skwdro : Distributionnally Robust Optimization for Statistical Learning

Franck Iutzeler

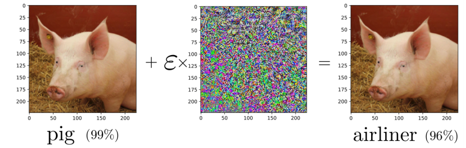
Tech'Session – Jul. 4, 2025

Decision under uncertainty

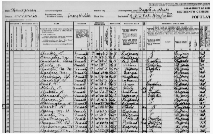
- Mathematical modelling
  - The **cost** f_x of a decision **parametrized** by $x \in \mathcal{X}$
  - depends on an **uncertain variable** $\xi \in \Xi$
- Why do we want **robustness** in practical applications?

Difficult-to-predict environments

Attacks against complex models



Biased, outdated, insufficient data



In phase with regulations

- Ben-Tal, Ghaoui, Nemirovski. *Robust optimization.* Princeton university press, 2009.
- Kolter, Madry. *Adversarial robustness - theory and practice.* NeurIPS tutorial https://adversarial-ml-tutorial.org/, 2018.

Decision under uncertainty

- Mathematical modelling
  - The **cost** f_x of a decision **parametrized** by $x \in \mathcal{X}$
  - depends on an **uncertain variable** $\xi \in \Xi$
- Why do we want **robustness** in **statistical learning**?
  - cost = model + loss f_x on data point ξ ex. least squares $f_x(\xi = (a, b)) = (\langle x, a \rangle - b)^2$
  - the uncertainty variable's **distribution** is known through **samples** $\xi_1, .., \xi_N$
  - Robustness is desirable for
    - Generalization guarantees on the true distribution of the samples
    - Distribution shifts between training and application

Popular approaches

- The *uncertain variable* ξ lives in some **uncertainty set** U

$$\min_{x \in \mathcal{X}} \sup_{\xi \in U} f_x(\xi) \qquad \text{(Worst-case robustness)}$$

  - U may be difficult to design
  - pessimistic decisions (unlikely values of ξ)

- The *uncertain variable* ξ is known though its **empirical distribution** $\hat{\mathbf{P}}_N = \frac{1}{N} \sum_{i=1}^{N} \delta_{\xi_i}$

$$\min_{x \in \mathcal{X}} \mathbb{E}_{\xi \sim \hat{\mathbf{P}}_N} [f_x(\xi)] \qquad \text{(Sample Average Approximation)}$$

  - also called Empirical Risk Minimization in machine learning
  - the empirical distribution $\hat{\mathbf{P}}_N$ may not be close to the true distribution of ξ in the target application

    too few samples, biased collection, distribution shifts

- Ben-Tal and Nemirovski. *Robust convex optimization.* Mathematics of operations research, 1998.
- Shapiro, Dentcheva, and Ruszczynski. *Lectures on stochastic programming: modeling and theory.* SIAM, 2021.

Distributionally Robust Optimization

- The empirical distribution data provides **partial information** about the encountered **distribution** of ξ
  - The uncertain variable's **distribution** lives in a **neighborhood** $\mathcal{U}(\hat{\mathbf{P}}_N)$ of its empirical distribution

$$\min_{x \in \mathcal{X}} \sup_{\substack{\mathbf{Q} \in \mathcal{P}(\Xi) \\ \mathbf{Q} \in \mathcal{U}(\hat{\mathbf{P}}_N)}} \mathbb{E}_{\xi \sim \mathbf{Q}}[f_x(\xi)] \tag{DRO}$$

  - Inner sup taken over the set $\mathcal{P}(\Xi)$ of probability measures on Ξ infinite dimensional
  - For some $\mathcal{U}(\hat{\mathbf{P}}_N)$, parametric (Gaussian) or not (ϕ-divergences), this leads to finite-dimension min-max problems efficient stochastic optimization methods
  - Enforces **model robustness at training**

- Scarf. *A min-max solution of an inventory problem.* Studies in the mathematical theory of inventory and production, 1958.
- Rahimian and Mehrotra. *Distributionally robust optimization: A review.* arXiv 1908.05659, 2019.
- Delage and Ye. *Distributionally robust optimization under moment uncertainty with application to data-driven problems.* Op. Res., 2010.
- Namkoong and Duchi. *Stochastic gradient methods for distributionally robust optimization with f-divergences.* NeurIPS, 2016.

