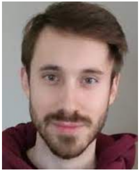
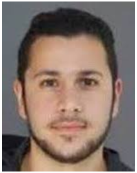
Distributionally robust optimization: regularization and applications in learning

Jérôme MALICK

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

ROADEF – Lyon – Feb. 2022

Based on joint work with
Waïss Azizian, Franck Iutzeler, Yassine Laguel

Robust ML/IA

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

Robust ML/IA

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

Example 1: Changes in environnements

Learning to drive in California vs. driving in the Alps

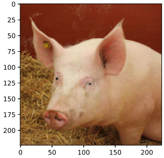
Robust ML/IA

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

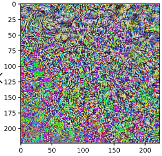
Example 1: Changes in environnements

Learning to drive in California vs. driving in the Alps

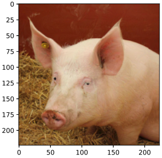
Example 2: Attacks [tutorial on robustness @ NeurIPS '18] (+ ROADEF '20 !)



$+\ \mathcal{E} \times$



$=$



pig (99%) airliner (96%)

Robust ML/IA

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

Example 1: Changes in environnements

Learning to drive in California vs. driving in the Alps

Example 2: Attacks [@ CVPR '18]

Robust ML/IA

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

Example 1: Changes in environnements

Learning to drive in California vs. driving in the Alps

Example 2: Attacks [@ ICLR '19]

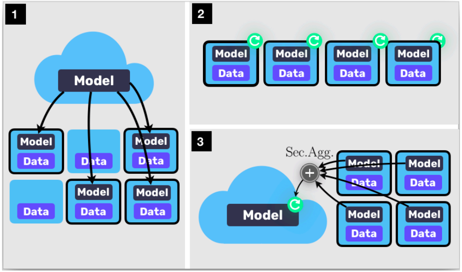
Robust ML/IA

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

Example 3: Data heterogeneity

E.g. in federated learning

Google, hospital consortium...



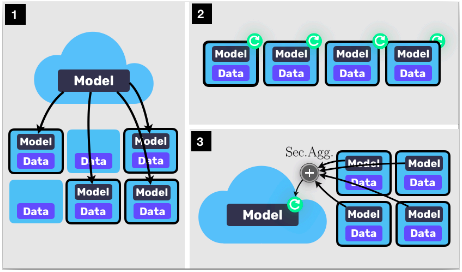
Robust ML/IA

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

Example 3: Data heterogeneity

E.g. in federated learning

Google, hospital consortium...

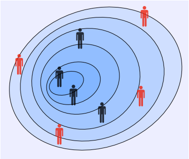


What about non-conforming users ?

Many issues !

(service quality ? fairness issues ?...)

more later...



Data-driven set-up

- Training data: $\xi_1, \ldots, \xi_N \sim \mathbb{P}$ (unknown)

  e.g. in supervised learning: $\xi_i = (a_i, y_i)$ feature, label

- Train model: x the parameter/decision, $f(x, \cdot)$ the loss

  e.g. least-square regression: $f\big(x, (a, y)\big) = (x^\top a - y)^2$

- 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 slightly different data ξ ?

  (generalisation, data shifts, adversarial examples,...)

  Take variation into account when optimizing/learning !

Data-driven set-up

- Training data: $\xi_1, \ldots, \xi_N \sim \mathbb{P}$ (unknown)
  e.g. in supervised learning: $\xi_i = (a_i, y_i)$ feature, label
- Train model: x the parameter/decision, $f(x, \cdot)$ the loss
  e.g. least-square regression: $f\big(x, (a, y)\big) = (x^\top a - y)^2$
- 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 slightly different data ξ ?
  (generalisation, data shifts, adversarial examples,...)
  Take variation into account when optimizing/learning !

Data-driven set-up

- Training data: $\xi_1, \ldots, \xi_N \sim \mathbb{P}$ (unknown)

  e.g. in supervised learning: $\xi_i = (a_i, y_i)$ feature, label

- Train model: x the parameter/decision, $f(x, \cdot)$ the loss

  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 slightly different data ξ ?

  (generalisation, data shifts, adversarial examples,...)

  Take variation into account when optimizing/learning !

