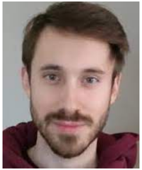
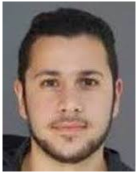
(Wasserstein) distributionally robust optim. in action

Jérôme MALICK

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

Data-driven optim. workshop – Ecole des Ponts – Oct. 2023

Based on joint work with
Waïss Azizian, Franck Iutzeler, Yassine Laguel, Zaid Harchaoui, Krishna Pillutla

Set-up: data-driven optimization under uncertainty

- Training data: $\xi_1, \ldots, \xi_N$ (in theory: sampled from $\mathbb{P}_{\text{train}}$ unknown)
  e.g. in supervised learning: labeled data $\xi_i = (a_i, y_i)$ feature, label

- Train model: $f(x, \cdot)$ the loss function with x the parameter/decision $(\omega, \beta, \theta, \ldots)$
  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 different data ξ
  - Distributional shifts: $\mathbb{P}_{\text{train}} \neq \mathbb{P}_{\text{test}}$
  - Adversarial attacks
  - Generalization: computations with $\widehat{\mathbb{P}}_N$ and guarantees on $\mathbb{P}_{\text{train}}$
  - Other situations, e.g. heterogeneous data

- Solution: take possible variations into account during training

Set-up: data-driven optimization under uncertainty

- Training data: $\xi_1, \dots, \xi_N$ (in theory: sampled from $\mathbb{P}_{\text{train}}$ unknown)

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

- Train model: $f(x, \cdot)$ the loss function with x the parameter/decision $(\omega, \beta, \theta, ...)$

  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 different data ξ
  - Distributional shifts: $\mathbb{P}_{\text{train}} \neq \mathbb{P}_{\text{test}}$
  - Adversarial attacks
  - Generalization: computations with $\widehat{\mathbb{P}}_N$ and guarantees on $\mathbb{P}_{\text{train}}$
  - Other situations, e.g. heterogeneous data

- Solution: take possible variations into account during training

(Distributionally) robust optimization

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

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

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

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

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

$$\boxed{\text{modeling vs. computational tractability}}$$

- $\mathcal{U} = \left\{\widehat{\mathbb{P}}_N\right\}$: $\min_{x} \frac{1}{N}\sum_{i=1}^{N} f(x,\xi_i)$ standard ERM
- $\mathcal{U} = \{\mathbb{Q} : \mathrm{supp}(\mathbb{Q}) \subset U\}$: $\min_{x} \max_{\xi\in U} f(x,\xi)$ standard robust optimization
- $\mathcal{U}$ 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]

Wasserstein balls as ambiguity sets

Def: Wasserstein distance (given a cost function c)

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

Wasserstein balls as ambiguity sets

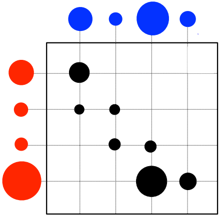
Def: Wasserstein distance (given a cost function c)

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

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 balls as ambiguity sets

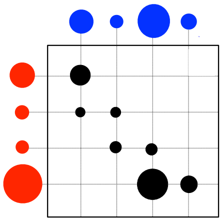
Def: Wasserstein distance (given a cost function c)

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

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

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

Wasserstein balls as ambiguity sets

Def: Wasserstein distance (given a cost function c)

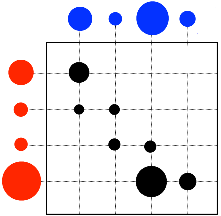
$$W(\mathbb{P}, \mathbb{Q}) = \min_{\boldsymbol{\pi}} \left\{ \mathbb{E}_{\boldsymbol{\pi}}[c(\xi, \xi')] : \boldsymbol{\pi} \text{ with marginals } [\boldsymbol{\pi}]_1 = \mathbb{P} \text{ and } [\boldsymbol{\pi}]_2 = \mathbb{Q} \right\}$$

