# Regularization for Wasserstein Distributionally Robust Optimization

Waïss Azizian

PhD student under the supervision of Franck Iutzeler, Jérôme Malick and Panayotis Mertikopoulos

May 2022

# Outline

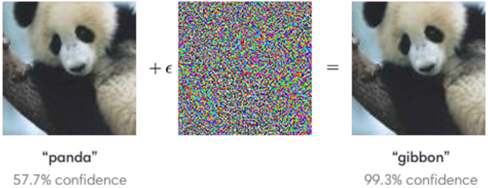
1. Quick introduction to WDRO
2. Regularizing WDRO
3. “Robust” generalization properties with WDRO

# Robust ML

We want ML models not to fail when applied in the real-world

## Shifts in distribution:

## Adversarial attacks: from (Goodfellow et al., 2015)



## Learning framework: from ERM to DRO

- Training data $\xi_1, \ldots, \xi_n \sim P_{train}$, where $P_{train}$ unknown, belgonging to $\Xi \subset \mathbb{R}^d$
  e.g., $\xi_i = (x_i, y_i)$ where $x_i$ input, $y_i$ label/target
- Objective $f_\theta : \Xi \to \mathbb{R}$, parameterized by $\theta$
  e.g., logistic regression $f_\theta(\xi) = f_\theta((x, y)) = \log\left(1 + e^{-y\langle\theta, x\rangle}\right)$
- Empirical Risk Minimization (ERM)

$$\min_\theta \frac{1}{n} \sum_{i=1}^{n} f_\theta(\xi_i)$$

# Learning framework: from ERM to DRO

- Training data $\xi_1, \ldots, \xi_n \sim P_{train}$, where $P_{train}$ unknown, belgonging to $\Xi \subset \mathbb{R}^d$
  e.g., $\xi_i = (x_i, y_i)$ where $x_i$ input, $y_i$ label/target

- Objective $f_\theta : \Xi \to \mathbb{R}$, parameterized by $\theta$
  e.g., logistic regression $f_\theta(\xi) = f_\theta((x, y)) = \log\left(1 + e^{-y\langle\theta, x\rangle}\right)$

- Empirical Risk Minimization (ERM)

$$\min_\theta \frac{1}{n} \sum_{i=1}^{n} f_\theta(\xi_i) = \mathbb{E}_{\xi \sim \hat{P}_n} f_\theta(\xi) \quad \text{with } \hat{P}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{\xi_i}$$

# Learning framework: from ERM to DRO

- Training data $\xi_1, \dots, \xi_n \sim P_{train}$, where $P_{train}$ unknown, belgonging to $\Xi \subset \mathbb{R}^d$
  e.g., $\xi_i = (x_i, y_i)$ where $x_i$ input, $y_i$ label/target

- Objective $f_\theta : \Xi \to \mathbb{R}$, parameterized by $\theta$
  e.g., logistic regression $f_\theta(\xi) = f_\theta((x, y)) = \log\left(1 + e^{-y\langle\theta, x\rangle}\right)$

- Empirical Risk Minimization (ERM)

$$\min_\theta \frac{1}{n} \sum_{i=1}^{n} f_\theta(\xi_i) = \mathbb{E}_{\xi \sim \hat{P}_n} f_\theta(\xi) \quad \text{with } \hat{P}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{\xi_i}$$

$\rightarrow$ Take into account uncertainty in the training data

- Distributionally Robust Optimization (DRO):

$$\min_\theta \sup_{Q \in \mathcal{U}(\hat{P}_n)} \mathbb{E}_{\xi \sim Q}[f_\theta(\xi)] \quad \text{where } \mathcal{U}(\hat{P}_n) \textit{ ambiguity set}$$

# Distributionally Robust Optimization

$$\min_{\theta} \sup_{Q \in \mathcal{U}(\hat{P}_n)} \mathbb{E}_{\xi \sim Q}[f_\theta(\xi)]$$

Choice of ambiguity set $\mathcal{U}(\hat{P}_n)$

- $\mathcal{U}(\hat{P}_n)$ defined by moment constraints (Delage and Ye, 2010).
- Through distance/divergence

$$\mathcal{U}(\hat{P}_n) = \{Q : \mathrm{dist}(Q, \hat{P}_n) \leq \rho\}$$

  with e.g., KL, MMD...

- This talk: Wasserstein distance

$$\mathcal{U}(\hat{P}_n) = \{Q : W_p(Q, \hat{P}_n) \leq \rho\}$$

  Popular recently: nice theoretical/practical properties (Mohajerin Esfahani and Kuhn, 2018)

# Wasserstein distributionally robust optimization (WDRO)

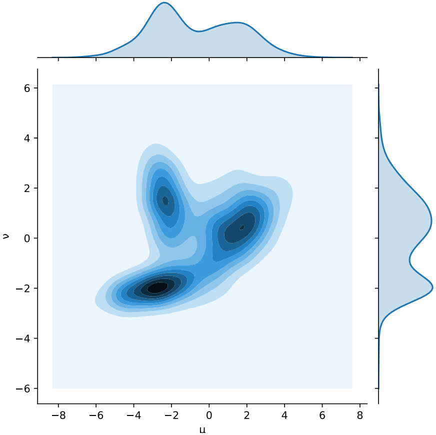
*p*-Wasserstein distance: for $P$, $Q$ probability distributions on $\Xi$,

$$W_p(P, Q) = \inf \left\{ \mathbb{E}_{(\xi,\zeta)\sim\pi} \|\xi - \zeta\|^p : \pi \in \mathcal{P}(\Xi^2), \pi_1 = P, \pi_2 = Q \right\}^{\frac{1}{p}}$$