Wasserstein Distributionally Robust Optimization

- The uncertain variable's **distribution** lives in a **Wasserstein neighborhood** of its empirical distribution

$$\min_{x \in \mathcal{X}} \sup_{\substack{\mathbf{Q} \in \mathcal{P}(\Xi) \\ W_c(\hat{\mathbf{P}}_N, \mathbf{Q}) \leq \rho}} \mathbb{E}_{\xi \sim \mathbf{Q}}[f_x(\xi)] \tag{WDRO}$$

  - For a cost function $c : \Xi \times \Xi \to \mathbb{R}_+$, the Wasserstein distance between $\hat{\mathbf{P}}_N$ and $\mathbf{Q}$ is defined as

$$W_c(\hat{\mathbf{P}}_N, \mathbf{Q}) = \inf \left\{ \mathbb{E}_{(\xi,\zeta)\sim\pi} \left[c(\xi, \zeta) \right] : \pi \in \mathcal{P}(\Xi \times \Xi), \pi_1 = \hat{\mathbf{P}}_N, \pi_2 = \mathbf{Q} \right\},$$

    with π_1 (resp. π_2) the first (resp. second) marginal of the transport plan π.
  - **Natural metric** to compare empirical and absolutely continuous distributions contrary to the Kullback-Leibler divergence and strong generalization/concentration results
  - Inner sup stays infinite dimensional and the constraint is itself linked to an optimization problem

- Esfahani and Kuhn. *Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations.* Mathematical Programming, 2018.
- Kuhn, Esfahani, Nguyen, and Shafieezadeh-Abadeh. *Wasserstein distributionally robust optimization: Theory and applications in machine learning.* In Operations Research & Management Science in the Age of Analytics, 2019.
- Blanchet and Murthy. *Quantifying distributional model risk via optimal transport.* Mathematics of Operations Research, 2019.
- Gao and Kleywegt. *Distributionally robust stochastic optimization with Wasserstein distance.* Mathematics of Operations Research, 2022.

- WDRO is an appealing framework for distributional robustness but difficult to optimize
  - Understand precisely the behavior of WDRO solutions
  - Study its statistical guarantees
  - Provide computationally tractable formulations for a large class of problems

Outline

Wasserstein Distributionally Robust Optimization

◇ **Formulation & Examples**

Dual problem

- **Duality** is at the core of modern WDRO
  - Lagrangian duality + Sup over (conditional) measure realized by a Dirac at the sup

$$\sup_{\substack{\mathbf{Q}\in\mathcal{P}(\Xi)\\ W_c(\hat{\mathbf{P}}_N,\mathbf{Q})\leq\rho}} \mathbb{E}_{\xi\sim\mathbf{Q}}[f_x(\xi)] = \inf_{\lambda\geq 0} \lambda\rho + \mathbb{E}_{\xi\sim\hat{\mathbf{P}}_N}\left[\sup_{\zeta\in\Xi}\{f_x(\zeta)-\lambda c(\xi,\zeta)\}\right] \tag{Duality}$$

- Main improvement: this is a finite-dimensional problem and λ is 1D!
  - **If** the sup is tractable, the Duality problem is solvable! and thus WDRO, but that's a big if
  - The optimal **worst-case distribution** is supported on $N+1$ atoms taken in $\arg\max_{\zeta\in\Xi}\{f_x(\zeta)-\lambda^\star c(\xi_i,\zeta)\}$ for $i=1,..,N$

- Esfahani and Kuhn. *Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations.* Mathematical Programming, 2018.
- Zhao and Guan. *Data-driven risk-averse stochastic optimization with Wasserstein metric.* Operations Research Letters, 2018.
- Blanchet and Murthy. *Quantifying distributional model risk via optimal transport.* Mathematics of Operations Research, 2019.
- Gao and Kleywegt. *Distributionally robust stochastic optimization with Wasserstein distance.* Mathematics of Operations Research, 2022.

Example I – the NewsVendor problem

- A NewsVendor has to decide how many papers he will buy for tomorrow
  - His buying price is $k = 5$ and his retail price is $u = 7$
  - He has a collection of sales data $\xi_1, .., \xi_N$
  - He wants to minimize its loss $f_x(\xi) = kx - u\min(x, \xi)$ by optimizing the number $x \in \mathbb{R}_+$ of newspaper bought, facing the uncertain demand of tomorrow $\xi \in \mathbb{R}_+$

- Taking a robust decision
  - Worst-case robustness leads to $x^\star_{WCR} = 0$ since $\xi = 0$ is possible
  - Sample Average Approximation leads to $x^\star_{SAA} > 0$ by minimizing the average loss over the past
  - What about WDRO?