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

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

Wasserstein balls as ambiguity sets

Def: Wasserstein distance (given a cost function c)

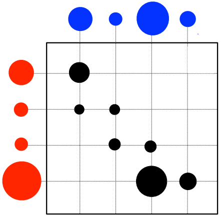
$$W(\mathbb{P},\mathbb{Q}) = \min_{\boldsymbol{\pi}}\left\{ \mathbb{E}_{\boldsymbol{\pi}}[c(\xi,\xi')] : \boldsymbol{\pi} \text{ with marginals } [\boldsymbol{\pi}]_1 = \mathbb{P} \text{ and } [\boldsymbol{\pi}]_2 = \mathbb{Q}\right\}$$

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

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

Wasserstein DRO (WDRO)

WDRO objective for given $\mathbb{P}$, ρ – and choice $c(\xi, \xi') = \|\xi - \xi'\|^2$

(Primal)
$$\begin{cases} \max_{\boldsymbol{\pi}} \quad \mathbb{E}_{[\boldsymbol{\pi}]_2}[f(\xi)] \\ \quad [\boldsymbol{\pi}]_1 = \mathbb{P} \\ \quad \mathbb{E}_{\boldsymbol{\pi}}[\|\xi - \xi'\|^2] \leqslant \rho \end{cases}$$

(Dual)
$$\min_{\lambda \geqslant 0} \; \lambda\rho + \mathbb{E}_{\mathbb{P}}[\, \max_{\xi'} \{f(\xi') - \lambda\|\xi - \xi'\|^2\} \,]$$

Wasserstein DRO (WDRO)

WDRO objective for given $\mathbb{P}$, ρ – and choice $c(\xi, \xi') = \|\xi - \xi'\|^2$

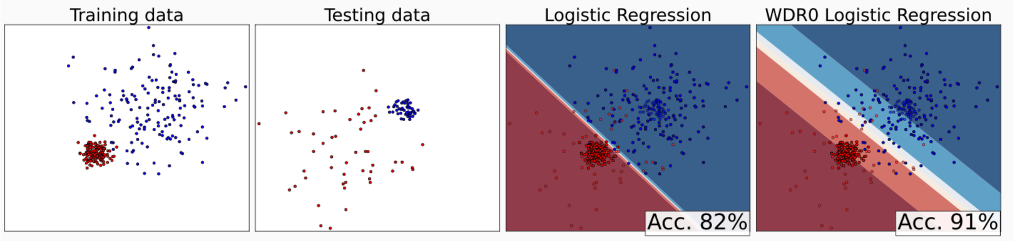
(Primal)
$$\begin{cases} \max_{\boldsymbol{\pi}} \quad \mathbb{E}_{[\boldsymbol{\pi}]_2}[f(\xi)] \\ \quad [\boldsymbol{\pi}]_1 = \mathbb{P} \\ \quad \mathbb{E}_{\boldsymbol{\pi}}[\|\xi - \xi'\|^2] \leqslant \rho \end{cases}$$

(Dual)
$$\min_{\lambda \geqslant 0} \; \lambda\rho + \mathbb{E}_{\mathbb{P}}[\, \max_{\xi'} \{f(\xi') - \lambda\|\xi - \xi'\|^2\} \,]$$

Ex: classification. Compute separator θ by min. logistic loss $f_\theta(\xi) = \log(1 + \exp(-y\, a^\top \theta))$

$\xi_i = (a_i, y_i) \in \mathbb{R}^2 \times \{-1, +1\}$ sampled from two Gaussian distributions

with variances $\sigma = 1$ and $\sigma = 5$ – reversed in testing !



WDRO : success !(?)

WDRO is very attractive

- Statistical/practical properties
  e.g. [Blanchet *et al.* '18] and [Blanchet and Shapiro '23]
- Computable in many cases
  e.g. [Kuhn *et al.* '18], [Zhao Guan '18]...
- Natural in many applications
  back to [Scarf 1958] ! + (...) + recent trend in learning, e.g. [Kuhn *et al.* '20]
- Interprets up to first-order as a penalization by $\|\nabla_\xi f(x, \xi)\|$ e.g. [Gao *et al.* '18]

WDRO : success !(?)