Transport plan between two probabilities on $\mathbb{R}$:

*"Transport a pile of sand onto another one:*

*$\pi(\xi, \zeta) =$ mass of sand taken from $P$ at $\xi$ to put at $\zeta$ for $Q$"*



# Wasserstein distributionally robust optimization (WDRO)

$p$-Wasserstein distance: for $P$, $Q$ probability distributions on $\Xi$,

$$W_p(P, Q) = \inf \left\{ \mathbb{E}_{(\xi,\zeta)\sim\pi} \|\xi - \zeta\|^p : \pi \in \mathcal{P}(\Xi^2), \pi_1 = P, \pi_2 = Q \right\}^{\frac{1}{p}}$$

WDRO objective:

$$\sup_{Q : W_p(P,Q) \le \rho} \mathbb{E}_{\xi\sim Q}[f_\theta(\xi)]$$

Dual: fundamental *both* in theory and practice

$$\inf_{\lambda \ge 0} \lambda\rho^p + \mathbb{E}_{\xi\sim P}\left[\sup_{\zeta\in\Xi}\{f_\theta(\zeta) - \lambda\|\xi - \zeta\|^p\}\right]$$

# Wasserstein distributionally robust optimization (WDRO)

$p$-Wasserstein distance: for $P$, $Q$ probability distributions on $\Xi$,

$$W_p(P,Q) = \inf \left\{ \mathbb{E}_{(\xi,\zeta)\sim\pi} \|\xi - \zeta\|^p : \pi \in \mathcal{P}(\Xi^2), \pi_1 = P, \pi_2 = Q \right\}^{\frac{1}{p}}$$

WDRO objective:

$$\sup_{Q : W_p(P,Q) \leq \rho} \mathbb{E}_{\xi\sim Q}[f_\theta(\xi)]$$

Dual: fundamental *both* in theory and practice

$$\inf_{\lambda \geq 0} \lambda\rho^p + \mathbb{E}_{\xi\sim P}\left[\sup_{\zeta\in\Xi}\{f_\theta(\zeta) - \lambda\|\xi - \zeta\|^p\}\right]$$

$\rightarrow$ For *structured* $f_\theta$, dual simplifies (solvable as min-max, recall S. Wright's talk)

# Illustration: logistic regression and distributional shift

$\xi = (x, y)$ with $y \in -1, +1$

$$f_\theta((x, y)) = \log\left(1 + e^{-y\langle\theta, x\rangle}\right)$$

Training:

$X|Y = -1 \sim N(\mu_-, 5)$

$X|Y = +1 \sim N(\mu_+, 1)$

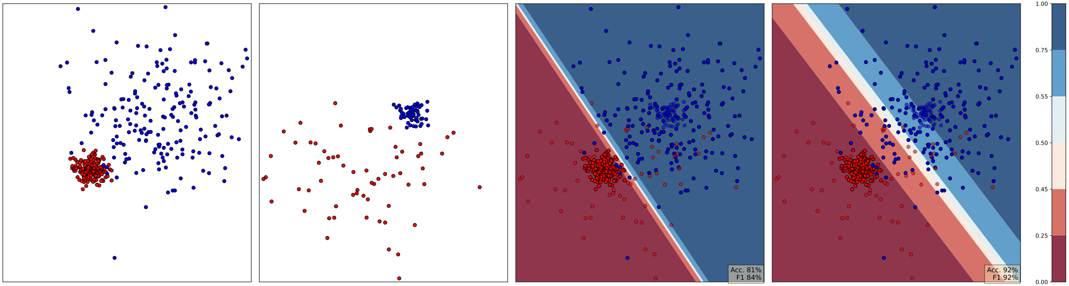
Testing:

$X|Y = -1 \sim N(\mu_-, 1)$

$X|Y = +1 \sim N(\mu_+, 5)$

Standard logistic regression
Test accuracy: 81%

WDRO Logistic regression
Test accuracy: 91%



Regularizing WDRO

# Regularization in optimal transport

$$\inf \left\{ \underbrace{\mathbb{E}_\pi c}_{\text{linear}} \qquad\qquad : \pi \in \mathcal{P}(\Xi^2), \pi_1 = P, \pi_2 = Q \right\}^{\frac{1}{p}},$$

# Regularization in optimal transport

$$\inf \left\{ \underbrace{\mathbb{E}_\pi c}_{\text{linear}} + \underbrace{R(\pi)}_{\text{strongly convex}} : \pi \in \mathcal{P}(\Xi^2), \pi_1 = P, \pi_2 = Q \right\}^{\frac{1}{p}},$$

## Most popular: entropic regularization

$$R(\pi) = \varepsilon KL(\pi | P \otimes Q) = \begin{cases} \varepsilon \int \log \frac{\mathrm{d}\pi}{\mathrm{d}P \otimes Q} \mathrm{d}P \otimes Q & \text{if } \pi \ll P \otimes Q \\ +\infty & \text{otherwise} \end{cases}$$

- Can be computed efficiently with the *Sinkhorn* algorithm

→ Popularized optimal transport in the ML community (Cuturi, 2013)