Example I – the NewsVendor problem

- A NewsVendor has to decide how many papers he will buy for tomorrow
  - His buying price is $k = 5$ and his retail price is $u = 7$
  - He has a collection of sales data $\xi_1, .., \xi_N$ in $\mathbb{R}_+ = \Xi$
  - He wants to minimize its loss $f_x(\xi) = kx - u\min(x, \xi)$ by optimizing the number $x \in \mathbb{R}_+$ of newspaper bought, facing the uncertain demand of tomorrow $\xi \in \mathbb{R}_+$

$$\min_{x \geq 0} \inf_{\lambda \geq 0} \lambda\rho + \frac{1}{N}\sum_{i=1}^{N} \sup_{\zeta \in \Xi} \left\{ kx - u\min(x, \zeta) - \lambda|\xi_i - \zeta| \right\}$$

- We can solve Duality with $c(\xi, \zeta) = |\xi - \zeta|$
  - If $\lambda^\star = 0$, the sup is attained at $\zeta_i^\star = 0$ for all ξ_i, leading to $x^\star = 0 \longrightarrow \rho$ **too large, worst-case**
  - If $\lambda^\star \geq u$, the sup is attained at $\zeta_i^\star = \xi_i$ for each $\xi_i \longrightarrow$ **SAA problem** linear cost/function cancel out
  - $\lambda \in (0, u)$ cannot be optimal gradient either positive or negative
- WDRO leads to $x^\star_{WCR} = 0$ or $x^\star_{SAA}$ depending on ρ!

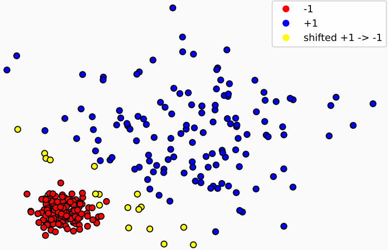
Example II – Logistic regression

- Standard classification problem
  - Labeled data $\xi_1, .., \xi_N$ of the form $\xi_i = (x_i, y_i) \in \mathbb{R}^d \times \{-1, +1\} = \Xi$
  - We minimize the loss $f_x(\xi = (x', y')) = \log(1 + \exp(-y'\langle x', x\rangle))$ by fitting separator $x \in \mathbb{R}^d$

$$\min_{x \in \mathbb{R}^d} \inf_{\lambda \geq 0} \lambda\rho + \frac{1}{N}\sum_{i=1}^{N} \sup_{\zeta=(z,v)\in\Xi} \left\{\log(1 + \exp(-y_i\langle x_i, x\rangle)) - \lambda\left(\|x_i - z\| + \kappa\mathbb{1}_{y_i \neq v}\right)\right\}$$

- We can solve Duality by disciplined convex programming
  - for this, $c(\xi = (x, y), \zeta = (z, v)) = \|x - z\| + \kappa\mathbb{1}_{y \neq v}$ if $\kappa = +\infty$, (WDRO) is ERM regularized by $\rho\|x\|_*$

$$\begin{aligned} \min_{x,\lambda,s} \quad & \lambda\rho + \frac{1}{N}\sum_{i=1}^{N} s_i \\ \text{s.t.} \quad & \log(1 + \exp(-y_i\langle x_i, x\rangle)) \leq s_i \quad \forall i \\ & \log(1 + \exp(y_i\langle x_i, x\rangle)) - \kappa\lambda \leq s_i \quad \forall i \\ & \|x\|_* \leq \lambda \end{aligned}$$



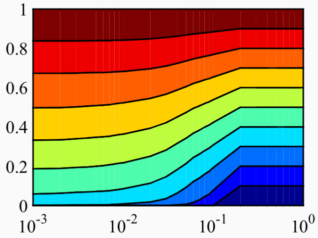
Example III – Portfolio selection

- Optimize a portfolio $x \in \{y \in \mathbb{R}^d_+ : \sum_{i=1}^d y[i] = 1\}$ over m assets subject to uncertain yearly returns
  - Return data $\xi_1, .., \xi_N$ in $\mathbb{R}^d = \Xi$
  - We minimize a risk-averse loss $f_x(\xi, \tau) = -\langle x, \xi\rangle + \eta\tau + \frac{\eta}{\alpha}\max(-\langle x, \xi\rangle - \tau; 0)$ with $\eta \geq 0$ is the risk aversion and $\alpha \in (0, 1]$ is the risk level ⇝ risk $\mathbb{E}[-\langle x, \xi\rangle] + \eta \,\mathrm{CVaR}_\alpha[-\langle x, \xi\rangle]$

