GANs, Optimal Transport, and Implicit Distribution Estimation

Tengyuan Liang

Econometrics and Statistics

The University of Chicago Booth School of Business
Tjalling C. Koopmans (1910-1985)


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
Implicit Distribution Estimation

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

$T(Z)$ close in dist.? $Y$
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).

equivalently

$$T(Z) \approx Y$$

$$T \# \mu \approx \nu$$
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, Binkowski, 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)
GENERICATIVE ADVERSARIAL NETWORKS

Generator $g_{\theta}$, Discriminator $f_{\omega}$

$U(\theta, \omega) = \mathbb{E}_{Y \sim \nu}/d_{\text{curly.alt2}}^{\text{target}} [f_{\omega}(Y)] - \mathbb{E}_{Z \sim \mu}/d_{\text{curly.alt2}}^{\text{input}} [f_{\omega}(g_{\theta}(Z))]$

GANs are widely used in practice, however...
**GENERATIVE ADVERSARIAL NETWORKS**

Generative adversarial networks (conceptual)

**Generator** $g_\theta$, **Discriminator** $f_\omega$

$$U(\theta, \omega) = \max_{\omega} \min_{\theta} \left[ \mathbb{E}_{Y \sim \nu} [f_\omega(Y)] - \mathbb{E}_{Z \sim \mu} [f_\omega(g_\theta(Z))] \right]$$

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/generlization error rate**?

- **Computational:**
  local convergence for practical optimization, how to stabilize?

- **Landscape:**
  are local saddle points good globally?
**FORMULATION**

\[ T_G \text{ class of } \textbf{generator} \text{ transformations, } F_D \text{ class of } \textbf{discriminator} \text{ functions} \]

\[ \nu \text{ target dist.} \]

population \[ g^* \in \arg \min_{g \in T_G} \max_{f \in F_D} \left\{ \mathbb{E}_{X \sim g \# \mu} [f(X)] - \mathbb{E}_{Y \sim \nu} [f(Y)] \right\} \]
FORMULATION

\(\mathcal{T}_G\) class of **generator** transformations, \(\mathcal{F}_D\) class of **discriminator** functions

\(\nu\) target dist.

**population**

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

\(\tilde{\nu}^n\) empirical dist.

**empirical**

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

\(\tilde{g} \# \mu\) as estimate for \(\nu\)
FORMULATION

\( \mathcal{T}_G \) class of **generator** transformations, \( \mathcal{F}_D \) class of **discriminator** functions

\( \nu \) target dist.

\[ \begin{align*}
\text{population} & \quad g^* \in \arg \min_{g \in \mathcal{T}_G} \max_{f \in \mathcal{F}_D} \left\{ \mathbb{E}_{X \sim g \# \mu} [f(X)] - \mathbb{E}_{Y \sim \nu} [f(Y)] \right\} \\
\text{empirical} & \quad \hat{g} \in \arg \min_{g \in \mathcal{T}_G} \max_{f \in \mathcal{F}_D} \left\{ \mathbb{E}_{X \sim \hat{g} \# \mu} [f(X)] - \mathbb{E}_{Y \sim \hat{\nu}^n} [f(Y)] \right\} \\
& \quad \hat{g} \# \mu \text{ as estimate for } \nu
\end{align*} \]

- 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)

\[
    d_{\mathcal{F}}(\mu, \nu) := \sup_{f \in \mathcal{F}} \left| \mathbb{E}_{X \sim \mu} f(X) - \mathbb{E}_{Y \sim \nu} f(Y) \right|
\]
**Discriminator metric**

Define the critic metric (IPM)

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

- \(\mathcal{F}\) 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 \(\beta\): 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 \(\nu\) with certain regularity.
## Summary of First Half of Talk

