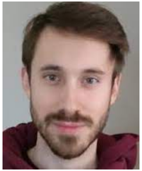
Distributionally robust optimization: Wasserstein ambiguity, regularization, and generalization

Jérôme MALICK

CNRS, Lab. Jean Kuntzmann & MIAI (Institut IA de Grenoble)

Journées SMAI-MODE – Limoges – June 2022

Based on joint work with
Waïss Azizian, Franck Iutzeler



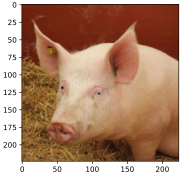
More robustness in ML/IA ?

we do not want machine-learned systems to fail when used in real-word

More robustness in ML/IA ?

we do not want machine-learned systems to fail when used in real-word

Example 1: Flying pigs (notebooks of NeurIPS 2018, tutorial on robustness)

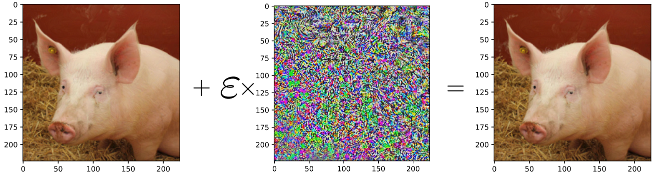


pig (99%)

More robustness in ML/IA ?

> we do not want machine-learned systems to fail when used in real-word

Example 1: Flying pigs (notebooks of NeurIPS 2018, tutorial on robustness)

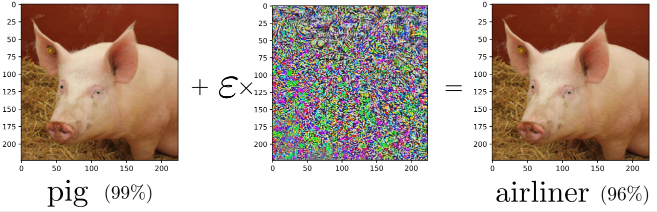


pig (99%)

More robustness in ML/IA ?

we do not want machine-learned systems to fail when used in real-word

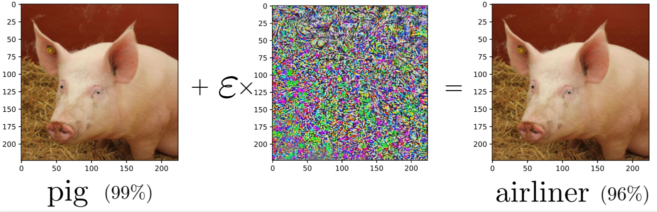
Example 1: Flying pigs (notebooks of NeurIPS 2018, tutorial on robustness)



More robustness in ML/IA ?

we do not want machine-learned systems to fail when used in real-word

Example 1: Flying pigs (notebooks of NeurIPS 2018, tutorial on robustness)

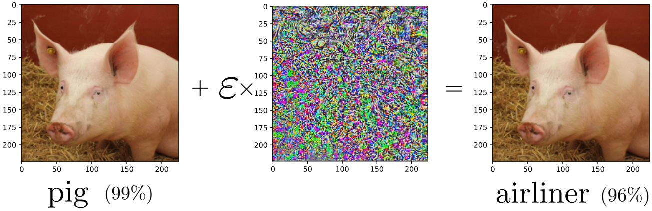


"ML is a wonderful technology: it makes pigs fly" [Kolter, Madry '18]

More robustness in ML/IA ?

we do not want machine-learned systems to fail when used in real-word

Example 1: Flying pigs (notebooks of NeurIPS 2018, tutorial on robustness)



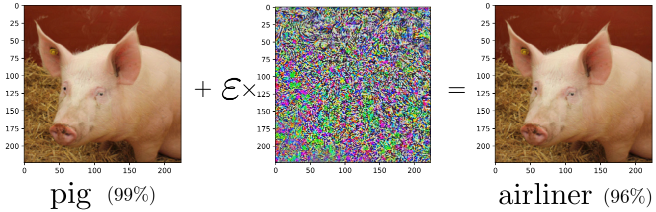
"ML is a wonderful technology: it makes pigs fly" [Kolter, Madry '18]