# Regularization in optimal transport

$$\inf \left\{ \underbrace{\mathbb{E}_\pi c}_{\text{linear}} + \underbrace{R(\pi)}_{\text{strongly convex}} : \pi \in \mathcal{P}(\Xi^2), \pi_1 = P, \pi_2 = Q \right\}^{\frac{1}{p}},$$

## Most popular: entropic regularization

$$R(\pi) = \varepsilon KL(\pi | P \otimes Q) = \begin{cases} \varepsilon \int \log \frac{\mathrm{d}\pi}{\mathrm{d}P\otimes Q} \mathrm{d}P \otimes Q & \text{if } \pi \ll P \otimes Q \\ +\infty & \text{otherwise} \end{cases}$$

- Can be computed efficiently with the *Sinkhorn* algorithm

$\rightarrow$ Popularized optimal transport in the ML community (Cuturi, 2013)

- Nice theoretical properties :
  - Provably approximates the unregularized Wasserstein distance (Genevay et al., 2019)
  - Resulting distance is smooth (Feydy et al., 2019)
  - Good statistical properties (Genevay et al., 2019)

# Regularizing the WDRO objective: but where?

WDRO objective: non-smooth as a function of $\theta$

$$\sup \left\{ \underbrace{\mathbb{E}_Q f_\theta}_{\text{linear function}} : Q \in \mathcal{P}(\Xi) \, , \, \underbrace{W_p(P, Q) \leq \rho}_{\text{non-smooth constraint}} \right\} = \inf_{\lambda \geq 0} \lambda \rho^p + \mathbb{E}_{\xi \sim P} \left[ \overbrace{\sup_{\zeta \in \Xi} \{ f_\theta(\zeta) - \lambda \| \xi - \zeta \|^p \}}^{\text{non-smooth}} \right] ,$$

# Regularizing the WDRO objective: but where?

WDRO objective: non-smooth as a function of $\theta$

$$\sup \left\{ \underbrace{\mathbb{E}_Q f_\theta}_{\text{linear function}} : Q \in \mathcal{P}(\Xi) \,,\, \underbrace{W_p(P, Q) \le \rho}_{\text{non-smooth constraint}} \right\} = \inf_{\lambda \ge 0} \lambda \rho^p + \mathbb{E}_{\xi \sim P} \left[ \overbrace{\sup_{\zeta \in \Xi} \{ f_\theta(\zeta) - \lambda \|\xi - \zeta\|^p \}}^{\text{non-smooth}} \right] ,$$

Reformulation: using the definition of $W_p(P, Q)$

$$\sup \left\{ \underbrace{\mathbb{E}_{\pi_2} f_\theta}_{\text{linear function}} : \pi \in \mathcal{P}(\Xi^2), \pi_1 = P \,,\, \underbrace{\mathbb{E}_{(\xi,\zeta) \sim \pi} \|\xi - \zeta\|^p \le \rho}_{\text{linear constraint}} \right\}$$

# Regularizing the WDRO objective

## Primal:

$$\sup \left\{ \underbrace{\mathbb{E}_{\pi_2} f_\theta}_{\text{linear function}} : \pi \in \mathcal{P}(\Xi^2), \pi_1 = P \,,\, \underbrace{\mathbb{E}_{(\xi,\zeta)\sim\pi}[\|\xi - \zeta\|^p]}_{\text{linear function}} \leq \rho \right\}$$

# Regularizing the WDRO objective

**Primal:** where $R, S : \mathcal{M}(\Xi^2) \to \mathbb{R} \cup \{+\infty\}$

$$\sup \left\{ \underbrace{\mathbb{E}_{\pi_2} f_\theta}_{\text{linear function}} - \underbrace{R(\pi)}_{\text{(strongly) convex}} : \pi \in \mathcal{P}(\Xi^2), \pi_1 = P\,, \underbrace{\mathbb{E}_{(\xi,\zeta)\sim\pi}[\|\xi - \zeta\|^p]}_{\text{linear function}} + \underbrace{S(\pi)}_{\text{(strongly) convex}} \leq \rho \right\}$$

# Regularizing the WDRO objective

Primal: where $R, S : \mathcal{M}(\Xi^2) \to \mathbb{R} \cup \{+\infty\}$

$$\sup \left\{ \underbrace{\mathbb{E}_{\pi_2} f_\theta}_{\text{linear function}} - \underbrace{R(\pi)}_{\text{(strongly) convex}} : \pi \in \mathcal{P}(\Xi^2), \pi_1 = P, \underbrace{\mathbb{E}_{(\xi,\zeta)\sim\pi}[\|\xi - \zeta\|^p]}_{\text{linear function}} + \underbrace{S(\pi)}_{\text{(strongly) convex}} \leq \rho \right\}$$

Dual:

$$\inf_{\lambda \geq 0} \inf_{\phi \in \mathcal{C}(\Xi^2)} \lambda\rho + \mathbb{E}_{\xi\sim P}\left[\sup_{\zeta\in\Xi} f(\zeta) - \lambda\|\xi - \zeta\|^p - \phi(\xi, \zeta)\right] + (R + \lambda S)^*(\phi)\,,$$

Idea of proof: on $\Xi$ compact to use duality $\mathcal{C}(\Xi^2)^* = \mathcal{M}(\Xi^2)$

- Lagrangian duality (Peypouquet, 2015)
- Fenchel duality (Bot et al., 2009)
- Exchange sup / $\mathbb{E}[\cdot]$ (Rockafellar and Wets, 1998)

