Day 2: Data-Driven Modeling and Optimization of Dynamical Systems under Uncertainty

Jia-Jie Zhu

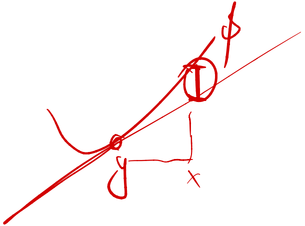
WIAS-Berlin

July 14, 2022

Mirror descent

$P \to P^*$

$\phi \quad X \to H$



We also define the Bregman divergence associated with ϕ as

$$B_\phi(x, y) = \phi(x) - \phi(y) - \nabla\phi(y)^\top(x - y).$$

Mirror descent update:

$$x_{t+1} = \underset{x}{\operatorname{argmin}} \left\{ f(x_t) + g_t{}^\top(x - x_t) + \frac{1}{\eta_k} B_\Phi(x, x_t) \right\}$$

$\min_x f(x)$

$\frac{1}{2}(x - x_t)\nabla^2 f(x - x_t)$

$g_t = \nabla f(x_t) \in \partial f(x_t)$

$1/\eta_k$ $\quad I$

$\|x_1 - x^*\|_2^2 \leq R^2$

Example of MD: Euclidean norm

Mirror map $\Phi(x) = \frac{1}{2}\|x\|_2^2$.

Example of MD: Euclidean norm

Mirror map $\Phi(x) = \frac{1}{2}\|x\|_2^2$.

MD:

$$x_{k+1} = \operatorname*{argmin}_x \left\{ f(x_k) + {g_k}^\top (x - x_k) + \frac{1}{2\eta_k}\|x - x_k\|_2^2. \right\}$$

This is equivalent to (sub-)GD with step η_k.

Example of MD: negative entropy

Mirror map $\Phi(x) = \sum_{i=1}^{d} x(i) \log x(i)$.

$x_1 \quad x_2 \quad \dots \quad x_d$

Example of MD: negative entropy

$(x(i) > 0, \forall i)$

Mirror map $\Phi(x) = \sum_{i=1}^{d} x(i) \log x(i)$. The resulting Bregman divergence is the KL-divergence $D_\Phi(x, y) = \sum_{i=1}^{d} x(i) \log \frac{x(i)}{y(i)}$.

Example of MD: negative entropy

Mirror map $\Phi(x) = \sum_{i=1}^{d} x(i) \log x(i)$. The resulting Bregman divergence is the KL-divergence $D_\Phi(x, y) = \sum_{i=1}^{d} x(i) \log \frac{x(i)}{y(i)}$.

If we restrict x to a (discrete probability) simplex $1^\top x = 1$, then the MD update

$$x_{t+1} = \operatorname*{argmin}_{1^\top x = 1} f(x_t) + g_t{}^\top (x - x_t) + \frac{1}{\eta} B_\Phi(x, x_t)$$

has the closed-form

$$x_{i,k+1} = \frac{x_{i,k} e^{-\eta_k g_k}}{\sum_{j=1}^{n} x_{j,k} e^{-\eta_k g_k}}, \quad i = 1, 2, \ldots, n$$

MD convergence

Let Φ be ρ-strongly convex in $\|\cdot\|$. Let $R^2 = \sup_x \Phi(x) - \Phi(x_1)$, and f be convex and L-Lipschitz w.r.t. $\|\cdot\|$. Then mirror descent with $\eta = \frac{R}{L}\sqrt{\frac{2\rho}{t}}$ satisfies

$$f\left(\frac{1}{t}\sum_{s=1}^{t} x_s\right) - f(x^*) \leq RL\sqrt{\frac{2}{\rho t}}.$$

MD convergence

Let Φ be ρ-strongly convex in $\|\cdot\|$. Let $R^2 = \sup_x \Phi(x) - \Phi(x_1)$, and f be convex and L-Lipschitz w.r.t. $\|\cdot\|$. Then mirror descent with $\eta = \frac{R}{L}\sqrt{\frac{2\rho}{t}}$ satisfies

$$f\left(\frac{1}{t}\sum_{s=1}^{t} x_s\right) - f(x^*) \leq RL\sqrt{\frac{2}{\rho t}}.$$

For example, for negative entropy $R \leq \sqrt{\log(d)}$.