WDRO is very attractive – in theory and in pratice ?

- Statistical/practical properties – warning : dimensionality ! (spotlight #1)
  e.g. [Blanchet *et al.* '18] and [Blanchet and Shapiro '23]
- Computable in many cases – but not always ! (spotlight #2)
  e.g. [Kuhn *et al.* '18], [Zhao Guan '18]...
- Natural in many applications – but not always ! (spotlight #3)
  back to [Scarf 1958] ! + (...) + recent trend in learning, e.g. [Kuhn *et al.* '20]
- Interprets up to first-order as a penalization by $\|\nabla_\xi f(x, \xi)\|$ e.g. [Gao *et al.* '18]

spotlight #1 : generalization guarantees

Azizian Waiss, Franck Iutzeler, and Jérôme Malick
Excat generalization guarantees for (regularized) WDRO models
Just accepted in NeurIPS, 2023

Existing generalization guarantees

- Suppose $\xi_1, \ldots, \xi_N \sim \mathbb{P}_{\text{train}}$ (where $\xi \in \mathbb{R}^d$)
- Computations with $\widehat{\mathbb{P}}_N = \frac{1}{N}\sum_{i=1}^{N} \delta_{\xi_i}$... guarantees with $\mathbb{P}_{\text{train}}$?
- We manipulate the WDRO risk $R_\rho(f) = \max_{W(\widehat{\mathbb{P}}_N, \mathbb{Q}) \leqslant \rho} \mathbb{E}_{\mathbb{Q}}[f(\xi)]$
- Obviously, if ρ, N large enough such that $W(\mathbb{P}_{\text{train}}, \widehat{\mathbb{P}}_N) \leqslant \rho$, then

$$\underbrace{R_\rho(f)}_{\text{can compute \& optimize}} \geqslant \underbrace{\mathbb{E}_{\mathbb{P}_{\text{train}}}[f(\xi)]}_{\text{cannot access}}$$

Existing generalization guarantees

- Suppose $\xi_1, \ldots, \xi_N \sim \mathbb{P}_{\text{train}}$ (where $\xi \in \mathbb{R}^d$)
- Computations with $\widehat{\mathbb{P}}_N = \frac{1}{N}\sum_{i=1}^{N} \delta_{\xi_i}$... guarantees with $\mathbb{P}_{\text{train}}$?
- We manipulate the WDRO risk $R_\rho(f) = \max_{W(\widehat{\mathbb{P}}_N, \mathbb{Q}) \leqslant \rho} \mathbb{E}_{\mathbb{Q}}[f(\xi)]$
- Obviously, if ρ, N large enough such that $W(\mathbb{P}_{\text{train}}, \widehat{\mathbb{P}}_N) \leqslant \rho$, then

$$\underbrace{R_\rho(f)}_{\text{can compute \& optimize}} \geqslant \underbrace{\mathbb{E}_{\mathbb{P}_{\text{train}}}[f(\xi)]}_{\text{cannot access}}$$

- But it requires $\rho \propto 1/\sqrt[d]{N}$ [Fournier and Guillin '15]
- Not optimal: $\rho \propto 1/\sqrt{N}$ suffices
  - asymptotically [Blanchet *et al* '22]
  - in particular cases [Shafieez-Adehabadeh *et al* '19]
  - or with error terms [Gao '22]

Extended exact generalization guarantees

By a direct approach (work direct to get a concentration result on the (dual) objective)

Theorem ([Azizian, Iutzeler, M. '23])

Assumptions : compactness on ξ + compactness on f + quad. growth of f near its minimizers

For $\delta \in (0,1)$, *if* $\rho \geqslant O\left(\sqrt{\frac{\log 1/\delta}{N}}\right)$