<table>
<thead>
<tr>
<th>Goal</th>
<th>Evaluation Metric</th>
<th>Results</th>
<th>Generator Class $\mathcal{G}$</th>
<th>Discriminator Class $\mathcal{F}$</th>
<th>Property</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adversarial Framework (nonparametric)</td>
<td>$d_\mathcal{F}$</td>
<td>$\inf_d$</td>
<td>Sobolev $W^\alpha$</td>
<td>Sobolev $W^\beta$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MMD GAN</td>
<td>$\sup_d$</td>
<td>smooth subspace in RKHS $\mathcal{H}$</td>
<td>RKHS $\mathcal{H}$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>oracle results</td>
<td>any</td>
<td>Sobolev $W^\beta$</td>
<td>$\mathcal{G}^+$</td>
<td></td>
</tr>
<tr>
<td>Generative Adversarial Networks (parametric)</td>
<td>$d_{TV}$</td>
<td>$\inf_d$</td>
<td>leaky-ReLU $\mathcal{G}$</td>
<td>leaky-ReLU $\mathcal{F}$</td>
<td>$\mathcal{F}^+, m^*$</td>
</tr>
<tr>
<td></td>
<td>leaky-ReLU GANs</td>
<td>$\sup_d$</td>
<td>leaky-ReLU $\mathcal{G}$</td>
<td>leaky-ReLU $\mathcal{F}$</td>
<td>$\mathcal{F}^+, m^*$</td>
</tr>
<tr>
<td></td>
<td>any GANs</td>
<td>oracle results</td>
<td>neural networks</td>
<td>neural networks</td>
<td>$\mathcal{G}^+, \mathcal{F}^+, m^*$</td>
</tr>
<tr>
<td></td>
<td>Lipschitz GANs</td>
<td>oracle results</td>
<td>Lipschitz neural networks</td>
<td>Lipschitz neural networks</td>
<td>$\mathcal{G}^+, \mathcal{F}^+, m^*$</td>
</tr>
</tbody>
</table>
## Summary of First Half of Talk

<table>
<thead>
<tr>
<th>Goal</th>
<th>Evaluation Metric</th>
<th>Results</th>
<th>Generator Class $\mathcal{G}$</th>
<th>Discriminator Class $\mathcal{F}$</th>
<th>Property</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adversarial Framework</td>
<td>$d_{\mathcal{F}}$</td>
<td>Sobolev minimax</td>
<td>Sobolev $W^\alpha$</td>
<td>Sobolev $W^\beta$</td>
<td></td>
</tr>
<tr>
<td>(nonparametric)</td>
<td></td>
<td>GAN upper optimal</td>
<td></td>
<td>smooth subspace in RKHS $\mathcal{H}$</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>oracle results</td>
<td></td>
<td>any Sobolev $W^\beta$</td>
<td>$\mathcal{G}^\dagger$</td>
</tr>
<tr>
<td>Generative Adversarial Networks</td>
<td>$d_{TV}$</td>
<td>leaky-ReLU upper bound</td>
<td>leaky-ReLU</td>
<td>leaky-ReLU</td>
<td>$\mathcal{F}^\ddagger, m^\star$</td>
</tr>
<tr>
<td>(parametric)</td>
<td></td>
<td>GANs</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$d_{TV}, d_{KL}, d_H$</td>
<td>any GANs oracle results</td>
<td>neural networks</td>
<td>neural networks</td>
<td>$\mathcal{G}^\dagger, \mathcal{F}^\ddagger, m^\star$</td>
</tr>
<tr>
<td></td>
<td>$d_W$</td>
<td>Lipschitz GANs</td>
<td>Lipschitz neural networks</td>
<td>Lipschitz neural networks</td>
<td>$\mathcal{G}^\dagger, \mathcal{F}^\ddagger, m^\star$</td>
</tr>
</tbody>
</table>

The symbols: $(\mathcal{G}^\dagger)$ and $(\mathcal{F}^\ddagger)$ to denote the mis-specification for the generator class and the discriminator class respectively, and $(m^\star)$ 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 $G := \{ \nu : \rho_\nu \in W^\alpha \}$ Sobolev space with smoothness $\alpha$, and the evaluation metric $F = W^\beta$ with smoothness $\beta$. 
Consider the target $\mathcal{G} := \{\nu : \rho_\nu \in W^\alpha\}$ Sobolev space with smoothness $\alpha$, and the evaluation metric $\mathcal{F} = W^\beta$ with smoothness $\beta$.