Stochastic (sub-)gradient descent

Stochastic oracle:

$$\mathbb{E}\,\tilde{g}(x) \in \boxed{\partial f(x)}.$$

$(= \nabla f(x))$

$g \in \partial f(x)$

(ERM)

$\nabla_x \left\{ \frac{1}{n} \sum_{i=1}^{n} f(x, \xi_i) \right\}$

Then, SGD update rule:

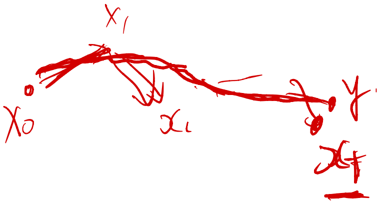
$$x_{t+1} = x_t - \eta_t \tilde{g}(x_t).$$

$\nabla_x f(x, \xi_i)$

$\xi_i \sim P_0$

$\nabla_x \mathbb{E}_{P_0} f(x, \xi)$

Stochastic (sub-)gradient descent



Stochastic oracle:

$$\mathbb{E}\, \tilde{g}(x) \in \partial f(x).$$

Then, SGD update rule:

$$x_{t+1} = x_t - \eta_t \tilde{g}(x_t).$$

Convergence. Let f be α-strongly convex, and assume that the stochastic oracle is such that $\mathbb{E}\|\tilde{g}(x)\|^2 \leq B^2$. Then SGD with $\eta_k = \frac{2}{\alpha(k+1)}$ satisfies

$$\mathbb{E} f\left(\sum_{k=1}^{t} \frac{2k}{t(t+1)} x_k\right) - f(x^*) \leq \frac{2B^2}{\alpha(t+1)}.$$

$f(x_t)$

(Skip) stochastic mirror descent (S-MD)

Let $x_1 \in \mathrm{argmin}_{\mathcal{X}} \Phi(x)$, and

$$x_{t+1} = \underset{x}{\mathrm{argmin}} \;\; f(x_t) + \tilde{g}(x_t)^\top (x - x_t) + \underline{\frac{1}{\eta} B_\Phi(x, x_t)}.$$

Convergence. Let Φ be a mirror map 1-strongly convex. Let $R^2 = \sup_x \Phi(x) - \Phi(x_1)$. Let f be convex and β-smooth w.r.t. $\|\cdot\|$. Assume that the stochastic oracle is such that $\mathbb{E}\|\nabla f(x) - \tilde{g}(x)\|_*^2 \leq \sigma^2$. Then S-MD with stepsize $\frac{1}{\beta + 1/\eta}$ and $\eta = \frac{R}{\sigma}\sqrt{\frac{2}{t}}$ satisfies

$$\mathbb{E} f\left(\frac{1}{t}\sum_{s=1}^{t} x_{s+1}\right) - f(x^*) \leq R\sigma\sqrt{\frac{2}{t}} + \frac{\beta R^2}{t}.$$

Compare the basic gradient descent with SGD for average loss functions

$f(x, \xi_i)$ $\{\xi_i\}$ $\{x_i, y_i\}$

$$\boxed{f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x).}$$ ER

$O(e^{-t})$

GD: (k)

$$x_{t+1} = x_t - \frac{\eta}{n}\sum_{i=1}^{n} \nabla f_i(x),$$

(Vanilla SGD)

$$x_{t+1} = x_t - \eta\nabla f_{i_t}(x),$$

$f(x, \xi_{i_t})$ Ep

$O(t^{-a})$

where i_t is drawn uniformly random.