Generalization guarantee: w.p. $1-\delta$, $\quad R_\rho(f) \geqslant \mathbb{E}_{\mathbb{P}_{train}}[f(\xi)]$

Distribution shifts: w.p. $1-\delta$,

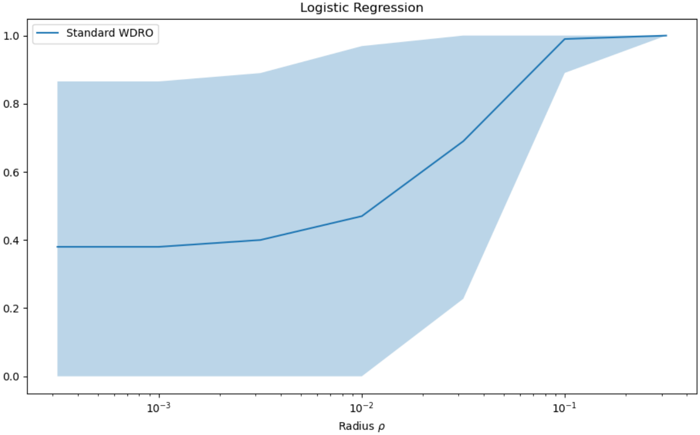
$$W(\mathbb{P},\mathbb{Q})^2 \leqslant \rho\left(\rho - O\left(\sqrt{\frac{\log 1/\delta}{N}}\right)\right) \quad \textit{it holds} \quad R_\rho(f) \geqslant \mathbb{E}_{\mathbb{Q}}[f(\xi)]$$

Assumptions valid in many cases: linear/logistic regression, kernel models, smooth neural networks, family of invertible mappings (e.g. normalizing flows)

Illustration

On logistic regression:

- for each ρ, sample 200 training datasets
- solve the WDRO problem on each of them [Blanchet *et al* '22]
- plot the proba of $R_\rho(f) - \mathbb{E}_{\mathbb{P}_{\text{train}}}[f] \geqslant 0$ (average, standard deviation)
- the training robust loss is indeed an upper-bound on the true loss



spotlight #2 – towards efficient computational toolkit: smoothed/regularized WDRO

Azizian Waiss, Franck Iutzeler, and Jérôme Malick
Regularization for Wasserstein distributionally robust optimization
ESAIM:COCV (Control Optim. Calculus of Variations), 2023

WDRO objective to be minimized

Dual WDRO is nonsmooth (which complicates resolution [Kuhn *et al.* '18])

$$R_\rho(f) = \min_{\lambda \geqslant 0} \; \lambda\rho + \mathbb{E}_{\mathbb{P}}\big[\max_{\xi'} \{f(\xi') - \lambda\|\xi - \xi'\|^2\}\big]$$

WDRO objective to be minimized

Dual WDRO is nonsmooth (which complicates resolution [Kuhn *et al.* '18])

$$R_\rho(f) = \min_{\lambda \geqslant 0} \ \lambda\rho + \mathbb{E}_{\mathbb{P}}[\max_{\xi'} \{f(\xi') - \lambda\|\xi - \xi'\|^2\}]$$

What about smoothing ? Smoothed counterpart

$$R_\rho^\varepsilon(f) = \min_{\lambda \geqslant 0} \ \lambda\rho + \varepsilon \, \mathbb{E}_{\mathbb{P}} \log \left(\mathbb{E}_{\xi' \sim \pi_0(\cdot|\xi)} e^{\frac{f(\xi') - \lambda\|\xi - \xi'\|^2}{\varepsilon}} \right)$$

(Nice interpretation as entropy-regularized WDRO)

Theorem (approximation bounds for WDRO [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

$$0 \ \leqslant \ R_\rho(f) - R_\rho^\varepsilon(f) \ \leqslant \ \left(C \, \varepsilon \log \frac{1}{\varepsilon} \right) d$$

Great ! but no computational resuls to show yet...