Example 2: Attacks against self-driving cars [@ CVPR '18]

More robustness in ML/IA ?

we do not want machine-learned systems to fail when used in real-word

Example 1: Flying pigs (notebooks of NeurIPS 2018, tutorial on robustness)



"ML is a wonderful technology: it makes pigs fly" [Kolter, Madry '18]

Example 2: Attacks against self-driving cars [@ ICLR '19]

Set-up: data-driven optimization under uncertainty

- Training data: $\xi_1, \ldots, \xi_N \sim \mathbb{P}_{\text{train}}$ (unknown)
  e.g. in supervised learning: $\xi_i = (a_i, y_i)$ feature, label
- Train model: x the parameter $f(x, \cdot)$ the objective function
  e.g. least-square regression: $f\left(x, (a, y)\right) = (x^\top a - y)^2$
- Compute x via empirical risk minimization (a.k.a SAA)
  (minimize the average loss on training data)

$$\min_x \frac{1}{N} \sum_{i=1}^{N} f(x, \xi_i)$$

- Prediction with x for different data ξ
  - Adversarial attacks (e.g. flying pigs)
  - Distributional shifts: $\mathbb{P}_{\text{train}} \neq \mathbb{P}_{\text{test}}$
  - Generalization: computations with $\widehat{\mathbb{P}}_N$ and guarantees on $\mathbb{P}_{\text{train}}$
  - Other situations...
- Solution: take possible variations into account during training (= when optimizing ☺)

Set-up: data-driven optimization under uncertainty

- Training data: $\xi_1, \ldots, \xi_N \sim \mathbb{P}_{\text{train}}$ (unknown)
  e.g. in supervised learning: $\xi_i = (a_i, y_i)$ feature, label
- Train model: x the parameter $f(x, \cdot)$ the objective function
  e.g. least-square regression: $f\left(x, (a, y)\right) = (x^\top a - y)^2$
- Compute x via empirical risk minimization (a.k.a SAA)
  (minimize the average loss on training data)

$$\min_x \frac{1}{N} \sum_{i=1}^{N} f(x, \xi_i) = \mathbb{E}_{\widehat{\mathbb{P}}_N}[f(x, \xi)] \qquad \text{with } \widehat{\mathbb{P}}_N = \frac{1}{N} \sum_{i=1}^{N} \delta_{\xi_i}$$

- Prediction with x for different data ξ
  - Adversarial attacks (e.g. flying pigs)
  - Distributional shifts: $\mathbb{P}_{\text{train}} \neq \mathbb{P}_{\text{test}}$
  - Generalization: computations with $\widehat{\mathbb{P}}_N$ and guarantees on $\mathbb{P}_{\text{train}}$
  - Other situations...
- Solution: take possible variations into account during training (= when optimizing ☺)

(Distributionally) robust optimization

Optimize expected loss for the worst probability in a set of perturbations

Instead of $\min_{x} \mathbb{E}_{\widehat{\mathbb{P}}_N}[f(x,\xi)]$ solve $\min_{x} \max_{\mathbb{Q}\in\mathcal{U}} \mathbb{E}_{\mathbb{Q}}[f(x,\xi)]$

with $\mathcal{U}$ a neighborhood of $\widehat{\mathbb{P}}_N$ (called ambiguity set)

(Distributionally) robust optimization

Optimize expected loss for the worst probability in a set of perturbations

Instead of $\min_{x} \mathbb{E}_{\widehat{\mathbb{P}}_N}[f(x,\xi)]$ solve $\min_{x} \max_{\mathbb{Q}\in\mathcal{U}} \mathbb{E}_{\mathbb{Q}}[f(x,\xi)]$

with $\mathcal{U}$ a neighborhood of $\widehat{\mathbb{P}}_N$ (called ambiguity set)