$$\min_{x\in\{\mathbb{R}^d_+:\sum_{i=1}^d x[i]=1\}} \min_{\tau\in\mathbb{R}} \inf_{\lambda\geq 0} \lambda\rho + \frac{1}{N}\sum_{i=1}^N \sup_{\zeta\in\Xi}\left\{-\langle x, \zeta\rangle + \eta\tau + \frac{\eta}{\alpha}\max(-\langle x, \zeta\rangle - \tau; 0) - \lambda\|\xi_i - \zeta\|\right\}$$

- We can again solve Duality by disciplined convex programming for $c(\xi, \zeta) = \|\xi - \zeta\|$

$$\begin{aligned}\min_{x,\tau,\lambda,s} \quad & \lambda\rho + \frac{1}{N}\sum_{i=1}^N s_i \\ \text{s.t.} \quad & \eta\tau - \langle x, \xi_i\rangle \leq s_i \quad \forall i \\ & \eta(1 - 1/\alpha)\tau - (1 + \eta/\alpha)\langle x, \xi_i\rangle \leq s_i \quad \forall i \\ & \|x\|_* \leq \lambda/\eta, \sum_{i=1}^d x[i] = 1, x \geq 0\end{aligned}$$



Portfolio as a function of ρ Source: Esfahani & Kuhn, 2018

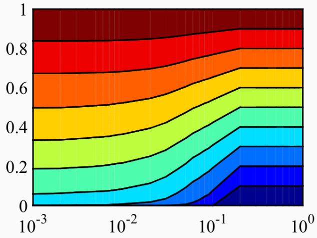
Example III – Portfolio selection

- Optimize a portfolio $x \in \{y \in \mathbb{R}^d_+ : \sum_{i=1}^d y[i] = 1\}$ over m assets subject to uncertain yearly returns
  - Return data $\xi_1, .., \xi_N$ in $\mathbb{R}^d = \Xi$
  - We minimize a risk-averse loss $f_x(\xi, \tau) = -\langle x, \xi\rangle + \eta\tau + \frac{\eta}{\alpha}\max(-\langle x, \xi\rangle - \tau; 0)$ with $\eta \geq 0$ is the risk aversion and $\alpha \in (0, 1]$ is the risk level $\rightsquigarrow$ risk $\mathbb{E}[-\langle x, \xi\rangle] + \eta\ \mathrm{CVaR}_\alpha[-\langle x, \xi\rangle]$

$$\min_{x\in\{\mathbb{R}^d_+ : \sum_{i=1}^d x[i]=1\}} \min_{\tau\in\mathbb{R}} \inf_{\lambda\geq 0} \lambda\rho + \frac{1}{N}\sum_{i=1}^N \sup_{\zeta\in\Xi}\left\{-\langle x, \zeta\rangle + \eta\tau + \frac{\eta}{\alpha}\max(-\langle x, \zeta\rangle - \tau; 0) - \lambda\|\xi_i - \zeta\|\right\}$$

- We can again solve Duality by disciplined convex programming for $c(\xi, \zeta) = \|\xi - \zeta\|$
- Recovers that optimality of equally weighted portfolio under high ambiguity

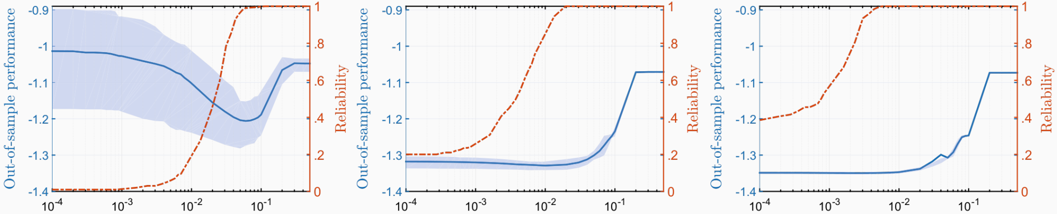
- Esfahani and Kuhn. *Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations.* Mathematical Programming, 2018.
- Pflug, Pichler, Wozabal. *The 1/N investment strategy is optimal under high model ambiguity.* J. Bank. Financ., 2012.
- Rockafellar and Uryasev. *Optimization of conditional value-at-risk.* J. Risk, 2000.