# Entropic regularization

**Corollary (A., Iutzeler, Malick, 2022)**

*With* $S = 0$, $R = \varepsilon KL(\cdot|\pi_0)$ *s.t.* $(\pi_0)_1 = P$

$$\sup_{\pi \in \mathcal{P}_P(\Xi^2): \mathbb{E}_{(\xi,\zeta)\sim\pi}[\|\xi-\zeta\|^p] \leq \rho} \mathbb{E}_{\pi_2} f - \varepsilon KL(\pi|\pi_0) = \inf_{\lambda \geq 0} \lambda\rho^p + \varepsilon \mathbb{E}_{\xi \sim P} \log \left( \mathbb{E}_{\zeta \sim \pi_0(\cdot|\xi)} e^{\frac{f(\zeta) - \lambda\|\xi-\zeta\|^p}{\varepsilon}} \right)$$

To compare with:

$$\sup_{Q \in \mathcal{P}(\Xi): W_p(P,Q) \leq \rho} \mathbb{E}_Q f = \inf_{\lambda \geq 0} \lambda\rho^p + \mathbb{E}_{\xi \sim P} \left[ \sup_{\zeta \in \Xi} \{ f(\zeta) - \lambda\|\xi-\zeta\|^p \} \right]$$

Similar expressions (from different perspectives) in Blanchet and Kang (2020) and Wang et al. (2021)

# Entropic regularization

## Corollary (A., Iutzeler, Malick, 2022)

*With* $S = 0$, $R = \varepsilon KL(\cdot|\pi_0)$ *s.t.* $(\pi_0)_1 = P$

$$\sup_{\pi \in \mathcal{P}_P(\Xi^2): \mathbb{E}_{(\xi,\zeta)\sim\pi}[\|\xi-\zeta\|^p] \leq \rho} \mathbb{E}_{\pi_2} f - \varepsilon KL(\pi|\pi_0) = \inf_{\lambda \geq 0} \lambda\rho^p + \varepsilon \mathbb{E}_{\xi\sim P} \log\left(\mathbb{E}_{\zeta\sim\pi_0(\cdot|\xi)} e^{\frac{f(\zeta)-\lambda\|\xi-\zeta\|^p}{\varepsilon}}\right)$$

To compare with:

$$\sup_{Q \in \mathcal{P}(\Xi): W_p(P,Q) \leq \rho} \mathbb{E}_Q f \;=\; \inf_{\lambda \geq 0} \lambda\rho^p + \mathbb{E}_{\xi\sim P}\left[\sup_{\zeta\in\Xi}\{f(\zeta) - \lambda\|\xi-\zeta\|^p\}\right]$$

Similar expressions (from different perspectives) in Blanchet and Kang (2020) and Wang et al. (2021)

# Choice of regularization measure

OT: when $P$, $Q$ fixed, entropic regularization w.r.t. $\pi_0 = P \otimes Q$ since

$$\pi_1 = P \text{ and } \pi_2 = Q \implies \pi \ll P \otimes Q$$

## Choice of regularization measure

OT: when $P$, $Q$ fixed, entropic regularization w.r.t. $\pi_0 = P \otimes Q$ since

$$\pi_1 = P \text{ and } \pi_2 = Q \implies \pi \ll P \otimes Q$$

WDRO: $\pi_2$ not fixed! Choose, with $(\pi_0)_1 = P$,

$$\pi_0(\mathrm{d}\xi, \mathrm{d}\zeta) \propto P(\mathrm{d}\xi)\, \mathbb{1}_{\zeta \in \Xi}\, e^{-\frac{\|\xi-\zeta\|^p}{\sigma}} \mathrm{d}\zeta$$

$$\pi_0(\mathrm{d}\zeta | \xi) \propto \mathbb{1}_{\zeta \in \Xi}\, e^{-\frac{\|\xi-\zeta\|^p}{\sigma}} \mathrm{d}\zeta$$

# Choice of regularization measure

OT: when $P$, $Q$ fixed, entropic regularization w.r.t. $\pi_0 = P \otimes Q$ since

$$\pi_1 = P \text{ and } \pi_2 = Q \implies \pi \ll P \otimes Q$$

WDRO: $\pi_2$ not fixed! Choose, with $(\pi_0)_1 = P$,

$$\pi_0(\mathrm{d}\xi, \mathrm{d}\zeta) \propto P(\mathrm{d}\xi)\, \mathbb{1}_{\zeta \in \Xi}\, e^{-\frac{\|\xi-\zeta\|^p}{\sigma}} \mathrm{d}\zeta$$

$$\pi_0(\mathrm{d}\zeta | \xi) \propto \mathbb{1}_{\zeta \in \Xi}\, e^{-\frac{\|\xi-\zeta\|^p}{\sigma}} \mathrm{d}\zeta$$

$\Rightarrow$ Enforces $\pi \ll$ Lebesgue

# Approximation bound

Inspired by Genevay et al. (2019) for OT, bound the approximation error between:

$$\sup_{\pi \in \mathcal{P}(\Xi^2): \pi_1 = P, \mathbb{E}_{(\xi,\zeta)\sim\pi}[\|\xi-\zeta\|^p] \leq \rho} \{\mathbb{E}_{\pi_2} f\} \tag{WDRO}$$

$$\sup_{\pi \in \mathcal{P}(\Xi^2): \pi_1 = P, \mathbb{E}_{(\xi,\zeta)\sim\pi}[\|\xi-\zeta\|^p] \leq \rho} \{\mathbb{E}_{\pi_2} f - \varepsilon KL(\pi|\pi_0)\} \tag{$\varepsilon$-WDRO}$$

