adaptivity**.

Thank you!

Liang, T. (2018). — On How Well Generative Adversarial Networks Learn Densities: Nonparametric and Parametric Results. *arXiv:1811.03179* *under review*

Liang, T. & Stokes, J. (2018). — Interaction Matters: A Note on Non-asymptotic Local Convergence of Generative Adversarial Networks. *arXiv:1802.06132* *AISTATS 2019*

Liang, T. (2019). — On the Minimax Optimality of Estimating the Wasserstein Metric. *arXiv:1908:10324*

Liang, T. & Sadhanala, V. (2019). — Working Paper.