The minimax optimal rate is

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

Here $\tilde{\nu}_n$ 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 $\rho_{\nu}$ in $\mathcal{G} = \{ \nu \mid \mathcal{T}^{-\frac{\alpha-1}{2}} \rho_{\nu} \|_{\mathcal{H}} \leq 1 \}$ with smoothness $\alpha$
**Minimax optimal rates: MMD GAN**

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

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

**Theorem (L. ’18, RKHS).**

The minimax optimal rate is

\[
\inf_{\tilde{\nu}_n} \sup_{\nu \in \mathcal{G}} \mathbb{E} d_{\mathcal{F}} (\nu, \tilde{\nu}_n) \approx n^{-(\alpha+1)\frac{\kappa}{2\alpha\kappa+2}} \lor n^{-\frac{1}{2}}
\]
**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 $\rho_\nu$ in $\mathcal{G} = \{ \nu \mid \|\mathcal{T}^{-\frac{\alpha - 1}{2}} \rho_\nu\|_{\mathcal{H}} \leq 1 \}$ with smoothness $\alpha$

**Theorem (L. ’18, RKHS).**

The minimax optimal rate is

$$\inf_{\tilde{\nu}_n} \sup_{\nu \in \mathcal{G}} \mathbb{E} d_{\mathcal{F}} (\nu, \tilde{\nu}_n) \lesssim n^{-\left(\frac{\alpha + 1}{2}\kappa + \frac{1}{2}\right)} \lor n^{-\frac{1}{2}}.$$  

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

$\kappa < 1$: sample complexity scales $n = e^{2+\frac{2}{\alpha+1}}\left(\frac{1}{\kappa} - 1\right)$, effective dim. $\frac{1}{\kappa}$.
**ORACLE INEQUALITY FOR GANs**

*Generator class $G$ may not contain the target $\nu$: oracle approach.*
**Oracle Inequality for GANs**

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

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

$\hat{g} \# \mu, \tilde{g} \# \mu$ are Implicit Density Estimators!

**Corollary (L. '17).**

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

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

attains a sub-optimal rate

$$\mathbb{E} d_{\mathcal{F}_D}(\hat{g} \# \mu, \nu) \leq \min_{g \in \mathcal{T}_G} d_{\mathcal{F}_D}(g \# \mu, \nu) + n^{-\frac{\beta}{d}} \sqrt{\frac{\log n}{n}}.$$

Canas and Rosasco (2012): $\beta = 1$
Oracle inequality for GANs

Generator class $G$ may not contain the target $\nu$: oracle approach.

Let $T_G$ be any generator transformation. The discriminator metric $F_D = W^\beta$, target density $\rho_\nu \in W^\alpha$.

Corollary (L. ’17).

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

$$\tilde{g} \in \arg\min_{g \in T_G} \max_{f \in F_D} \left\{ \mathbb{E}_{X \sim g^\# \mu} [f(X)] - \mathbb{E}_{Y \sim \tilde{\nu}^n} [f(Y)] \right\},$$

a faster rate is attainable

$$\mathbb{E} d_{F_D} (\tilde{g}^\# \mu, \nu) \leq \min_{g \in T_G} d_{F_D} (g^\# \mu, \nu) + n^{-\alpha + \beta \over 2\alpha + d} \vee \frac{1}{\sqrt{n}}.$$

Canas and Rosasco (2012): $\beta = 1$
**SUB-OPTIMALITY AND REGULARIZATION**

Regularization helps achieve faster rate

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

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

\[ \hat{\theta}_{m,n} \in \arg \min_{\theta : g_0 \in G} \max_{\omega : f_\omega \in F} \{ \mathbb{E}_m f_\omega (g_0(Z)) - \mathbb{E}_n f_\omega (Y) \} , \]

with \( m \) generator samples and \( n \) target samples.