modeling vs. computational tractability

- $\mathcal{U} = \left\{\widehat{\mathbb{P}}_N\right\}$: $\min_{x} \frac{1}{N}\sum_{i=1}^{N} f(x,\xi_i)$ standard ERM
- $\mathcal{U}$ defined by moments e.g. [Delage, Ye, '10]
- $\mathcal{U} = \left\{\mathbb{Q} : d(\widehat{\mathbb{P}}_N,\mathbb{Q}) \leqslant \rho\right\}$ for various distances or divergences
  E.g. KL-div., χ_2-div., max-mean-discrepancy... e.g. [Namkoong, Duchi '17]
- $\mathcal{U} = \left\{\mathbb{Q} : W(\widehat{\mathbb{P}}_N,\mathbb{Q}) \leqslant \rho\right\}$ Wasserstein distance [Kuhn *et al.* '18] (in this talk)
- and Sinkhorn ? not considered yet ?! because not clear... (more on that later)

WDRO: DRO with Wasserstein balls as ambiguity sets

Notation: p-Wasserstein distance

$$W(\mathbb{P},\mathbb{Q}) = \min_{\boldsymbol{\pi}}\{\,\mathbb{E}_{\boldsymbol{\pi}}[\|\xi-\xi'\|^p] : \boldsymbol{\pi} \text{ with marginals } [\boldsymbol{\pi}]_1 = \mathbb{P} \text{ and } [\boldsymbol{\pi}]_2 = \mathbb{Q}\}^{\frac{1}{p}}$$

WDRO objective for given $\mathbb{P}$ and ρ

$$\begin{cases} \max_{\mathbb{Q}} & \mathbb{E}_{\mathbb{Q}}[f(\xi)] \\ & W(\mathbb{P},\mathbb{Q}) \leqslant \rho \end{cases}$$

WDRO: DRO with Wasserstein balls as ambiguity sets

Notation: p-Wasserstein distance

$$W(\mathbb{P},\mathbb{Q}) = \min_{\boldsymbol{\pi}}\{\, \mathbb{E}_{\boldsymbol{\pi}}[\|\xi-\xi'\|^p] : \boldsymbol{\pi} \text{ with marginals } [\boldsymbol{\pi}]_1 = \mathbb{P} \text{ and } [\boldsymbol{\pi}]_2 = \mathbb{Q}\}^{\frac{1}{p}}$$

WDRO objective for given $\mathbb{P}$ and ρ

$$\begin{cases} \max_{\mathbb{Q}} \quad \mathbb{E}_{\mathbb{Q}}[f(\xi)] \\ \quad W(\mathbb{P},\mathbb{Q}) \leqslant \rho \end{cases} \iff \begin{cases} \max_{\mathbb{Q},\boldsymbol{\pi}} \quad \mathbb{E}_{\mathbb{Q}}[f(\xi)] \\ \quad [\boldsymbol{\pi}]_1 = \mathbb{P}, [\boldsymbol{\pi}]_2 = \mathbb{Q} \\ \quad \min_{\boldsymbol{\pi}} \mathbb{E}_{\boldsymbol{\pi}}[\|\xi-\xi'\|^p] \leqslant \rho^p \end{cases}$$

WDRO: DRO with Wasserstein balls as ambiguity sets

Notation: p-Wasserstein distance

$$W(\mathbb{P},\mathbb{Q}) = \min_{\boldsymbol{\pi}}\{\,\mathbb{E}_{\boldsymbol{\pi}}[\|\xi-\xi'\|^p] : \boldsymbol{\pi} \text{ with marginals } [\boldsymbol{\pi}]_1 = \mathbb{P} \text{ and } [\boldsymbol{\pi}]_2 = \mathbb{Q}\}^{\frac{1}{p}}$$

WDRO objective for given $\mathbb{P}$ and ρ

$$\begin{cases} \max_{\mathbb{Q}} \quad \mathbb{E}_{\mathbb{Q}}[f(\xi)] \\ \quad W(\mathbb{P},\mathbb{Q}) \leqslant \rho \end{cases} \iff \begin{cases} \max_{\mathbb{Q},\boldsymbol{\pi}} \quad \mathbb{E}_{\mathbb{Q}}[f(\xi)] \\ \quad [\boldsymbol{\pi}]_1 = \mathbb{P}, [\boldsymbol{\pi}]_2 = \mathbb{Q} \\ \quad \min_{\boldsymbol{\pi}} \mathbb{E}_{\boldsymbol{\pi}}[\|\xi-\xi'\|^p] \leqslant \rho^p \end{cases} \iff \boxed{\begin{cases} \max_{\boldsymbol{\pi}} \quad \mathbb{E}_{[\boldsymbol{\pi}]_2}[f(\xi)] \\ \quad [\boldsymbol{\pi}]_1 = \mathbb{P} \\ \quad \mathbb{E}_{\boldsymbol{\pi}}[\|\xi-\xi'\|^p] \leqslant \rho^p \end{cases}}$$