## Proposition (A., Iutzeler, Malick, 2022)

*Under regularity assumptions on $f$ and $\Xi \subset \mathbb{R}^d$ compact, with,* $\pi_0(\mathrm{d}\xi, \mathrm{d}\zeta) \propto P(\mathrm{d}\xi)\, \mathbb{1}_{\zeta \in \Xi}\, e^{-\frac{\|\xi-\zeta\|^p}{\sigma}} \mathrm{d}\zeta$ *then,*

$$0 \leq val(WDRO) - val(\varepsilon\text{-}WDRO) \leq \mathcal{O}\left(\varepsilon d \log \frac{1}{\varepsilon}\right)$$

# Approximation bound

Inspired by Genevay et al. (2019) for OT, bound the approximation error between:

$$\sup_{\pi \in \mathcal{P}(\Xi^2): \pi_1 = P, \mathbb{E}_{(\xi,\zeta)\sim\pi}[\|\xi-\zeta\|^p] \leq \rho} \{\mathbb{E}_{\pi_2} f\} \tag{WDRO}$$

$$\sup_{\pi \in \mathcal{P}(\Xi^2): \pi_1 = P, \mathbb{E}_{(\xi,\zeta)\sim\pi}[\|\xi-\zeta\|^p] \leq \rho} \{\mathbb{E}_{\pi_2} f - \varepsilon KL(\pi|\pi_0)\} \tag{$\varepsilon$-WDRO}$$

## Proposition (A., Iutzeler, Malick, 2022)

*Under regularity assumptions on $f$ and $\Xi \subset \mathbb{R}^d$ compact, with,* $\pi_0(\mathrm{d}\xi, \mathrm{d}\zeta) \propto P(\mathrm{d}\xi)\, \mathbb{1}_{\zeta\in\Xi}\, e^{-\frac{\|\xi-\zeta\|^p}{\sigma}} \mathrm{d}\zeta$ *then,*

$$0 \leq val(WDRO) - val(\varepsilon\text{-}WDRO) \leq \mathcal{O}\left(\varepsilon d \log \frac{1}{\varepsilon}\right)$$

Conclusion of the first part: regularize the WDRO objective

- Smooth and still tractable dual
- Provably close to original
- Interesting in practice (to be done)
- Interesting in theory (now in the second part!)

## “Robust” generalization properties of WDRO

# Statistical properties of WDRO

With $\hat{P}_n = \frac{1}{n}\sum_{i=1}^n \delta_{\xi_i}$ where $\xi_i \sim P_{train}$ i.i.d. in $\Xi \subset \mathbb{R}^d$

- Initial statistical guarantee for WDRO (Mohajerin Esfahani and Kuhn, 2018)

$$\text{if } \rho \geq \mathcal{O}\left(n^{-\frac{1}{d}}\right), \quad \text{with high probability,}$$

$$\underbrace{\sup_{Q : W_p(\hat{P}_n, Q) \leq \rho} \mathbb{E}_{\xi \sim Q}[f(\xi)]}_{\text{can compute and optimize!}} \geq \underbrace{\mathbb{E}_{\xi \sim P_{train}} f(\xi)}_{\text{cannot access}}$$

# Statistical properties of WDRO

With $\hat{P}_n = \frac{1}{n}\sum_{i=1}^n \delta_{\xi_i}$ where $\xi_i \sim P_{train}$ i.i.d. in $\Xi \subset \mathbb{R}^d$

- Initial statistical guarantee for WDRO (Mohajerin Esfahani and Kuhn, 2018)

  if $\rho \geq \mathcal{O}\left(n^{-\frac{1}{d}}\right)$, with high probability,

$$\underbrace{\sup_{Q: W_p(\hat{P}_n, Q) \leq \rho} \mathbb{E}_{\xi \sim Q}[f(\xi)]}_{\text{can compute and optimize!}} \geq \underbrace{\mathbb{E}_{\xi \sim P_{train}} f(\xi)}_{\text{cannot access}}$$

- Consequence of standard OT theory (Fournier and Guillin, 2015): with high probability

$$W_p(\hat{P}_n, P_{train}) \leq \mathcal{O}\left(n^{-\frac{1}{d}}\right)$$

# Statistical properties of WDRO

With $\hat{P}_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{\xi_i}$ where $\xi_i \sim P_{train}$ i.i.d. in $\Xi \subset \mathbb{R}^d$

- Initial statistical guarantee for WDRO (Mohajerin Esfahani and Kuhn, 2018)

  if $\rho \geq \mathcal{O}\left(n^{-\frac{1}{d}}\right)$, with high probability,

$$\underbrace{\sup_{Q:W_p(\hat{P}_n,Q)\leq\rho} \mathbb{E}_{\xi\sim Q}[f(\xi)]}_{\text{can compute and optimize!}} \geq \underbrace{\mathbb{E}_{\xi\sim P_{train}} f(\xi)}_{\text{cannot access}}$$

- Consequence of standard OT theory (Fournier and Guillin, 2015): with high probability

$$W_p(\hat{P}_n, P_{train}) \leq \mathcal{O}\left(n^{-\frac{1}{d}}\right)$$

$\rightarrow$ But exponential dependance in $d$...