- (Distributionally) robust optimization

  (optimize expected loss for a the worst case in a set of perturbation)

$$\min_x \max_{\mathbb{Q} \in \mathcal{U}} \mathbb{E}_{\mathbb{Q}}[f(x, \xi)]$$

Modeling issues

E.g. ambiguity/incertainty set $\mathcal{U}$: $$\min_{x} \max_{\mathbb{Q}\in\mathcal{U}} \mathbb{E}_{\mathbb{Q}}[f(x,\xi)]$$

- $\mathcal{U} = \left\{\widehat{\mathbb{P}}_N\right\}$: $\quad \min_{x} \frac{1}{N}\sum_{i=1}^{N} f(x,\xi_i)$
- $\mathcal{U} = \{\mathbb{Q} : \mathrm{supp}(\mathbb{Q}) \subset U\}$: $\quad \min_{x} \max_{\xi\in U} f(x,\xi)$
- $\mathcal{U}$ defined by moments e.g. [Delage, Ye, '10]
- $\mathcal{U} = \left\{\mathbb{Q} : \ d(\mathbb{Q},\widehat{\mathbb{P}}_N) \leqslant \rho\right\}$ for various distances or divergences
  E.g. KL-div., χ_2-div., max-mean-discrepancy... e.g. [Namkoong, Duchi '17]
- $\mathcal{U} = \left\{\mathbb{Q} : \ W(\mathbb{Q},\widehat{\mathbb{P}}_N) \leqslant \rho\right\}$ Wasserstein distance (in this talk)
  Good statistical/practical properties... e.g. [Kuhn *et al.* '18]

DRO in action #1 : toy example

Least-square linear regression

Data : $\xi_1, \xi_2, \ldots, \xi_N$ with $\xi_i = (a_i, y_i)$ in two groups (majority vs. minority)

$y_i = \bar{x}^\top a_i + \varepsilon_i$ with $\varepsilon_i \sim \beta \mathcal{N}^{\text{major}} + (1-\beta)\mathcal{N}^{\text{minor}}$

Compute from data:

standard regression x^{ERM} vs. DRO regression x^{DRO} (KL-regularized)

DRO in action #1 : toy example

Least-square linear regression

Data : $\xi_1, \xi_2, \ldots, \xi_N$ with $\xi_i = (a_i, y_i)$ in two groups (majority vs. minority)

$y_i = \bar{x}^\top a_i + \varepsilon_i$ with $\varepsilon_i \sim \beta \mathcal{N}^{\text{major}} + (1-\beta)\mathcal{N}^{\text{minor}}$

Compute from data:

standard regression x^{ERM} vs. DRO regression x^{DRO} (KL-regularized)

Generate new data $\xi'_1, \ldots, \xi'_M$

Test the regression errors given by x^{ERM} vs x^{DRO} $(r_i = |x^\top a_i - y_i|)$

DRO in action #1 : toy example

Least-square linear regression

Data : $\xi_1, \xi_2, \ldots, \xi_N$ with $\xi_i = (a_i, y_i)$ in two groups (majority vs. minority)

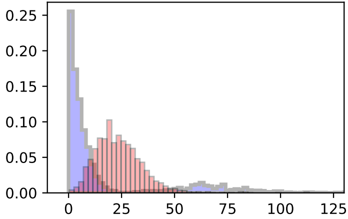
$y_i = \bar{x}^\top a_i + \varepsilon_i$ with $\varepsilon_i \sim \beta\mathcal{N}^{\text{major}} + (1-\beta)\mathcal{N}^{\text{minor}}$

Compute from data:

standard regression x^{ERM} vs. DRO regression x^{DRO} (KL-regularized)

Generate new data $\xi'_1, \ldots, \xi'_M$

Test the regression errors given by x^{ERM} vs x^{DRO} $(r_i = |x^\top a_i - y_i|)$



Histogram of the regression errors for unseen data

DRO in action #1 : toy example

Least-square linear regression

Data : $\xi_1, \xi_2, \ldots, \xi_N$ with $\xi_i = (a_i, y_i)$ in two groups (majority vs. minority)

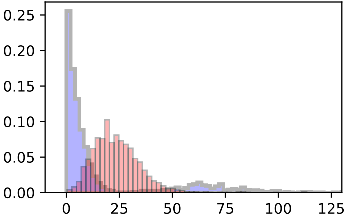
$y_i = \bar{x}^\top a_i + \varepsilon_i$ with $\varepsilon_i \sim \beta\mathcal{N}^{\text{major}} + (1-\beta)\mathcal{N}^{\text{minor}}$

Compute from data:

standard regression x^{ERM} vs. DRO regression x^{DRO} (KL-regularized)

Generate new data $\xi'_1, \ldots, \xi'_M$

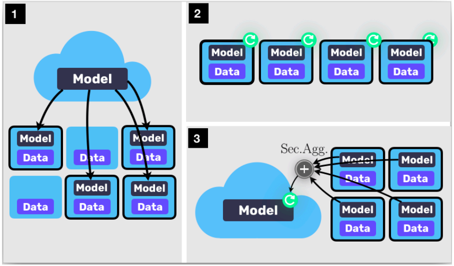
Test the regression errors given by x^{ERM} vs x^{DRO} $(r_i = |x^\top a_i - y_i|)$