How well GANs learn the distribution, under objective evaluation metric, say \( d_{TV} \)?
**GENERALIZED ORACLE INEQUALITY**

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

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

Pdim(·) the pseudo-dimension of the neural network function.
**GENERALIZED ORACLE INEQUALITY**

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

sto. err. \[ S_{n,m}(\mathcal{F}, \mathcal{G}) := \sqrt{\text{Pdim}(\mathcal{F}) \frac{\log(m \wedge n)}{m \wedge n}} + \sqrt{\text{Pdim}(\mathcal{F} \circ \mathcal{G}) \frac{\log(m)}{m}}, \]

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

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

\[
\mathbb{E} d^2_{TV}(\nu, (g_{\theta_{m,n}}) \# \mu), \mathbb{E} d^2_W(\nu, (g_{\theta_{m,n}}) \# \mu),
\]

\[
\mathbb{E} d_{KL}(\nu || (g_{\theta_{m,n}}) \# \mu) + \mathbb{E} d_{KL}((g_{\theta_{m,n}}) \# \mu || \nu)
\]

\[
\leq A_1(\mathcal{F}, \mathcal{G}, \nu) + A_2(\mathcal{G}, \nu) + S_{n,m}(\mathcal{F}, \mathcal{G}) .
\]
**Generalized Oracle Inequality**

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

sto. err. \[ S_{n,m}(F, G) := \sqrt{\text{Pdim}(F) \frac{\log(m \wedge n)}{m \wedge n}} + \sqrt{\text{Pdim}(F \circ G) \frac{\log(m)}{m}}, \]

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

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

\[
\mathbb{E} d_{TV}^2(\nu, (g_{\Theta_{m,n}})_\# \mu), \mathbb{E} d_{W}^2(\nu, (g_{\Theta_{m,n}})_\# \mu), \\
\mathbb{E} d_{KL}(\nu \| (g_{\Theta_{m,n}})_\# \mu) + \mathbb{E} d_{KL}((g_{\Theta_{m,n}})_\# \mu \| \nu) \\
\leq A_1(F, G, \nu) + A_2(G, \nu) + S_{n,m}(F, G). 
\]

We emphasize on the interplay between \((G, F)\) as a pair of tuning parameters for regularization.
approx. err. $A_1(\mathcal{F}, \mathcal{G}, \nu) := \sup_{\theta} \inf_\omega \left\| \frac{\sqrt{\rho_{\nu}} - \sqrt{\rho_{\mu_\theta}}}{\sqrt{\rho_{\nu}} + \sqrt{\rho_{\mu_\theta}}} - f_\omega \right\|$, 

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

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

\[ \mathbb{E} d^2_{TV} \left( \nu, (\mathcal{G}_{\theta_{m,n}}) \# \mu \right), \mathbb{E} d^2_H \left( \nu, (\mathcal{G}_{\theta_{m,n}}) \# \mu \right), \]

\[ \leq A_1(\mathcal{F}, \mathcal{G}, \nu) + A_2(\mathcal{G}, \nu) + S_{n,m}(\mathcal{F}, \mathcal{G}) . \]

similar result for Hellinger $d_H$, for non-absolutely continuous $(\mathcal{G}_\theta) \# \mu$ and $\nu$. 
**PAIR REGULARIZATION**

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

fix $F$, as $G$ increase: $A_1(F, G, \nu)$ increase, $A_2(G, \nu)$ decrease, $S_{n,m}(F, G)$ increase.
Applications of pair regularization
APPLICATION I: PARAMETRIC RATES FOR LEAKY ReLU NETWORKS

When the generator $G$ 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.

$$\log \rho_{(g_0)_\#\mu}(x) = c_1 \sum_{l=1}^{L-1} \sum_{i=1}^{d} \mathbf{1}_{m_{li}(x) \geq 0} + c_0,$$