- To do better: treat the WDRO objective as a *whole*

  e.g., (An and Gao, 2021) : guarantees with $\rho \propto n^{-\frac{1}{2}}$

# Statistical properties of WDRO

With $\hat{P}_n = \frac{1}{n}\sum_{i=1}^n \delta_{\xi_i}$ where $\xi_i \sim P_{train}$ i.i.d. in $\Xi \subset \mathbb{R}^d$

- Initial statistical guarantee for WDRO (Mohajerin Esfahani and Kuhn, 2018)

  if $\rho \geq \mathcal{O}\left(n^{-\frac{1}{d}}\right)$, with high probability,

$$\underbrace{\sup_{Q: W_p(\hat{P}_n, Q) \leq \rho} \mathbb{E}_{\xi \sim Q}[f(\xi)]}_{\text{can compute and optimize!}} \geq \underbrace{\mathbb{E}_{\xi \sim P_{train}} f(\xi)}_{\text{cannot access}}$$

- Consequence of standard OT theory (Fournier and Guillin, 2015): with high probability

$$W_p(\hat{P}_n, P_{train}) \leq \mathcal{O}\left(n^{-\frac{1}{d}}\right)$$

$\rightarrow$ But exponential dependance in $d$...

- To do better: treat the WDRO objective as a *whole*

  e.g., (An and Gao, 2021) : guarantees with $\rho \propto n^{-\frac{1}{2}}$
- But we can do even better, especially with regularization!

# What we would like

Define,

$$F_\rho^\varepsilon(f, P) = \sup_{\pi \in \mathcal{P}(\Xi^2): \pi_1 = P, \mathbb{E}_{(\xi,\zeta)\sim\pi}[\|\xi-\zeta\|^p] \leq \rho} \{\mathbb{E}_{\pi_2} f - \varepsilon KL(\pi|\pi_0)\}$$

and recall $\hat{P}_n = \frac{1}{n}\sum_{i=1}^n \delta_{\xi_i}$ where $\xi_i \sim P_{train}$

**Ideal result**

*With high probability, for all $f \in \mathcal{F}$,*

$$F_\rho^\varepsilon(f, \hat{P}_n) \geq F_{\rho-\rho_n}^\varepsilon(f, P_{train})$$

*with* $\rho_n = \mathcal{O}\left(n^{-\frac{1}{2}}\right)$, $\varepsilon \geq 0$

- Optimal requirement on radius when $n \to \infty$ (Blanchet, Murthy, et al., 2021)
- Guarantee on the WDRO objective and $\rho$ can be non-vanishing

# Nice consequences of ideal result, e.g. case $\varepsilon = 0$

$\hat{P}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{\xi_i}$ with $\xi_i \sim P_{train}$

1. Generalization bound:

$$\text{with high probability,} \quad F_\rho(f, \hat{P}_n) \;\geq\; F_{\rho-\rho_n}(f, P_{train}) \;\geq\; \mathbb{E}_{P_{train}} f$$

# Nice consequences of ideal result, e.g. case $\varepsilon = 0$

$\hat{P}_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{\xi_i}$ with $\xi_i \sim P_{train}$

1. Generalization bound:

$$\text{with high probability,} \quad F_\rho(f, \hat{P}_n) \;\geq\; F_{\rho-\rho_n}(f, P_{train}) \;\geq\; \mathbb{E}_{P_{train}} f$$

2. Distribution shift: $P_{train} \neq P_{test}$ i.e. $W_2(P_{train}, P_{test}) > 0$

$$\begin{aligned} \text{with high probability,} \quad F_\rho(f, \hat{P}_n) \;&\geq\; F_{\rho-\rho_n}(f, P_{train}) \\ &\geq\; \mathbb{E}_{P_{test}} f \\ \text{when } \rho - \rho_n \geq W_2(P_{train}, P_{test}) & \end{aligned}$$

# Can we have this ideal result?

Yes!

Existing works:

- In very restricted settings (Shafieezadeh-Abadeh et al., 2019)
- With error terms and obligatory vanishing $\rho$ (An and Gao, 2021)

# Can we have this ideal result?

Yes!

Existing works:

- In very restricted settings (Shafieezadeh-Abadeh et al., 2019)
- With error terms and obligatory vanishing $\rho$ (An and Gao, 2021)

Our work: version of the ideal result (A., Iutzeler, Malick, 2022)

- $\Xi$ compact and $p = 2$
- $\varepsilon > 0$ (at least today)
- + assumptions about $\mathcal{F}$, etc...