Histogram of the regression errors for unseen data

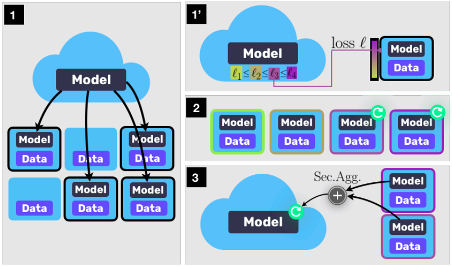
DRO re-shapes histograms towards more fairness ☺

DRO in action #2 : federated learning with heterogeneous users



Federated Learning by Google = FedAvg

DRO in action #2 : federated learning with heterogeneous users



Federated Learning by Google = FedAvg vs. DRO FedAvg

[Laguel, Pillutla, M., Harchaoui '21]

Illustration:

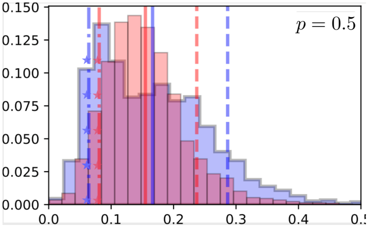
Classification task by ConvNet

with EMNIST dataset

(1730 users, 179 images/users)

Histogram over users
of test misclassification error

(dashed lines: 10%/90% -percentiles)



Research topic: extend the (W)DRO toolkit

- DRO works well ☺
- Trade-off : $\boxed{\text{modeling vs. computational tractability}}$
- Wasserstein-DRO is popular...
  Good statistical/practical properties, e.g. [Kuhn *et al.* '18]
- ...but has some limitations ! news results
- We propose: Regularized WDRO [Azizian, Iutzeler, M. '22]
- Why regularizing ? it helps computationnally !
  One of the main reasons of the popularity of OT in ML [Cuturi '13]
- On-going research...

DRO with Wasserstein balls as ambiguity sets

Def: Wasserstein distance (given a cost function c)

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

DRO with Wasserstein balls as ambiguity sets

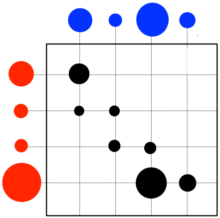
Def: Wasserstein distance (given a cost function c)

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

Demystification: in the discrete case

e.g. $\mathbb{P} = (p_1,\dots,p_N)$ and $\mathbb{Q} = (q_1,\dots,q_N)$ in the simplex

$$\begin{cases} \min_{\boldsymbol{\pi}} \quad \sum_{i,j=1}^N c_{i,j}\,\boldsymbol{\pi}_{i,j} \\ \quad \sum_{j=1}^N \boldsymbol{\pi}_{i,j} = p_i \quad i=1,\dots,N \\ \quad \sum_{i=1}^N \boldsymbol{\pi}_{i,j} = q_j \quad j=1,\dots,N \\ \quad \boldsymbol{\pi}_{i,j} \geqslant 0 \quad i,j=1,\dots,N \end{cases}$$



linear assignment !

DRO with Wasserstein balls as ambiguity sets

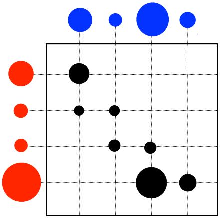
Def: Wasserstein distance (given a cost function c)

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

Demystification: in the discrete case

e.g. $\mathbb{P} = (p_1, \ldots, p_N)$ and $\mathbb{Q} = (q_1, \ldots, q_N)$ in the simplex

$$\begin{cases} \min_{\boldsymbol{\pi}} \quad \sum_{i,j=1}^{N} c_{i,j}\, \boldsymbol{\pi}_{i,j} \\ \quad \sum_{j=1}^{N} \boldsymbol{\pi}_{i,j} = p_i \quad i = 1, \ldots, N \\ \quad \sum_{i=1}^{N} \boldsymbol{\pi}_{i,j} = q_j \quad j = 1, \ldots, N \\ \quad \boldsymbol{\pi}_{i,j} \geqslant 0 \quad i,j = 1, \ldots, N \end{cases}$$



linear assignment !