Computable in many cases

e.g. [Kuhn *et al.* '18]

WDRO: DRO with Wasserstein balls as ambiguity sets

Notation: p-Wasserstein distance

$$W(\mathbb{P},\mathbb{Q}) = \min_{\boldsymbol{\pi}}\{\, \mathbb{E}_{\boldsymbol{\pi}}[\|\xi - \xi'\|^p] : \boldsymbol{\pi} \text{ with marginals } [\boldsymbol{\pi}]_1 = \mathbb{P} \text{ and } [\boldsymbol{\pi}]_2 = \mathbb{Q}\}^{\frac{1}{p}}$$

WDRO objective for given $\mathbb{P}$ and ρ

$$\begin{cases} \max_{\mathbb{Q}} & \mathbb{E}_{\mathbb{Q}}[f(\xi)] \\ & W(\mathbb{P},\mathbb{Q}) \leqslant \rho \end{cases} \iff \begin{cases} \max_{\mathbb{Q},\boldsymbol{\pi}} & \mathbb{E}_{\mathbb{Q}}[f(\xi)] \\ & [\boldsymbol{\pi}]_1 = \mathbb{P}, [\boldsymbol{\pi}]_2 = \mathbb{Q} \\ & \min_{\boldsymbol{\pi}} \mathbb{E}_{\boldsymbol{\pi}}[\|\xi - \xi'\|^p] \leqslant \rho^p \end{cases} \iff \boxed{\begin{cases} \max_{\boldsymbol{\pi}} & \mathbb{E}_{[\boldsymbol{\pi}]_2}[f(\xi)] \\ & [\boldsymbol{\pi}]_1 = \mathbb{P} \\ & \mathbb{E}_{\boldsymbol{\pi}}[\|\xi - \xi'\|^p] \leqslant \rho^p \end{cases}}$$

Computable in many cases

e.g. [Kuhn *et al.* '18]

Dual $\displaystyle \min_{\lambda \geqslant 0} \ \lambda\rho^p + \mathbb{E}_{\mathbb{P}}[\, \max_{\xi'}\{f(\xi') - \lambda\|\xi - \xi'\|^p\}\,]$

WDRO: DRO with Wasserstein balls as ambiguity sets

Notation: p-Wasserstein distance

$$W(\mathbb{P},\mathbb{Q}) = \min_{\boldsymbol{\pi}} \{ \mathbb{E}_{\boldsymbol{\pi}}[\|\xi - \xi'\|^p] : \boldsymbol{\pi} \text{ with marginals } [\boldsymbol{\pi}]_1 = \mathbb{P} \text{ and } [\boldsymbol{\pi}]_2 = \mathbb{Q} \}^{\frac{1}{p}}$$

WDRO objective for given $\mathbb{P}$ and ρ

$$\begin{cases} \max_{\mathbb{Q}} & \mathbb{E}_{\mathbb{Q}}[f(\xi)] \\ & W(\mathbb{P},\mathbb{Q}) \leqslant \rho \end{cases} \iff \begin{cases} \max_{\mathbb{Q},\boldsymbol{\pi}} & \mathbb{E}_{\mathbb{Q}}[f(\xi)] \\ & [\boldsymbol{\pi}]_1 = \mathbb{P}, [\boldsymbol{\pi}]_2 = \mathbb{Q} \\ & \min_{\boldsymbol{\pi}} \mathbb{E}_{\boldsymbol{\pi}}[\|\xi - \xi'\|^p] \leqslant \rho^p \end{cases} \iff \boxed{\begin{cases} \max_{\boldsymbol{\pi}} & \mathbb{E}_{[\boldsymbol{\pi}]_2}[f(\xi)] \\ & [\boldsymbol{\pi}]_1 = \mathbb{P} \\ & \mathbb{E}_{\boldsymbol{\pi}}[\|\xi - \xi'\|^p] \leqslant \rho^p \end{cases}}$$