$\xi_i \sim P_z$ N ξ_i

$x_t = (XX^T)^{-1}X^T Y$

Optimization under distributional uncertainty

Empirical risk minimization

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} l(f_\theta, \xi_i), \quad \xi_i \sim P_0$$

$$\inf_{f\in H} E_{P_0} l(f, \xi)$$

- Robust under statistical fluctuation, e.g., we can bound $\mathbb{E}_{P_0} l(f_{\hat{\theta}}, \xi)$

Optimization under distributional uncertainty

Empirical risk minimization

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} l(f_{\theta}, \xi_i), \quad \xi_i \sim P_0$$

- Robust under statistical fluctuation, e.g., we can bound $\mathbb{E}_{P_0} l(f_{\hat{\theta}}, \xi)$
- Not robust under data distribution shifts, when Q $(\neq P_0)$

Optimization under distributional uncertainty

Empirical risk minimization

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} l(f_\theta, \xi_i), \quad \xi_i \sim P_0$$

- Robust under statistical fluctuation, e.g., we can bound $\mathbb{E}_{P_0} l(f_{\hat{\theta}}, \xi)$
- Not robust under data distribution shifts, when $Q \; (\neq P_0)$

Distributionally robust optimization and learning

$$\min_{f \in \mathcal{H}} \overset{\text{(max)}}{\sup_{Q \in \mathcal{M}}} \mathbb{E}_Q \, l(f, \xi)$$

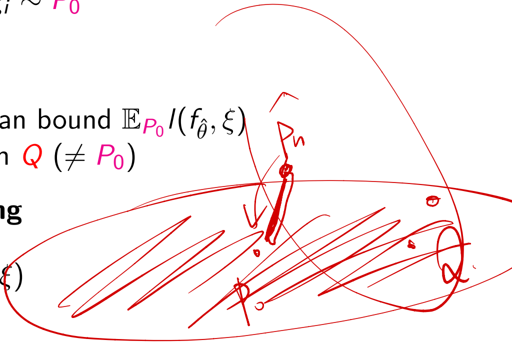
- Minimize risk under a local worst-case distribution Q

Optimization under distributional uncertainty

Empirical risk minimization

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} l(f_\theta, \xi_i), \quad \xi_i \sim P_0$$

- Robust under statistical fluctuation, e.g., we can bound $\mathbb{E}_{P_0} l(f_{\hat{\theta}}, \xi)$
- Not robust under data distribution shifts, when Q $(\neq P_0)$



Distributionally robust optimization and learning

$$\min_{f \in \mathcal{H}} \sup_{Q \in \mathcal{M}} \mathbb{E}_Q L(f, \xi)$$

- Minimize risk under a local worst-case distribution Q
- Distribution shift described by an ambiguity set $\mathcal{M}$

Optimization under distributional uncertainty

Empirical risk minimization

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} l(f_\theta, \xi_i), \quad \xi_i \sim P_0$$

- Robust under statistical fluctuation, e.g., we can bound $\mathbb{E}_{P_0} l(f_{\hat{\theta}}, \xi)$
- Not robust under data distribution shifts, when Q $(\neq P_0)$

Distributionally robust optimization and learning

$$\min_{f \in \mathcal{H}} \sup_{Q \in \mathcal{M}} \mathbb{E}_Q L(f, \xi)$$

- Minimize risk under a local worst-case distribution Q
- Distribution shift described by an ambiguity set $\mathcal{M}$
- We can bound performance beyond statistical fluctuation (classical learning theory)

DRO with divergence or metric ball constraints

$\mathcal{P}$

$\mathcal{M}^+$

Let $\mathcal{D}$ denote a divergence measure or metric on the $\mathcal{P}$, we consider the *data-driven DRO problem*

$$\boxed{\min_{\theta} \max_{\mathcal{D}(P,\hat{P})\leq\epsilon} \mathbb{E}_P l(\theta,\xi)}$$

$\hat{P} = \frac{1}{n}\sum_{i=1}^{n} \delta_{\xi_i}$

$\cdot \hat{P}$

Φ-divergence (skip in class)

Tsybakov. – .