Portfolio as a function of ρ Source: Esfahani & Kuhn, 2018

- Sample 200 training datasets of size $N = \{30, 300, 3000\}$ from the same distribution
  - for each of them, solve WDRO to get optimal point $\hat{x}^\star$ and value $\widehat{\mathcal{R}}_\rho(f_{\hat{x}^\star})$
- **Reliability** = pc. of datasets s.t. the WDRO value is greater than the loss at the WDRO optimal point:

estimated by taking $N = 30000$ **target** $\mathbb{E}_{\xi\sim\mathbf{P}}[f_{\hat{x}^\star}(\xi)] \leq \widehat{\mathcal{R}}_\rho(f_{\hat{x}^\star})$ **computed**



Out-of-sample performance $\mathbb{E}_{\xi\sim\mathbf{P}}\left[f_{\hat{x}^\star}(\xi)\right]$ and reliability as a function of ρ Source: Esfahani & Kuhn, 2018

- To get a fixed reliability, no need to scale as $\frac{1}{N^{1/10}}$, $\frac{1}{\sqrt{N}}$ seems enough!

Conclusion on WDRO

- An appealing modeling framework...
  - Actually models robustness in distribution
  - Natural metric without prior
- ...with some caveats
  - The (dual) problem is only tractable for specific combinations of objectives and cost functions
  - Discrete worst cases despite encompassing all kind of distributions
  - Can suffer from a bang-bang effect between worst-cases and SAA
- WDRO models control the true risk with high probability
  - Radius ρ should be intuitively taken proportional to $1/\sqrt{N}$
  - Uniform in the model f_x

Conclusion on WDRO

- An appealing modeling framework...
  - Actually models robustness in distribution
  - Natural metric without prior
- ...with some caveats
  - **The (dual) problem is only tractable for specific combinations of objectives and cost functions**
  - Discrete worst cases despite encompassing all kind of distributions
  - Can suffer from a bang-bang effect between worst-cases and SAA
- WDRO models control the true risk with high probability
  - Radius ρ should be intuitively taken proportional to $1/\sqrt{N}$
  - Uniform in the model f_x

Wasserstein Distributionally Robust Optimization

◊ **WDRO in practical ML**

Entropic regularization

- We draw inspiration from entropic transport and regularize by entropy wrt. a reference coupling π_0
  - In optimal transport, entropic regularization with $\mathrm{KL}(\pi \,|\, \mathbf{P} \otimes \mathbf{Q})$ π_0 is the product of marginals
  - In WDRO, the second marginal is **not fixed** but optimized to get our adversarial distribution
  - We choose $\pi_0(\mathrm{d}\xi, \mathrm{d}\zeta) \propto \hat{\mathbf{P}}_N(\mathrm{d}\xi) e^{-\frac{\|\xi-\zeta\|^p}{2^{p-1}\sigma}} \mathbb{1}_{\zeta\in\Xi}\, \mathrm{d}\zeta$

$$\widehat{\mathcal{R}}_\rho(f_x) = \inf_{\lambda\geq 0} \lambda\rho + \mathbb{E}_{\xi\sim\hat{\mathbf{P}}_N}\left[\sup_{\zeta\in\Xi}\left\{f_x(\zeta) - \lambda\|\xi-\zeta\|^p\right\}\right] \tag{WDRO}$$

$$\widehat{\mathcal{R}}^\varepsilon_\rho(f_x) = \inf_{\lambda\geq 0} \lambda\rho + \varepsilon\, \mathbb{E}_{\xi\sim\hat{\mathbf{P}}_N}\left[\log\left(\mathbb{E}_{\zeta\sim\pi_0(\cdot|\xi)}\left[e^{\frac{f_x(\zeta)-\lambda\|\xi-\zeta\|^p}{\varepsilon}}\right]\right)\right] \tag{ε-WDRO}$$

Theorem (Azizian, I., Malick'22)
If $\Xi \subset \mathbb{R}^d$ is compact, convex, with nonempty interior and f_x is Lipschitz continuous, then as ε goes to 0

$$0 \leq \widehat{\mathcal{R}}_\rho(f_x) - \widehat{\mathcal{R}}^\varepsilon_\rho(f_x) \leq \mathcal{O}\left(\varepsilon d \log\left(\frac{1}{\varepsilon}\right)\right)$$

◇ Genevay, Chizat, Bach, Cuturi, and Peyré. *Sample complexity of sinkhorn divergences.* AIStats, 2019.