Computable in many cases
e.g. [Kuhn *et al.* '18]

Dual $\quad \min_{\lambda \geqslant 0} \; \lambda\rho^p + \mathbb{E}_{\mathbb{P}}[\, \max_{\xi'} \{ f(\xi') - \lambda\|\xi - \xi'\|^p \} \,]$

Success: WDRO is popular

- Natural in many applications, e.g. in learning [Kuhn *et al.* '20]
- Good statistical/practical properties, e.g. [Blanchet *et al.* '18]
- Interprets up to first-order as a penalization by $\|\nabla_\xi f(x,\xi)\|$, e.g. [Gao *et al.* '18]

But WDRO also has some limitations: further work needed to extend the WDRO toolkit

Our work: regularization for WDRO

Inspired by [Paty, Cuturi '20] (study of general regularization for OT)

We propose to regularize WDRO with general convex functions $(R, S : \mathcal{M}(\Xi \times \Xi) \to \mathbb{R} \cup \{+\infty\})$

$$\begin{cases} \max_{\boldsymbol{\pi}} \quad \mathbb{E}_{[\boldsymbol{\pi}]_2}[f(\xi)] - R(\boldsymbol{\pi}) \\ \quad [\boldsymbol{\pi}]_1 = \mathbb{P} \\ \quad \mathbb{E}_{\boldsymbol{\pi}}[\|\xi - \xi'\|^p] + S(\boldsymbol{\pi}) \leqslant \rho^p \end{cases}$$

Our work: regularization for WDRO

Inspired by [Paty, Cuturi '20] (study of general regularization for OT)

We propose to regularize WDRO with general convex functions $(R, S \colon \mathcal{M}(\Xi \times \Xi) \to \mathbb{R} \cup \{+\infty\})$

$$\begin{cases} \max_{\boldsymbol{\pi}} \quad \mathbb{E}_{[\boldsymbol{\pi}]_2}[f(\xi)] - R(\boldsymbol{\pi}) \\ \quad [\boldsymbol{\pi}]_1 = \mathbb{P} \\ \quad \mathbb{E}_{\boldsymbol{\pi}}[\|\xi - \xi'\|^p] + S(\boldsymbol{\pi}) \leqslant \rho^p \end{cases}$$

Dual regularized WDRO

$$\min_{\lambda \geqslant 0} \min_{\varphi} \; \lambda \rho^p + \mathbb{E}_{\mathbb{P}}[\, \max_{\xi'} \{f(\xi') - \lambda \|\xi - \xi'\|^p - \varphi(\xi, \xi')\}\,] + (R + \lambda S)_*(\varphi)$$

Quite abstract...

Our work: regularization for WDRO

Inspired by [Paty, Cuturi '20] (study of general regularization for OT)

We propose to regularize WDRO with general convex functions $(R, S : \mathcal{M}(\Xi \times \Xi) \to \mathbb{R} \cup \{+\infty\})$

$$\begin{cases} \max_{\boldsymbol{\pi}} \quad \mathbb{E}_{[\boldsymbol{\pi}]_2}[f(\xi)] - R(\boldsymbol{\pi}) \\ \quad [\boldsymbol{\pi}]_1 = \mathbb{P} \\ \quad \mathbb{E}_{\boldsymbol{\pi}}[\|\xi - \xi'\|^p] + S(\boldsymbol{\pi}) \leqslant \rho^p \end{cases}$$

Dual regularized WDRO

$$\min_{\lambda \geqslant 0} \min_{\varphi} \lambda \rho^p + \mathbb{E}_{\mathbb{P}}[\max_{\xi'} \{ f(\xi') - \lambda \|\xi - \xi'\|^p - \varphi(\xi, \xi') \}] + (R + \lambda S)_*(\varphi)$$