Smoothed ~~W~~DRO in action

Superquantile $\mathbb{S}_\theta$ [Rockfellar *et al* '00] (a.k.a. Conditional Value-at-Risk)

Risk measure with dual formulation

$$R_\theta(x) = \max_{q \in \Delta_n} \left\{ \sum_{i=1}^{n} q_i \, \ell(y_i, \varphi(x, a_i)) : 0 \leqslant q_i \leqslant \frac{1}{n(1-\theta)} \right\}$$

DRO with (smoothed) superquantile in Pytorch

https://github.com/krishnap25/sqwash

```
import torch.nn.functional as F
from sqwash import reduce_superquantile

for x, y in dataloader:
    y_hat = model(x)
    batch_losses = F.cross_entropy(y_hat, y, reduction='none')  # must set `reduction='none'`
    loss = reduce_superquantile(batch_losses, superquantile_tail_fraction=0.5)  # Additional line
    loss.backward()  # Proceed as usual from here
    ...
```

spotlight #3 : ~~W~~DRO for federated learning

Krishna Pillutla, Yassine Laguel, Jérôme Malick, Zaid Harchaoui
Federated Learning with Superquantile Aggregation for Heterogeneous Data
Machine Learning Journal, 2023

Federated learning in a nutschell

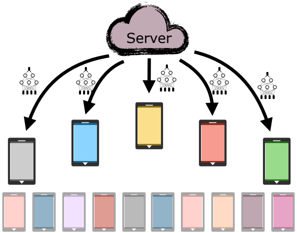
- Standard learning : get all the data and learn your model on it
- Efficient... but is privacy invasive (hospitals, european laws...)
- Idea : move the model not the data !

Federated learning in a nutschell

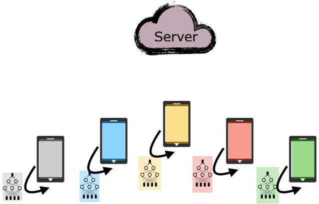
- Standard learning : get all the data and learn your model on it
- Efficient... but is privacy invasive (hospitals, european laws...)
- Idea : move the model not the data !
- Usual approach : FedAvg [McMahan *et al* 2017]

  (based on old ideas, e.g. [Mangasarian 1995])

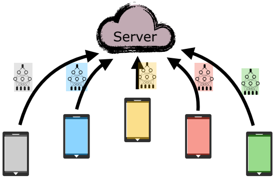
Step 1 of 3: Server broadcasts global model to sampled clients



Step 2 of 3: Clients perform some local SGD steps on their local data

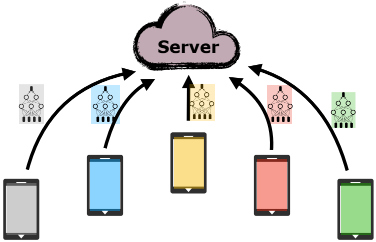


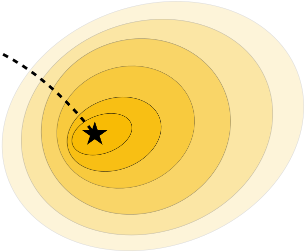
Step 3 of 3: Aggregate client updates securely



Issue of heterogeneous users

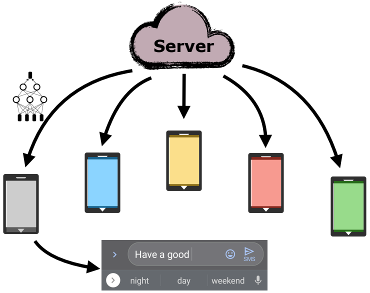
Global model is trained on *average distribution* across clients (ERM)