Solving generic WDRO problems

- Leverage the entropic regularization

$$\min_{x \in \mathcal{X}} \inf_{\lambda \geq 0} \lambda\rho + \varepsilon \frac{1}{N} \sum_{i=1}^{N} \left[\log \left(\mathbb{E}_{\zeta \sim \pi_0(\cdot|\xi_i)} \left[e^{\frac{f_x(\zeta) - \lambda \|\xi_i - \zeta\|^2}{\varepsilon}} \right] \right) \right]$$

  - Gradients in x and λ are available

$$\frac{1}{N} \sum_{i=1}^{N} \left[\frac{\mathbb{E}_{\zeta \sim \pi_0(\cdot|\xi_i)} \nabla_x f_x(\zeta) e^{\frac{f_x(\zeta) - \lambda \|\xi_i - \zeta\|^2}{\varepsilon}}}{\mathbb{E}_{\zeta \sim \pi_0(\cdot|\xi_i)} e^{\frac{f_x(\zeta) - \lambda \|\xi_i - \zeta\|^2}{\varepsilon}}} \right] \text{ and } \rho - \frac{1}{N} \sum_{i=1}^{N} \left[\frac{\mathbb{E}_{\zeta \sim \pi_0(\cdot|\xi_i)} \|\xi_i - \zeta\|^2 e^{\frac{f_x(\zeta) - \lambda \|\xi_i - \zeta\|^2}{\varepsilon}}}{\mathbb{E}_{\zeta \sim \pi_0(\cdot|\xi_i)} e^{\frac{f_x(\zeta) - \lambda \|\xi_i - \zeta\|^2}{\varepsilon}}} \right]$$

- **Crude approach:** sample some points from $\pi_0(\cdot|\xi_i) \propto e^{\frac{\|\xi_i - \zeta\|^2}{2\sigma}} \mathbb{1}_{\zeta \in \Xi}$ and minimize the sampled loss
  - This is a biased approximation with poor performance in practice except for $d = 1$
- **Better approach:** sample the expectation at each iteration by (Metropolis-adjusted) Langevin
  - "Robustifies" but unstable behavior of λ
- **Implemented approach:** additionally use importance sampling towards $\nabla_{\xi_i} f_x(\xi_i)$
  - Much more stable, when initialized with the ERM solution

- https://github.com/iutzeler/skwdro + pip/conda
- Two interfaces (see the Documentation)
    - scikit-learn models

```
from sklearn.linear_model import LinearRegression # sklearn's regressor
from skwdro.linear_models import LinearRegression as RobustLinearRegression

X,y     = ... # Training data

# === ERM ===
lin     = LinearRegression()
lin.fit(X,y)

# === DRO ===
rob_lin = RobustLinearRegression(rho=0.1) # WDRO with radius 0.1
rob_lin.fit(X,y)
```

- https://github.com/iutzeler/skwdro + pip/conda
- Two interfaces (see the Documentation)
  - wrapper over pytorch modules

```
import torch as pt
from skwdro.wrap_problem import dualize_primal_loss

model = nn.Linear(...)  # Inference model is a pytorch Module
loss_fn = pt.nn.MSELoss(reduction='none')  # quadratic loss function

wdro_loss = dualize_primal_loss(
    loss_fn, model,
    rho = pt.tensor(0.1), # Robustness radius
    X, y # Provide some "warmup" samples
) # Replaces the loss of the model by the dual WDRO loss
wdro_loss.get_initial_guess_at_dual(X, y)  # Choice of a starting lambda

optimizer = torch.optim.XXX # Optimizer of your choice
for _ in range(...): # training loop
    for X, y in train_batches:
        optimizer.zero_grad()

        # === ERM ===  Here is what you would do usualy to optimize the loss:
        # loss = loss_fn(model(X), y).mean()   # Standard loss on batch
        # loss.backward()
        # optimizer.step()  # Standard optimization step

        # === DRO === Here is the new version:
        rob_loss = wdro_loss(X, y).mean()  # Robust loss
        rob_loss.backward()
        optimizer.step()  # Robust optimization step
```