Wasserstein distance

$\in P.$ $\quad E_P|X|^p < \infty$

p-Wasserstein distance between probability measures $\cancel{\mu_0}$ P, $\cancel{\mu_1}$ Q on $\mathbb{R}^d$ (with p finite moments) is defined through the following Kantorovich problem

$$W_p(P, Q)^p := \min \left\{ \int_{\mathbb{R}^d \times \mathbb{R}^d} |x_0 - x_1|^p \, d\Pi \,\middle|\, \pi^{(1)}_{\#}\Pi = P,\ \pi^{(2)}_{\#}\Pi = Q \right\}.$$

Wasserstein distance

p-Wasserstein distance between probability measures μ_0, μ_1 on $\mathbb{R}^d$ (with p finite moments) is defined through the following Kantorovich problem

$$W_p(P,Q)^p := \min\left\{ \int_{\mathbb{R}^d\times\mathbb{R}^d} |x_0 - x_1|^p \,\mathrm{d}\Pi \,\middle|\, \pi^{(1)}_{\#}\Pi = P,\ \pi^{(2)}_{\#}\Pi = Q \right\}.$$

$\|$

∞-dim duality

The dual Kantorovich problem

$$\max\left\{ \int_X \phi \,\mathrm{d}P + \int_X \psi \,\mathrm{d}Q \ :\ \phi,\psi \in C(X),\ \underbrace{\phi(x)+\psi(y) \le \overset{|x-y|^p}{c(x,y)}}_{},\ \forall x,y \right\}.$$

Wasserstein distance

p-Wasserstein distance between probability measures μ_0, μ_1 on $\mathbb{R}^d$ (with p finite moments) is defined through the following Kantorovich problem

$$W_p(P,Q)^p := \min \left\{ \int_{\mathbb{R}^d \times \mathbb{R}^d} |x_0 - x_1|^p \, \mathrm{d}\Pi \,\middle|\, \pi^{(1)}_{\#}\Pi = P,\ \pi^{(2)}_{\#}\Pi = Q \right\}.$$

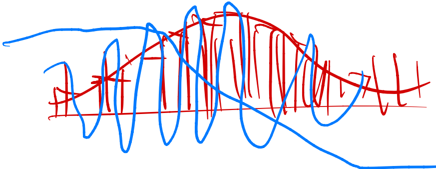
The dual Kantorovich problem

$$\max \left\{ \int_X \phi \, \mathrm{d}P + \int_X \psi \, \mathrm{d}Q \ :\ \phi, \psi \in C(X),\ \phi(x) + \psi(y) \le c(x,y),\ \forall x, y \right\}.$$

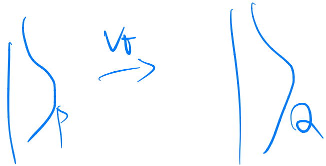
Convergence depends on dimensions, e.g., (See also [Weed and Bach 2017])

$$W_1(\hat{P}, P_0) \ge \mathcal{O}(n^{-\frac{1}{d}}).$$

$\xi_i \sim P_0$



(Skip) The dynamic formulation of the Wasserstein distance



$$W_2(P,Q)^2 = \min\left\{ \int_0^1 \int_{\mathbb{R}^d} |v_t|^2 \,\mathrm{d}\mu_t \,\mathrm{d}t \,\middle|\, \mu_0 = P, \mu_1 = Q, \frac{\mathrm{d}}{\mathrm{d}t}\mu_t + \mathrm{div}\,(v_t\mu_t) = 0 \right\}$$

$c-$transform

Given a function $f : X \to \overline{\mathbb{R}}$ we define its $\{c-transform\}$ *(or $c-$conjugate function)* by

$$f^c(y) = \underbrace{\inf_{x\in X} c(x,y) - f(x)}_{\langle x, y\rangle}.$$

- *Moreover, we say that a function ψ is $c-$concave if there exists ϕ such that $\psi = \phi^c$ and we denote by $\Psi_c(X)$ the set of $c-$concave functions.*

$c-$transform

Given a function $f : X \to \overline{\mathbb{R}}$ we define its $\{c-transform\}$ *(or $c-$conjugate function) by*

$$f^c(y) = \inf_{x \in X} c(x,y) - f(x).$$

- *Moreover, we say that a function ψ is $c-$concave if there exists ϕ such that $\psi = \phi^c$ and we denote by $\Psi_c(X)$ the set of $c-$concave functions.*
- *Using this transform, we can write down the so-called semi-dual formulation*

$$\boxed{\max \left\{ \int_X \phi \, \mathrm{d}\mu + \int_X \phi^c \, \mathrm{d}\nu \ : \ \phi \in \Psi_c(X) \right\}.}$$