Bai et al. (2018)
APPLICATION I: PARAMETRIC RATES FOR LEAKY ReLU NETWORKS

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

**Theorem (L. ’18, leaky ReLU).**

$$
\mathbb{E} d_{TV}^2 \left( \nu, (\tilde{g}_{\tilde{\theta}_{m,n}}) \# \mu \right) \lesssim \sqrt{d^2 L^2 \log(dL) \left( \frac{\log m}{m} \vee \frac{\log n}{n} \right)}.
$$

The results hold for very deep networks with depth $L = o(\sqrt{n/\log n})$. 
APPLICATION II: LEARNING MULTIVARIATE 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,

$$\mathbb{E} d_{TV}^2 (\nu, (g_{\tilde{\theta}_{m,n}}) \neq \mu) \lesssim \sqrt{\frac{d^2 \log d}{n \wedge m}}.$$
PAIR REGULARIZATION: WHY GANs MIGHT BE BETTER

nonparametric density estimation

\( A_2(G, \nu) = 0 \)
\( A_1(\mathcal{F}, G, \nu) = 0 \)
and \( A_2(G, \nu) = 0 \) dominated by \((G_*, \mathcal{F}_*)\)

data-memorization, empirical deviation

Generator Class \( G \)

Discriminator Class \( \mathcal{F} \)
classic parametric models
Optimization
(local convergence)
FORMULATION

Generator $g_{\theta}$, Discriminator $f_{\omega}$

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

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

- global optimization for general $U(\theta, \omega)$ is hard \cite{Singh2000, Pfau2016, Salimans2016}
**FORMULATION**

Generator $g_\theta$, Discriminator $f_\omega$

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

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

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

Local saddle point $(\theta_*, \omega_*)$ such that no incentive to deviate locally

$$U(\theta_*, \omega) \leq U(\theta_*, \omega_*) \leq U(\theta, \omega_*)$$

for $(\theta, \omega)$ in an open neighborhood of $(\theta_*, \omega_*)$.

- 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)

\[
\begin{align*}
\theta_{t+1} &= \theta_t - 2\eta \nabla_\theta U(\theta_t, \omega_t) + \eta \nabla_\theta U(\theta_{t-1}, \omega_{t-1}) \\
\omega_{t+1} &= \omega_t + 2\eta \nabla_\omega U(\theta_t, \omega_t) - \eta \nabla_\omega U(\theta_{t-1}, \omega_{t-1})
\end{align*}
\]

Rakhlin and Sridharan (2013)

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

shown in Daskalakis et al. (2017): \[
T \geq \epsilon^{-4} \log \frac{1}{\epsilon} \cdot \text{Poly} \left( \frac{\lambda_{\max}(CC^T)}{\lambda_{\min}(CC^T)} \right)
\]
**GEOMETRICALLY FAST CONVERGENCE TO UNSTABLE EQUILIBRIUM**

**OMD** proposed in Daskalakis et al. (2017)

\[
\begin{align*}
\theta_{t+1} &= \theta_t - 2\eta \nabla_{\theta} U(\theta_t, \omega_t) + \eta \nabla_{\theta} U(\theta_{t-1}, \omega_{t-1}) \\
\omega_{t+1} &= \omega_t + 2\eta \nabla_{\omega} U(\theta_t, \omega_t) - \eta \nabla_{\omega} U(\theta_{t-1}, \omega_{t-1})
\end{align*}
\]

Rakhlin and Sridharan (2013)

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

shown in Daskalakis et al. (2017):

\[
T \gtrsim \frac{1}{\epsilon^4} \log \frac{1}{\epsilon} \cdot \text{Poly} \left( \frac{\lambda_{\text{max}}(C^T C)}{\lambda_{\text{min}}(C^T C)} \right)
\]

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

we proved:

\[
T \gtrsim \log \frac{1}{\epsilon} \cdot \frac{\lambda_{\text{max}}(C^T C)}{\lambda_{\text{min}}(C^T C)}
\]

further generalized beyond bi-linear game in Mokhtari et al. (2019).
GEOMETRICALLY FAST CONVERGENCE TO UNSTABLE EQUILIBRIUM

**Theorem** (L. & Stokes, ’18).

\[ T \gg \log \frac{1}{\epsilon} \cdot \frac{\lambda_{\text{max}}(CCT)}{\lambda_{\text{min}}(CCT)} \]

we proved : 

Further generalized beyond bi-linear game in Mokhtari et al. (2019).
• **Statistical**  :  
  given $n$ samples, what is the statistical/generalization error rate?