Quite abstract... but more concrete expressions when specialized

e.g. $R(\pi) = \varepsilon \,\mathrm{KL}(\pi | \pi_0)$ and $S(\pi) = \delta \,\mathrm{KL}(\pi | \pi_0)$ KL div. : $\mathrm{KL}(\mu | \nu) = \begin{cases} \int \log \frac{d\mu}{d\nu} \, d\mu & \mu \ll \nu \\ +\infty & \text{otherwise} \end{cases}$

for a given π_0 such that $[\pi_0]_1 = \mathbb{P}$

$$\min_{\lambda \geqslant 0} \quad \lambda \rho^p + (\varepsilon + \lambda \delta) \, \mathbb{E}_{\mathbb{P}} \log \left(\mathbb{E}_{\xi' \sim \pi_0(\cdot|\xi)} e^{\frac{f(\xi') - \lambda \|\xi - \xi'\|^p}{\varepsilon + \lambda \delta}} \right)$$

smooth ☺

(Similar expressions in [Blanchet *et al* '21] [Wang *et al* '21])

Entropic regularization: OT & WDRO

OT: Sinkhorn distance, very popular from [Cuturi '13]

$$\min_{\boldsymbol{\pi}} \{ \mathbb{E}_{\boldsymbol{\pi}}[\|\xi - \xi'\|^p] + \varepsilon \,\mathrm{KL}(\boldsymbol{\pi}|\pi_0) : \boldsymbol{\pi} \text{ with marginals } [\boldsymbol{\pi}]_1 = \mathbb{P} \text{ and } [\boldsymbol{\pi}]_2 = \mathbb{Q}\}$$

WDRO: entropic regularization, seemingly new [Azizian, Iutzeler, M. '21]

$$\begin{cases} \max_{\boldsymbol{\pi}} \quad \mathbb{E}_{[\boldsymbol{\pi}]_2}[f(\xi)] - \varepsilon \,\mathrm{KL}(\boldsymbol{\pi}|\pi_0) \\ \quad [\boldsymbol{\pi}]_1 = \mathbb{P} \\ \quad \mathbb{E}_{\boldsymbol{\pi}}[\|\xi - \xi'\|^p] + \delta \,\mathrm{KL}(\boldsymbol{\pi}|\pi_0) \leqslant \rho^p \end{cases}$$

Subtility:

in OT, take $\pi_0 = \mathbb{P} \otimes \mathbb{Q}$

$[\boldsymbol{\pi}]_1 = \mathbb{P}, [\boldsymbol{\pi}]_2 = \mathbb{Q} \Rightarrow \boldsymbol{\pi} \ll \pi_0$

vs

but in WDRO, $[\pi_0]_2$ not fixed !

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

Entropic regularization: OT & WDRO

OT: Sinkhorn distance, very popular from [Cuturi '13]

$$\min_{\boldsymbol{\pi}} \{ \mathbb{E}_{\boldsymbol{\pi}}[\|\xi - \xi'\|^p] + \varepsilon \, \mathrm{KL}(\boldsymbol{\pi}|\pi_0) : \boldsymbol{\pi} \text{ with marginals } [\boldsymbol{\pi}]_1 = \mathbb{P} \text{ and } [\boldsymbol{\pi}]_2 = \mathbb{Q} \}$$

WDRO: entropic regularization, seemingly new [Azizian, Iutzeler, M. '21]

$$\begin{cases} \max_{\boldsymbol{\pi}} & \mathbb{E}_{[\boldsymbol{\pi}]_2}[f(\xi)] - \varepsilon \, \mathrm{KL}(\boldsymbol{\pi}|\pi_0) \\ & [\boldsymbol{\pi}]_1 = \mathbb{P} \\ & \mathbb{E}_{\boldsymbol{\pi}}[\|\xi - \xi'\|^p] + \delta \, \mathrm{KL}(\boldsymbol{\pi}|\pi_0) \leqslant \rho^p \end{cases}$$

Subtility:

in OT, take $\pi_0 = \mathbb{P} \otimes \mathbb{Q}$

$[\boldsymbol{\pi}]_1 = \mathbb{P}, [\boldsymbol{\pi}]_2 = \mathbb{Q} \Rightarrow \boldsymbol{\pi} \ll \pi_0$

vs

but in WDRO, $[\pi_0]_2$ not fixed !