$c-$transform

Given a function $f : X \to \overline{\mathbb{R}}$ we define its $\{c-transform\}$ *(or c−conjugate function) by*

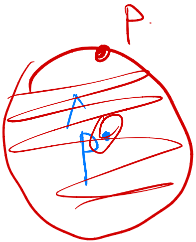
$$f^c(y) = \inf_{x \in X} c(x,y) - f(x).$$

- *Moreover, we say that a function ψ is c−concave if there exists ϕ such that $\psi = \phi^c$ and we denote by $\Psi_c(X)$ the set of c−concave functions.*
- *Using this transform, we can write down the so-called semi-dual formulation*

$$\max \left\{ \int_X \phi \, \mathrm{d}\mu + \int_X \phi^c \, \mathrm{d}\nu \ : \ \phi \in \Psi_c(X) \right\}.$$

- *Exercise. Can you see the relation with the α-strong convexity and β-smoothness we talked about?*

Dual reformulation for 2-Wasserstein DRO



The primal DRO problem is intractable

$$\min_{\theta} \sup_{W_2(P,\hat{P})\le\epsilon} \mathbb{E}_P l(\theta,\xi).$$

$\Phi(P,Q) \ge 0$

when $P \neq Q$

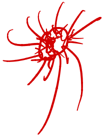
$D(\cdot,\cdot) := 0$

inf. C

$C \in \mathbb{R}$.

s.t. $E_P l(\theta,\xi) \le C, \forall P \in B_{W_2}(\hat{P})$

$\max_{\xi} l(\theta, \boxed{\xi})$



Moment Constr. $E_P x^p = E_Q x^p$

Dual reformulation for 2-Wasserstein DRO

The primal DRO problem is intractable

$\hat{P}_N := \frac{1}{N}\sum_{i=1}^{N} \delta_{\xi_i}$

$$\min_{\theta} \sup_{W_2(P,\hat{P})\leq\epsilon} \mathbb{E}_P l(\theta,\xi).$$

Fortunately, it has a strong dual as follows, i.e., the two problems have the same optimal value

$\mathbb{E}_{\hat{P}}$

$$\min_{\theta,\lambda\geq 0} \frac{1}{N}\sum_{i=1}^{N} (l_\theta)^{\lambda\|\cdot\|^2}(\xi_i) + \lambda\epsilon^2$$

c-transform of $l(f_\theta, \cdot)$

$\sup_{\xi} \{ l(f_\theta, \xi) - c(\xi, \xi_i) \}$

$(P) \leq g(\lambda)$

$\lambda \quad \|\|_2^2$

Dual reformulation for 2-Wasserstein DRO

The primal DRO problem is intractable

$$\min_{\theta} \sup_{W_2(P,\hat{P})\leq\epsilon} \mathbb{E}_P l(\theta, \xi).$$

Fortunately, it has a strong dual as follows, i.e., the two problems have the same optimal value

$$\min_{\theta,\lambda>0} \frac{1}{N}\sum_{i=1}^{N}(l_\theta)^{\lambda\|\cdot\|^2}(\xi_i) + \lambda\epsilon^2$$

This can be shown to motivate a stochastic gradient algorithm for DRO.

PA

① sup.
② inf

Kernel maximum mean discrepancy (MMD)

$$\mathrm{MMD}(P, Q) := \| \int k(x, \cdot) dP_{(x)} - \int k(x, \cdot) dQ_{(x)} \|_{\mathcal{H}}.$$

$$\int k(x, \cdot)\, d\, P(x) - Q(x)$$

Kernel maximum mean discrepancy (MMD)

$$\mathrm{MMD}(P, Q) := \left\| \int k(x,\cdot)dP - \int k(x,\cdot)dQ \right\|_{\mathcal{H}}.$$