Wasserstein-DRO (WDRO) objective for given $\mathbb{P}$ and ρ

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

DRO with Wasserstein balls as ambiguity sets

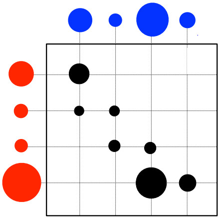
Def: Wasserstein distance (given a cost function c)

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

Demystification: in the discrete case

e.g. $\mathbb{P} = (p_1,\ldots,p_N)$ and $\mathbb{Q} = (q_1,\ldots,q_N)$ in the simplex

$$\begin{cases} \min_{\boldsymbol{\pi}} \quad \sum_{i,j=1}^N c_{i,j}\,\boldsymbol{\pi}_{i,j} \\ \quad \sum_{j=1}^N \boldsymbol{\pi}_{i,j} = p_i \quad i=1,\ldots,N \\ \quad \sum_{i=1}^N \boldsymbol{\pi}_{i,j} = q_j \quad j=1,\ldots,N \\ \quad \boldsymbol{\pi}_{i,j} \geqslant 0 \quad i,j=1,\ldots,N \end{cases}$$



linear assignment !

Wasserstein-DRO (WDRO) objective for given $\mathbb{P}$ and ρ

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

DRO with Wasserstein balls as ambiguity sets

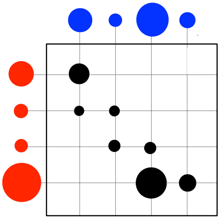
Def: Wasserstein distance (given a cost function c)

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

Demystification: in the discrete case

e.g. $\mathbb{P} = (p_1,\dots,p_N)$ and $\mathbb{Q} = (q_1,\dots,q_N)$ in the simplex

$$\begin{cases} \min_{\boldsymbol{\pi}} \quad \sum_{i,j=1}^N c_{i,j}\,\boldsymbol{\pi}_{i,j} \\ \quad \sum_{j=1}^N \boldsymbol{\pi}_{i,j} = p_i \quad i=1,\dots,N \\ \quad \sum_{i=1}^N \boldsymbol{\pi}_{i,j} = q_j \quad j=1,\dots,N \\ \quad \boldsymbol{\pi}_{i,j} \geqslant 0 \quad i,j=1,\dots,N \end{cases}$$



linear assignment !

Wasserstein-DRO (WDRO) objective for given $\mathbb{P}$ and ρ

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

WDRO : duality

Primal WDRO

$$\begin{cases} \max_{\boldsymbol{\pi}} \quad \mathbb{E}_{[\boldsymbol{\pi}]_2}[f(\xi)] \\ \quad [\boldsymbol{\pi}]_1 = \mathbb{P} \\ \quad \mathbb{E}_{\boldsymbol{\pi}}[c(\xi,\xi')] \leqslant \rho \qquad \leftarrow \lambda \geqslant 0 \end{cases}$$

Dual WDRO

$$\min_{\lambda \geqslant 0} \qquad \lambda\rho + \mathbb{E}_{\mathbb{P}}[\max_{\xi'} f(\xi') - \lambda c(\xi,\xi')]$$

WDRO : duality

Primal WDRO regularized (with two convex functions R, S)

$$\left\{\begin{array}{ll} \max_{\boldsymbol{\pi}} & \mathbb{E}_{[\boldsymbol{\pi}]_2}[f(\xi)] - R(\boldsymbol{\pi}) \\ & [\boldsymbol{\pi}]_1 = \mathbb{P} \\ & \mathbb{E}_{\boldsymbol{\pi}}[c(\xi,\xi')] + S(\boldsymbol{\pi}) \leqslant \rho \qquad \leftarrow \lambda \geqslant 0 \end{array}\right.$$

Dual WDRO when regularized

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

WDRO : duality

Primal WDRO regularized (with two convex functions R, S)

$$\left\{\begin{array}{l} \max_{\boldsymbol{\pi}} \quad \mathbb{E}_{[\boldsymbol{\pi}]_2}[f(\xi)] - R(\boldsymbol{\pi}) \\ \quad [\boldsymbol{\pi}]_1 = \mathbb{P} \\ \quad \mathbb{E}_{\boldsymbol{\pi}}[c(\xi,\xi')] + S(\boldsymbol{\pi}) \leqslant \rho \qquad \leftarrow \lambda \geqslant 0 \end{array}\right.$$

Dual WDRO when regularized

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

Quite abstract... but more concrete expressions when specialized
e.g. with $R(\pi) = \varepsilon\,\mathrm{KL}(\pi|\pi_0)$ and $S(\pi) = \delta\,\mathrm{KL}(\pi|\pi_0)$ for a given π_0

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

WDRO: approximation result

Dual WDRO:

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

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

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

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

Under mild assumptions (non-degeneracy, lipschitz, $c = \|\cdot\|^p$, special form of π_0), 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 $\overline{\lambda} = \frac{2\sup_{\Xi}|f|}{\rho - \mathbb{E}_{\pi_0} c}$ an explicit dual bound.