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

Bottomline: (entropic) regularization – for both OT and WDRO

- helps numerically (on-going research for WDRO...)
- helps in theory, especially against curse of dimension (end of this talk)

Regularization helps in theory #1, sanity check: approximation bounds for WDRO

Inspired by [Genevay, Chizat, et al. '19] (bound on the approximation error for regularized OT)

Dual WDRO: $$(P) \quad \min_{\lambda \geqslant 0} \; \lambda \rho^p + \mathbb{E}_{\mathbb{P}}[\max_{\xi'} \{f(\xi') - \lambda \|\xi - \xi'\|^p\}]$$

Dual WDRO regularized by $R(\pi) = \varepsilon\, \mathrm{KL}(\pi|\pi_0)$ and $S(\pi) = \delta\, \mathrm{KL}(\pi|\pi_0)$

$$(P_{\varepsilon,\delta}) \quad \min_{\lambda \geqslant 0} \; \lambda \rho^p + (\varepsilon + \lambda\delta)\mathbb{E}_{\mathbb{P}} \log \left(\mathbb{E}_{\xi' \sim \pi_0(\cdot|\xi)} e^{\frac{f(\xi') - \lambda \|\xi - \xi'\|^p}{\varepsilon + \lambda \delta}} \right)$$

Theorem ([Azizian, Iutzeler, M. '21])

Under mild assumptions (non-degeneracy, lipschitz), if the support of $\mathbb{P}$ is contained in a compact convex set $\Xi \subset \mathbb{R}^d$, then

$$0 \;\leqslant\; \mathrm{val}(P) - \mathrm{val}(P_{\varepsilon,\delta}) \;\leqslant\; C\, d\, (\varepsilon + \overline{\lambda}\delta) \log \frac{1}{\varepsilon + \overline{\lambda}\delta}$$

where $\overline{\lambda} = \frac{2 \sup_{\Xi} |f|}{\rho^p - \mathbb{E}_{\pi_0} c}$ an explicit dual upper bound.

(the proof uses techniques from [Carlier *et al* '17])

Regularization helps in theory #2: generalization results for WDRO

Data $\xi_1, \ldots, \xi_N \sim \mathbb{P}_{\text{train}}$; computation with $\widehat{\mathbb{P}}_N = \frac{1}{N}\sum_{i=1}^N \delta_{\xi_i}$; guarantees with $\mathbb{P}_{\text{train}}$?

Regularization helps in theory #2: generalization results for WDRO

Data $\xi_1, \dots, \xi_N \sim \mathbb{P}_{\text{train}}$; computation with $\widehat{\mathbb{P}}_N = \frac{1}{N}\sum_{i=1}^N \delta_{\xi_i}$; guarantees with $\mathbb{P}_{\text{train}}$?

OT theory: $W(\mathbb{P}_{\text{train}}, \widehat{\mathbb{P}}_N) \leqslant O(1/\sqrt[d]{N})$ (with high probability) [Fournier, Guillin '15]

WDRO consequence [Esfahani, Kuhn '18]: if $\rho \geqslant O(1/\sqrt[d]{N})$, for all $f \in \mathcal{F}$

$$\mathbb{E}_{\mathbb{P}_{\text{train}}}[f(\xi)] \leqslant \max_{W(\widehat{\mathbb{P}}_N, \mathbb{Q}) \leqslant \rho} \mathbb{E}_{\mathbb{Q}}[f(\xi)] \qquad \text{(with high probability)}$$

Regularization helps in theory #2: generalization results for WDRO

Data $\xi_1, \ldots, \xi_N \sim \mathbb{P}_{\text{train}}$; computation with $\widehat{\mathbb{P}}_N = \frac{1}{N}\sum_{i=1}^N \delta_{\xi_i}$; guarantees with $\mathbb{P}_{\text{train}}$?