Given two samples from the distribution of interest
$x_i \sim P, i = 1 \dots M; y_j \sim Q, j = 1 \dots N,$

$$\mathrm{MMD}(P, Q)^2 = \mathbb{E}_{x,x' \sim P} k(x, x') + \mathbb{E}_{y,y' \sim Q} k(y, y') - 2\mathbb{E}_{x \sim P, y \sim Q} k(x, y)$$

$$\approx \frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} k(x_i, x_j') + \frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} k(y_i, y_j') - 2 \frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} k(x_i, y_j)$$

This is particularly handy in, e.g., training deep generative models.

bedroom $\hat{P} = \frac{1}{n} \sum \delta_{x_i}$

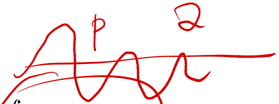
MMD-GAN

W-GAN X

$\min_Q (P, Q)$

$\frac{1}{m} \sum_{i=1}^{m} G_\theta(z_i)$

Kernel maximum mean discrepancy (MMD)



$$\mathrm{MMD}(P, Q) := \left\| \int k(x, \cdot)dP - \int k(x, \cdot)dQ \right\|_{\mathcal{H}}.$$

$\cdot\ \varepsilon$

Given two samples from the distribution of interest
$x_i \sim P, i = 1 \ldots M; y_j \sim Q, j = 1 \ldots N,$

- $k(x,x)$
- $K^T\alpha = Y$

$$\mathrm{MMD}(P, Q)^2 = \mathbb{E}_{x,x' \sim P} k(x, x') + \mathbb{E}_{y,y' \sim Q} k(y, y') - 2\mathbb{E}_{x \sim P, y \sim Q} k(x, y)$$

$$\approx \frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} k(x_i, x_j') + \frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} k(y_i, y_j') - 2\frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} k(x_i, y_j)$$

This is particularly handy in, e.g., training deep generative models.

Just like W_p, the MMD has a dual formulation

$\cancel{\int f\, dP + \int f\, dQ}$

$$\mathrm{MMD}(P, Q) = \sup_{\|f\|_{\mathcal{H}} \le 1} \int f \, d(P - Q).$$

Kernel maximum mean discrepancy (MMD)

$$\mathrm{MMD}(P,Q) := \| \int \underline{k}(x,\cdot)dP - \int \underline{k}(x,\cdot)dQ \|_{\mathcal{H}}.$$

Given two samples from the distribution of interest
$x_i \sim P, i = 1 \ldots M; y_j \sim Q, j = 1 \ldots N,$

$$\begin{aligned}\mathrm{MMD}(P,Q)^2 &= \mathbb{E}_{x,x' \sim P} k(x,x') + \mathbb{E}_{y,y' \sim Q} k(y,y') - 2\mathbb{E}_{x \sim P, y \sim Q} k(x,y) \\ &\approx \frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} k(x_i, x_j') + \frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} k(y_i, y_j') - 2\frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} k(x_i, y_j)\end{aligned}$$

This is particularly handy in, e.g., training deep generative models.

Just like W_p, the MMD has a dual formulation

$$\mathrm{MMD}(P,Q) = \sup_{\|f\|_{\mathcal{H}} \leq 1} \int f \, d(P - Q).$$

Convergence

$$\mathrm{MMD}(\hat{P}, P_0) \leq \mathcal{O}(n^{-\frac{1}{2}}).$$

$\frac{1}{d}$

Dual reformulation for Kernel DRO

$\ell \notin \mathcal{H}$

The DRO problem with MMD constraint

$$\boxed{\min_{\theta} \sup_{\mathrm{MMD}(P,\hat{P})\le\epsilon} \mathbb{E}_P\, l(\theta,\xi)}$$

$\phi(x) + \psi(y) \le c(x,y)$

can be reformulated using a Kantorovich-type duality as

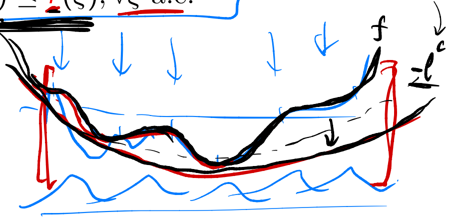
cvx

$\sum \alpha_i k(x_i, \cdot)$

cvx $\quad$ cvx $\quad + R$

"Moreau"

$$\min_{\theta, f\in\mathcal{H}} \frac{1}{N}\sum_{i=1}^{N} f(\xi_i) + \epsilon \|f\|_{\mathcal{H}} \quad \boxed{\text{s.t. } l(\theta,\xi) \le f(\xi), \forall \xi \text{ a.e.}}$$