• **Approximation**  :-(
  what dist. can be approximated by the generator $g_\theta(Z)$?

• **Computational**  :-O
  local convergence for practical optimization, how to stabilize?

• **Landscape**  :-(
  are local saddle points good globally?

Other approach? theory of optimal transport $\Rightarrow$ GANs?
**Optimal Transport**

Wasserstein-$p$ metric,

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

\(\Pi(\mu, \nu)\) all couplings

Theorem (Brenier, '87, \(p = 2\))

Peyré et al. (2019)
**Optimal Transport**

Wasserstein-$p$ metric,

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

$\Pi(\mu, \nu)$ all couplings

**Theorem (Brenier, ‘87, $p = 2$).**

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

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

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

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

Peyré et al. (2019)
**Optimal Transport**

Consider $[0, 1]^d, Z \sim \text{Unif}([0, 1]^d)$, with a convex $\psi$ 
$(\nabla \psi)(Z)$ can represent distribution $\nu$!

**Theorem** (Brenier, ’87, $p = 2$).

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

$$
\frac{1}{2} W_2^2(\mu, \nu) = \inf_{\pi \in \Pi(\mu, \nu)} \int \frac{1}{2} \| x - y \|^2 d\pi \\
= \int \left( \frac{\| x \|^2}{2} - \psi_{\text{opt}}(x) \right) \mu(dx) + \int \left( \frac{\| y \|^2}{2} - \psi^*_{\text{opt}}(y) \right) \nu(dy) \\
= \int \frac{1}{2} \| x - (\nabla \psi_{\text{opt}})(x) \|^2 \mu(dx), \quad \nu = (\nabla \psi_{\text{opt}}) \# \mu
$$

Peyré et al. (2019)
Optimal Transport

Recall input measure $\mu$ given, empirical target measure $\bar{\nu}^n$

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

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 $\epsilon$-entropic regularization

$$\frac{1}{2} W^2_{2, \epsilon} (\mu, \bar{\nu}^n) = \sup_{\phi} \left\{ \int \phi_c^\epsilon (x) \mu(dx) + \int \phi(y) \bar{\nu}^n(dy) \right\}$$

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

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 $\epsilon$, solving $W^2_{2,\epsilon}(\mu, \nu^n)$ induced transportation map

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

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 \( \mu \) given, empirical target measure \( \tilde{\nu}^n \)

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

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

Analogy to GANs:

\( \phi : \mathbb{R}^d \to \mathbb{R} \) as discriminator function

\( \text{Id} - \nabla \phi^c : \mathbb{R}^d \to \mathbb{R}^d \) as generator transformation
Optimal Transport and Pair Regularization

Recall input measure $\mu$ given, empirical target measure $\hat{\nu}^n$

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

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

Analogy to GANs:

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

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

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

$$W_2((Id - \nabla \phi^c)\# \mu, \nu) = W_2(\hat{\nu}^n, \nu) \asymp n^{-\frac{1}{d}}$$
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 \( \mathcal{F}_\star = \{ \phi, \text{ regular} \}, \mathcal{G}_\star = \{ Id - \nabla \phi^c, \text{ regular} \} \) for better statistical rate
Estimating Transportation Cost
Another Application of Pair Regularization

Regularity in OT Caffarelli (1992, 1991): \( \mu, \nu \in C^\alpha \) 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 \( \mu \sim \text{Unif}([0, 1]^d) \) known.