We control the error... Next step: solve $(P_{\varepsilon,\delta})$ efficiently, another story...

Conclusion

Main take-aways

- DRO is a rich field + promising approach in ML
- Our work : extend the toolkit of DRO
- Proposal : use regularized WDRO !

  general duality, approximation results, worst-case distribution...

On-going work on regularized WDRO

- Towards scalable algorithms...
- Statistical guarantees ?
- Applications ? (fairness?)

Conclusion

Main take-aways

- DRO is a rich field + promising approach in ML
- Our work : extend the toolkit of DRO
- Proposal : use regularized WDRO !

  general duality, approximation results, worst-case distribution...

On-going work on regularized WDRO

- Towards scalable algorithms...
- Statistical guarantees ?
- Applications ? (fairness?)

thanks !!

Some references

Some important references (on DRO and related topics)

Daniel Kuhn, P.M. Esfahani, V. Anh Nguyen and S. Shafieezadeh-Abadeh
Wasserstein distributionally robust optimization: Theory and applications in ML
In Operations Research & Management Science in the Age of Analytics, 2019

Gabriel Peyré and Marco Cuturi
Computational optimal transport and applications to data science
Foundations and Trends in Machine Learning, 2019

Terry Rockafellar, Johannes Royset, Sofia Miranda.
Superquantile regression with applications to reliability, uncertainty quantification, and (...)
European Journal of Operational Research, 2014

Some references on our work

Yassine Laguel, Jérôme Malick, and Zaid Harchaoui
Optimization for Superquantile-based Supervised Learning
30th Workshop on Machine Learning for Signal Processing, 2020

Krishna Pillutla, Yassine Laguel, Jérôme Malick, and Zaid Harchaoui
Federated Learning with Heterogeneous Data: A Superquantile Optimization Approach
Submitted to Machine Learning Research, 2021

Azizian Waiss, Franck Iutzeler, and Jérôme Malick
Regularization for Wasserstein distributionnally robust optimization
Submitted to ESAIM: Control, Optimization, and Calculus of Variations, 2022

SFL comparison w. state-of-the-art

From [Laguel, Pillutla, M., Harchaoui '21]

		90th Percentile		Average	
		Linear	ConvNet	Linear	ConvNet
	Δ-FL $p = 0.5$	$\mathbf{46.48} \pm 0.38$	$\mathbf{23.69} \pm 0.94$	35.02 ± 0.20	$\mathbf{15.49} \pm 0.30$
$\mathbb{E}$	FedAvg	49.66 ± 0.67	28.46 ± 1.07	34.38 ± 0.38	16.64 ± 0.50
prox	FedProx	49.15 ± 0.74	27.01 ± 1.86	$\mathbf{33.82} \pm 0.30$	16.02 ± 0.54
$\Vert \cdot \Vert_q^q$ $(q > 1)$	q-FFL	49.90 ± 0.58	28.02 ± 0.80	34.34 ± 0.33	16.59 ± 0.30
max	AFL	51.62 ± 0.28	45.08 ± 1.00	39.33 ± 0.27	33.01 ± 0.37

Regularized WDRO

From [Azizian, Iutzeler, M. '22]

- Recall : KL (Kullback-Lieber divergence)

$$\mathrm{KL}(\mu|\nu) = \begin{cases} \int \log \frac{\mathrm{d}\mu}{\mathrm{d}\nu}\ \mathrm{d}\mu & \text{if } \mu,\nu \geqslant 0 \text{ and } \mu \ll \nu \\ +\infty & \text{otherwise} \end{cases}$$

In the discrete case: $\mathbb{P} = (p_1, \ldots, p_N)$ and $\mathbb{Q} = (q_1, \ldots, q_N)$

$$\mathrm{KL}(\mathbb{P}|\mathbb{Q}) = \sum_{i=1}^{N} p_i \log \frac{p_i}{q_i}$$

- Explicit reference measure

$$\pi_0(\mathrm{d}\xi, \mathrm{d}\xi') \ \propto \ \mathbb{P}(\mathrm{d}\xi)\ \mathbb{I}_{\xi' \in \Xi} e^{-\frac{\|\xi - \xi'\|^p}{2^{p-1}\sigma}}\ \mathrm{d}\xi'$$

- Worst-case distribution

$$\mathbb{P}^* = (...) \text{ supported on the whole space}$$

vs. WDRO where the worst-case is finitely supported...

(WDRO hedges against wrong set of distributions ?)