Idea of proof:

1. Why we need to lower bound $\lambda$
2. How we lower bound $\lambda$

# Idea of proof 1: Why we need to lower bound $\lambda$

Recall, for $\varepsilon > 0$,

$$F_\rho^\varepsilon(f, P) = \sup_{\pi \in \mathcal{P}(\Xi^2) : \pi_1 = P, \mathbb{E}_{(\xi,\zeta)\sim\pi}\left[\|\xi-\zeta\|^2\right] \leq \rho} \{\mathbb{E}_{\pi_2} f - \varepsilon KL(\pi|\pi_0)\}$$

$$= \inf_{\lambda \geq 0} \lambda\rho^2 + \mathbb{E}_{\xi\sim\hat{P}_n}\left[\log\left(\mathbb{E}_{\zeta\sim\pi_0(\cdot|\xi)}\left[e^{\frac{f(\zeta)-\lambda\|\xi-\zeta\|^2}{\varepsilon}}\right]\right)\right]$$

**Lemma**

*For $\rho > 0$, $\varepsilon > 0$ assume that there is some $\underline{\lambda}(\rho) > 0$ such that, with high probability,*

$$\forall f \in \mathcal{F}, \quad F_\rho^\varepsilon(f, \hat{P}_n) = \inf_{\lambda \geq \underline{\lambda}(\rho)} \lambda\rho^2 + \mathbb{E}_{\xi\sim\hat{P}_n}\left[\log\left(\mathbb{E}_{\zeta\sim\pi_0(\cdot|\xi)}\left[e^{\frac{f(\zeta)-\lambda\|\xi-\zeta\|^2}{\varepsilon}}\right]\right)\right]$$

*then we get the ideal result: with high probability, for all $f \in \mathcal{F}$,*

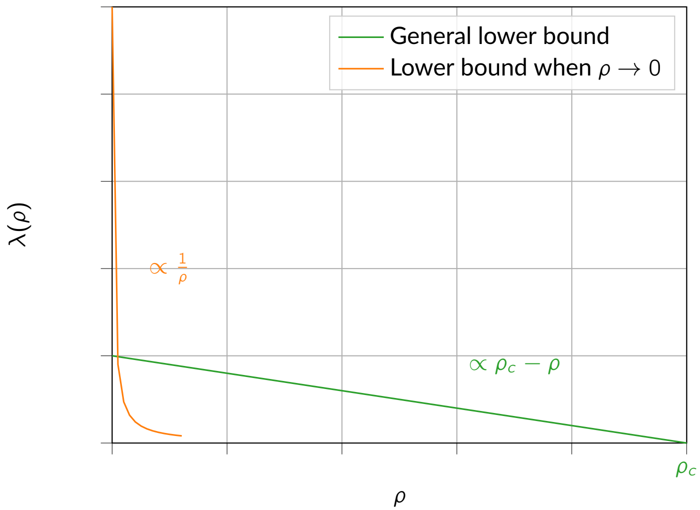
$$F_\rho^\varepsilon(f, \hat{P}_n) \geq F_{\rho-\rho_n}^\varepsilon(f, P_{train})$$

*with*

$$\rho_n = \mathcal{O}\left(\frac{1}{\underline{\lambda}(\rho)\rho\sqrt{n}}\right)$$

$\Rightarrow$ Need a lower bound $\underline{\lambda}(\rho)$ on the optimal dual multiplier for $\hat{P}_n$

# Idea of proof 2: How we lower bound $\lambda$




Recall: $\lambda$ dual multiplier for

$$W_2(\hat{P}_n, Q) \le \rho$$

When $\rho$ large enough, the constraint becomes inactive and $\lambda = 0$

# Ideal theorem

**Theorem (informal) (A., Iutzeler, Malick, 2022)**

*For $\varepsilon \propto \rho$, with*

$$\rho_n = \mathcal{O}\left(\frac{1}{\sqrt{n}}\right),$$

*if*

$$\rho_n \leq \rho \leq \frac{\rho_c}{2} - \mathcal{O}\left(n^{-\frac{1}{2}}\right), \quad \rho_c \geq \mathcal{O}\left(n^{-\frac{1}{6}}\right)$$

*then, with high probability,*

$$\forall f \in \mathcal{F}, \quad F_\rho^\varepsilon(f, \hat{P}_n) \geq F_{\rho-\rho_n}^\varepsilon(f, P_{train})$$

# Ideal theorem

**Theorem (informal) (A., Iutzeler, Malick, 2022)**

*For $\varepsilon \propto \rho$, with*

$$\rho_n = \mathcal{O}\left(\frac{1}{\sqrt{n}}\right),$$

*if*

$$\rho_n \leq \rho \leq \frac{\rho_c}{2} - \mathcal{O}\left(n^{-\frac{1}{2}}\right), \quad \rho_c \geq \mathcal{O}\left(n^{-\frac{1}{6}}\right)$$

*then, with high probability,*

$$\forall f \in \mathcal{F}, \quad F_\rho^\varepsilon(f, \hat{P}_n) \geq F_{\rho-\rho_n}^\varepsilon(f, P_{train})$$

Remark: extends to unregularized ($\varepsilon = 0$) with stronger assumptions on $\mathcal{F}$

# Conclusion

Main takeaways:

- Present regularization for WDRO: smooth dual and still provably close to the original
- New generalization bounds for WDRO, especially for regularized WDRO

# Conclusion

Main takeaways:

- Present regularization for WDRO: smooth dual and still provably close to the original
- New generalization bounds for WDRO, especially for regularized WDRO

Future work:

- Wrap up the paper ☺
- Generalize the current generalization bounds (non-compact, $p \neq 2$, other regularizations...)
- Efficient and scalable computational methods

Azizian, Iutzeler, Malick (2022). “Regularization for Wasserstein Distributionally Robust Optimization”. *arXiv:2205.08826, submitted*.

Azizian, Iutzeler, Malick (2022). “Robust Generalization Bounds for Wasserstein Distributionally Robust Optimization”. *to be submitted*.