**Lemma** (L. & Sadhanala, ’19).

\[
\sup_{\nu \in C^\alpha} \mathbb{E} |\tilde{W}_n - W_2^2(\mu, \nu)| \lesssim n^{-\frac{2\alpha + 2}{2\alpha + d}} + n^{-\frac{1}{2}}
\]
ANOTHER APPLICATION OF PAIR REGULARIZATION

Regularity in OT Caffarelli (1992, 1991): $\mu, \nu \in C^\alpha$ 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 $\mu \sim \text{Unif}([0,1]^d)$ known.

Lemma (L. & Sadhanala, ’19).

$$\sup_{\nu \in C^\alpha} \mathbb{E} |\tilde{W}_n - W^2_2(\mu, \nu)| \lesssim n^{-\frac{2\alpha+2}{2\alpha+d}} + n^{-\frac{1}{2}}$$

Elbow phenomenon: $\alpha \geq \frac{d}{2} - 2$, one gets parametric rate
**Another Application of Pair Regularization**

Regularity in OT [Caffarelli (1992, 1991): \( \mu, \nu \in C^\alpha \) 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 \( \mu \sim \text{Unif}([0, 1]^d) \) known.

**Lemma (L. & Sadhanala, ’19).**

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

**Pair regularization:** \( \phi \in C^{\alpha + 2}, Id - \nabla \phi^c \in C^{\alpha + 1} \), by [Caffarelli (1992, 1991)]
Another Application of Pair Regularization

Regularity in OT Caffarelli (1992, 1991): $\mu, \nu \in C^\alpha$ 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 $\mu \sim \text{Unif}([0, 1]^d)$ known.

Lemma (L. & Sadhanala, ’19).

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

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

Hütter and Rigollet (2019)
BACK TO THE ADVERSARIAL FRAMEWORK

Two related problems

Estimate under the metric/loss

Theorem (L.,'17).

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

\[
\mathcal{G} = W^\alpha, \mathcal{F} = W^\beta
\]

No elbow phenomenon on \(\alpha\).

Liang (2017); Singh et al. (2018); Weed and Berthet (2019)
Back to the Adversarial Framework

Two related problems

Estimate under the metric/loss

Estimating the metric/loss itself

**Theorem (L., ’17).**

\[
\inf \sup \mathbb{E} d_{\mathcal{F}}^2 (\nu, \tilde{\nu}_n) \asymp n^{-\frac{2\alpha + 2\beta}{2\alpha + d}} \vee n^{-1}
\]

\[G = W^\alpha, \mathcal{F} = W^\beta\]

No elbow phenomenon on \(\alpha\).

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

**Theorem (L. & Sadhanala, ’19).**

\[
\inf \sup \mathbb{E} |\tilde{W}_n - d_{\mathcal{F}}^2 (\mu, \nu)|^2 \asymp n^{-\frac{8\alpha + 8\beta}{4\alpha + d}} \vee n^{-1}
\]

\[G = W^\alpha, \mathcal{F} = W^\beta\]

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 $\nu$, 

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

as we know

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

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

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

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 $\nu$,

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

as we know

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

- 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 $d(\hat{T}_# \mu, \nu)$ and $\hat{d}(\mu, \nu)$, with $\nu = T^* \mu$.

  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.

  induces plug-in estimate

  harder $d(\hat{T}_# \mu, \nu) \quad \overset{\text{induces}}{\longrightarrow} \quad \hat{d}(\mu, \nu)$ easier

  sometimes induces a transportation map

- Idea of **pair regularization**

  what GANs have over classical nonparametrics.
SUMMARY

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

  **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.

  \[ \text{harder} \quad d(\hat{T}_# \mu, \nu) \quad \xrightarrow{\text{induces plug-in estimate}} \quad \hat{d}(\mu, \nu) \quad \text{easier} \]

  sometimes induces a transportation map

• 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!

*arXiv:1811.03179 under review*

*arXiv:1802.06132 AISTATS 2019*

*arXiv:1908:10324*