OT theory: $W(\mathbb{P}_{\text{train}}, \widehat{\mathbb{P}}_N) \leqslant O(1/\sqrt[d]{N})$ (with high probability) [Fournier, Guillin '15]

WDRO consequence [Esfahani, Kuhn '18]: if $\rho \geqslant O(1/\sqrt[d]{N})$, for all $f \in \mathcal{F}$

$$\mathbb{E}_{\mathbb{P}_{\text{train}}}[f(\xi)] \leqslant \max_{W(\widehat{\mathbb{P}}_N, \mathbb{Q}) \leqslant \rho} \mathbb{E}_{\mathbb{Q}}[f(\xi)] \qquad \text{(with high probability)}$$

WDRO direct: [An, Gao '22] drops the dependence on d by taking $\rho \propto 1/\sqrt{N}$

Regularization helps in theory #2: generalization results for WDRO

Data $\xi_1, \ldots, \xi_N \sim \mathbb{P}_{\text{train}}$; computation with $\widehat{\mathbb{P}}_N = \frac{1}{N}\sum_{i=1}^N \delta_{\xi_i}$; guarantees with $\mathbb{P}_{\text{train}}$?

OT theory: $W(\mathbb{P}_{\text{train}}, \widehat{\mathbb{P}}_N) \leqslant O(1/\sqrt[d]{N})$ (with high probability) [Fournier, Guillin '15]

WDRO consequence [Esfahani, Kuhn '18]: if $\rho \geqslant O(1/\sqrt[d]{N})$, for all $f \in \mathcal{F}$

$$\mathbb{E}_{\mathbb{P}_{\text{train}}}[f(\xi)] \leqslant \max_{W(\widehat{\mathbb{P}}_N, \mathbb{Q}) \leqslant \rho} \mathbb{E}_{\mathbb{Q}}[f(\xi)] \qquad \text{(with high probability)}$$

WDRO direct: [An, Gao '22] drops the dependence on d by taking $\rho \propto 1/\sqrt{N}$

With regularization: we can do even better ! (when $p = 2$, $\varepsilon > 0$, and Ξ compact)

Theorem (very informal, [Azizian, Iutzeler, M. '22])

If $\rho \geqslant \rho_N = O(1/\sqrt{N})$, then for all $f \in \mathcal{F}$

$$\mathbb{E}_{\mathbb{P}_{train}}[f] \leqslant F^{\varepsilon}_{\rho-\rho_N}(f, \mathbb{P}_{train}) \leqslant F^{\varepsilon}_{\rho}(f, \widehat{\mathbb{P}}_N) \qquad \textit{(with high probability)}$$

$$F^{\varepsilon}_{\rho}(f, \mathbb{P}) = \begin{cases} \max_{\boldsymbol{\pi}} & \mathbb{E}_{[\boldsymbol{\pi}]_2}[f(\xi)] - \varepsilon\,\mathrm{KL}(\boldsymbol{\pi}|\pi_0) \\ & [\boldsymbol{\pi}]_1 = \mathbb{P} \\ & \mathbb{E}_{\boldsymbol{\pi}}[\|\xi - \xi'\|^2] \leqslant \rho^2 \end{cases}$$

Conclusion

Main take-aways

- More work is needed on robustness in learning
- Distributionally robust optimization DRO is rich, active topic
- Our current work: extend the toolkit of DRO by regularization, inspired by OT
  (general duality, approximation results, worst-case distribution, statistical guarantees)

On-going work

- Wrap up the paper on generalisation
- Further investigate the computational aspects !
- Further investigate applications... (in fairness?)
  (first sucess in federated learning [Laguel, Pillutla, M., Harchaoui])

Conclusion

Main take-aways

- More work is needed on robustness in learning
- Distributionally robust optimization DRO is rich, active topic
- Our current work: extend the toolkit of DRO by regularization, inspired by OT
  (general duality, approximation results, worst-case distribution, statistical guarantees)

On-going work

- Wrap up the paper on generalisation
- Further investigate the computational aspects !
- Further investigate applications... (in fairness?)
  (first sucess in federated learning [Laguel, Pillutla, M., Harchaoui])

thank you all !