# Bibliography I


An, Yang and Rui Gao (2021). "Generalization Bounds for (Wasserstein) Robust Optimization". In: *Advances in Neural Information Processing Systems* 34.

Blanchet, Jose and Yang Kang (2020). "Semi-Supervised Learning Based on Distributionally Robust Optimization". In: *Data Analysis and Applications 3*. John Wiley & Sons, Ltd, pp. 1–33. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119721871.ch1.

Blanchet, Jose, Karthyek Murthy, and Nian Si (Mar. 3, 2021). "Confidence Regions in Wasserstein Distributionally Robust Estimation". URL: http://arxiv.org/abs/1906.01614.

Bot, Radu Ioan, Sorin-Mihai Grad, and Gert Wanka (2009). *Duality in Vector Optimization*. Vector Optimization. Berlin, Heidelberg: Springer Berlin Heidelberg. URL: http://link.springer.com/10.1007/978-3-642-02886-1.

Carlier, Guillaume et al. (Jan. 1, 2017). "Convergence of Entropic Schemes for Optimal Transport and Gradient Flows". In: *SIAM J. Math. Anal.* 49, pp. 1385–1418. URL: https://epubs.siam.org/doi/10.1137/15M1050264.

Cuturi, Marco (2013). "Sinkhorn Distances: Lightspeed Computation of Optimal Transport". In: *Advances in Neural Information Processing Systems*. Vol. 26. Curran Associates, Inc. URL: https://papers.nips.cc/paper/2013/hash/af21d0c97db2e27e13572cbf59eb343d-Abstract.html.

# Bibliography II


Delage, Erick and Yinyu Ye (June 1, 2010). "Distributionally Robust Optimization Under Moment Uncertainty with Application to Data-Driven Problems". In: *Operations Research* 58, pp. 595–612. URL: https://pubsonline.informs.org/doi/10.1287/opre.1090.0741.

Feydy, Jean et al. (Apr. 11, 2019). "Interpolating between Optimal Transport and MMD Using Sinkhorn Divergences". In: *The 22nd International Conference on Artificial Intelligence and Statistics*. The 22nd International Conference on Artificial Intelligence and Statistics. PMLR, pp. 2681–2690. URL: https://proceedings.mlr.press/v89/feydy19a.html.

Fournier, Nicolas and Arnaud Guillin (Aug. 1, 2015). "On the Rate of Convergence in Wasserstein Distance of the Empirical Measure". In: *Probab. Theory Relat. Fields* 162, pp. 707–738. URL: https://doi.org/10.1007/s00440-014-0583-7.

Genevay, Aude et al. (Apr. 11, 2019). "Sample Complexity of Sinkhorn Divergences". In: *The 22nd International Conference on Artificial Intelligence and Statistics*. The 22nd International Conference on Artificial Intelligence and Statistics. PMLR, pp. 1574–1583. URL: https://proceedings.mlr.press/v89/genevay19a.html.

Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy (Mar. 20, 2015). "Explaining and Harnessing Adversarial Examples". URL: http://arxiv.org/abs/1412.6572.

# Bibliography III


Kwon, Yongchan et al. (Nov. 21, 2020). "Principled Learning Method for Wasserstein Distributionally Robust Optimization with Local Perturbations". In: *International Conference on Machine Learning*. International Conference on Machine Learning. PMLR, pp. 5567–5576. URL: https://proceedings.mlr.press/v119/kwon20a.html.

Li, Jiajin, Caihua Chen, and Anthony Man-Cho So (2020). "Fast Epigraphical Projection-based Incremental Algorithms for Wasserstein Distributionally Robust Support Vector Machine". In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., pp. 4029–4039. URL: https://proceedings.neurips.cc/paper/2020/hash/2974788b53f73e7950e8aa49f3a306db-Abstract.html.

Mohajerin Esfahani, Peyman and Daniel Kuhn (Sept. 1, 2018). "Data-Driven Distributionally Robust Optimization Using the Wasserstein Metric: Performance Guarantees and Tractable Reformulations". In: *Math. Program.* 171, pp. 115–166. URL: https://doi.org/10.1007/s10107-017-1172-1.

Peypouquet, Juan (2015). *Convex Optimization in Normed Spaces: Theory, Methods and Examples*. SpringerBriefs in Optimization. Springer International Publishing. URL: https://www.springer.com/gp/book/9783319137094.

Rockafellar, R. Tyrrell and Roger J.-B. Wets (1998). *Variational Analysis*. Grundlehren Der Mathematischen Wissenschaften. Berlin Heidelberg: Springer-Verlag. URL: https://www.springer.com/gp/book/9783540627722.

# Bibliography IV


Shafieezadeh Abadeh, Soroosh, Peyman Mohajerin Mohajerin Esfahani, and Daniel Kuhn (2015). "Distributionally Robust Logistic Regression". In: *Advances in Neural Information Processing Systems*. Vol. 28. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2015/hash/cc1aa436277138f61cda703991069eaf-Abstract.html.

Shafieezadeh-Abadeh, Soroosh, Daniel Kuhn, and Peyman Mohajerin Esfahani (2019). "Regularization via Mass Transportation". In: *Journal of Machine Learning Research* 20, pp. 1–68. URL: http://jmlr.org/papers/v20/17-633.html.

Wang, Jie, Rui Gao, and Yao Xie (Sept. 24, 2021). "Sinkhorn Distributionally Robust Optimization". URL: http://arxiv.org/abs/2109.11926.