$f \in \mathcal{H}$

~~$f(x) \in F$~~

Dual reformulation for Kernel DRO

The DRO problem with MMD constraint

$$\min_{\theta} \sup_{\mathrm{MMD}(P,\hat{P})\le\epsilon} \mathbb{E}_P l(\theta,\xi)$$

can be reformulated using a Kantorovich-type duality as

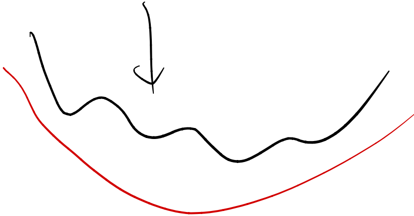
$$\min_{\theta, f\in\mathcal{H}} \frac{1}{N}\sum_{i=1}^{N} f(\xi_i) + \epsilon\|f\|_{\mathcal{H}} \quad \text{s.t. } \underline{l(\theta,\xi) \le f(\xi), \forall\xi} \text{ a.e.}$$

Optionally, we may consider to solve the problem with a relaxed (albeit with statistical guarantee) constraint

$\cancel{y_i = f(\xi_i)}$

$$l(\theta,\underline{\xi_i}) \,\underline{\le}\, f(\underline{\xi_i}),\ \ i = 1,\ldots N.$$

Dual reformulation for Kernel DRO



The DRO problem with MMD constraint

$$\min_{\theta} \sup_{\mathrm{MMD}(P,\hat{P})\leq\epsilon} \mathbb{E}_P l(\theta,\xi)$$

can be reformulated using a Kantorovich-type duality as

$$\min_{\theta, f\in\mathcal{H}} \frac{1}{N}\sum_{i=1}^{N} f(\xi_i) + \epsilon\|f\|_{\mathcal{H}} \quad \text{s.t. } l(\theta,\xi) \leq f(\xi), \forall\xi \text{ a.e.}$$

Optionally, we may consider to solve the problem with a relaxed (albeit with statistical guarantee) constraint

$$l(\theta,\xi_i) \cancel{\leq f} (\xi_i), \quad i = 1,\ldots N.$$

The solution has the nice property of being

- a kernel interpolant of the loss $l(\theta,\xi)$ at the data points ξ_i's.

Dual reformulation for Kernel DRO

The DRO problem with MMD constraint $\hat{P}_N$

$\xi \sim P$

$$\min_{\theta} \sup_{\mathrm{MMD}(P,\hat{P})\leq\epsilon} \mathbb{E}_P l(\theta, \xi)$$

$\int f \, dP - \hat{P}$

can be reformulated using a Kantorovich-type duality as

$$\min_{\theta, f\in\mathcal{H}} \frac{1}{N}\sum_{i=1}^{N} f(\xi_i) + \epsilon\|f\|_{\mathcal{H}} \quad \text{s.t. } l(\theta,\xi) \leq f(\xi), \forall\xi \text{ a.e.}$$

Optionally, we may consider to solve the problem with a relaxed (albeit with statistical guarantee) constraint

$$l(\theta, \xi_i) \leq f(\xi_i), \; i = 1, \ldots N.$$

The solution has the nice property of being

- a kernel interpolant of the loss $l(\theta, \xi)$ at the data points ξ_i's.
- a witness (optimal test) function between $\hat{P}$ and the underlying worst-case distribution P.

Going beyond DRO

Going beyond DRO

- Big gap between theory and practice in large-scale learning models

Going beyond DRO

- Big gap between theory and practice in large-scale learning models
- Continuous optimization and gradient flows of nonlinear functionals of measures and distributions

$$\min_{\mu \in \mathcal{M}} F(\mu).$$

$$E_\mu \ell$$

End!