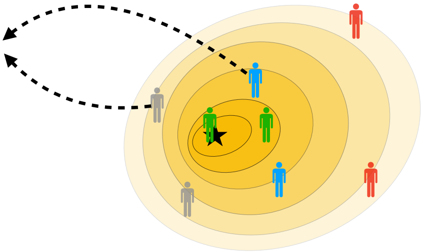




Issue of heterogeneous users

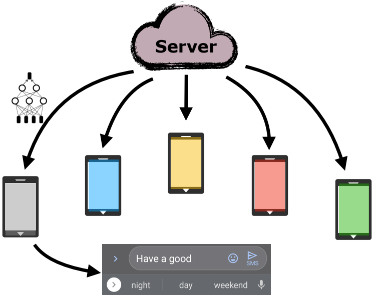
Global model is deployed on *individual* clients

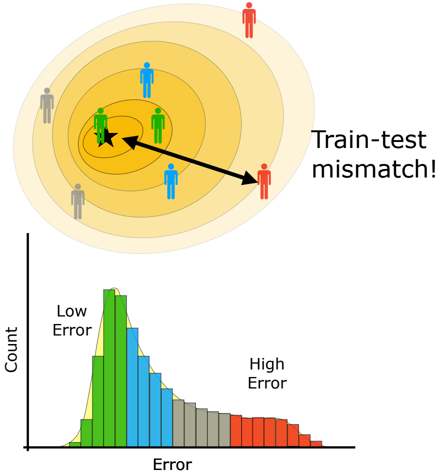




Issue of heterogeneous users

Global model is deployed on *individual* clients





Our solution : superquantile minimization

ERM Algorithm (FedAvg):

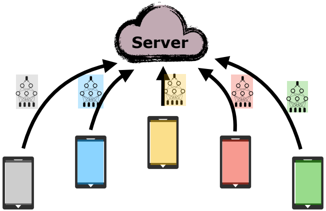
$$\min_{w} \quad \frac{1}{n}\sum_{i=1}^{n} F_i(w)$$

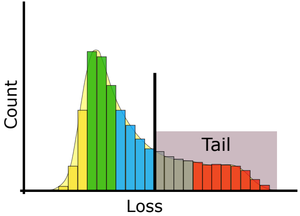
*Step 3 of 3: Aggregate updates contributed by **all clients***

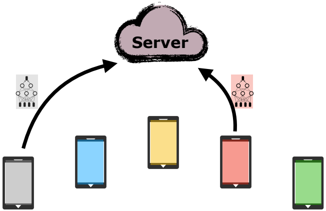
Simplicial-FL Algorithm:

$$\min_{w} \mathbb{S}_{\theta}\Big(\big(F_1(w), \cdots, F_n(w)\big) \Big)$$

*Step 3 of 3: Aggregate updates contributed by **tail clients** only*





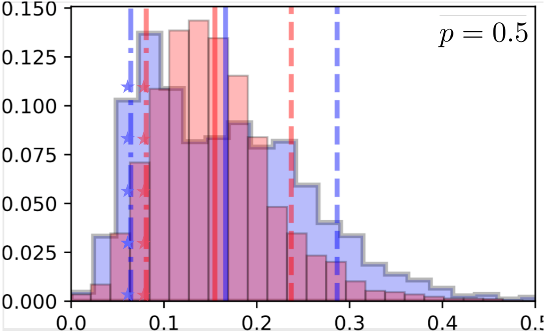


- Compatible with secure aggregation and differential privacy
- Analysis of the entropy-regularized version (both cvx and non-cvx)

Illustration

Classification task – ConvNet with EMNIST dataset (1730 users, 179 images/users)

Histogram over users of test misclassification error: standard vs. DRO
(dashed lines: 10%/90% -percentiles)



DRO reshapes test histograms

Conclusion

Main take-aways

- Distributionally robust optimization DRO is rich, active topic
  and has real-life applications, as in federated learning
- WDRO has nice generalization properties
- smoothed WDRO has nice properties
  (general duality, approximation results, worst-case distribution, generalization)

On-going work

- Show that WDRO is not just a nice theory
- Further investigate applications... (in fairness?)

thank you all !