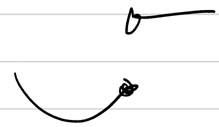
Coherent risk measures

Convex Analysis

$f: \mathbb{R}^d \to \mathbb{R}$

Lower semi-continuity (lsc) $\quad \liminf_{x' \to x} f(x') \geq f(x)$



FACT Any lsc convex function is the supremum of all affine functions minorizing f

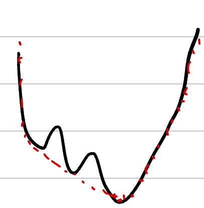
i.e. $f(x) = \sup_{a \leq f} a(x)$ where sup is over all affine a.

cf. Proof of this uses the separating hyperplane theorem. Argues any point outside the epigraph has an affine minorant between it and the epigraph.

The Fenchel conjugate of f is $f^*(s) := \sup_t \{\langle s,t\rangle - f(t)\}$.

The biconjugate is simply the conjugate of f^*, which we denote by f^{**}.

The biconjugate f^{**} is the largest lsc convex function minorizing f.



Lemma $\quad f^{**}(x) = \sup_{a \in \mathbb{R}^d, b \in \mathbb{R}} \{\langle a,x\rangle - b : \langle a,t\rangle - b \leq f(t) \ \forall t\}$

Pf)

Let $A \subset \mathbb{R}^d \times \mathbb{R}$ be set of all (a,b) minorizing f: $f(t) \geq \langle a,t\rangle - b \quad \forall t$.

Then, $(a,b) \in A \iff f(t) \geq \langle a,t\rangle - b \ \forall t \iff b \geq \langle a,t\rangle - f(t) \ \forall t \iff b \geq f^*(a),\ a \in \operatorname{dom} f^*$

and RHS $= \sup\{\langle a,x\rangle - b : a \in \operatorname{dom} f^*, -b \leq -f^*(a)\} = \sup\{\langle a,x\rangle - f^*(a)\}$. $\boxtimes$

In particular, if f is convex, then $f^{**} = f$. ✓ Fenchel–Moreau

Coherence

Consider a risk measure that maps random losses to a risk value $\mathcal{R}: L^{k_0}(P) \to \mathbb{R}$. The simplest such measure is $\mathcal{R}(W) = \mathbb{E}_P W$.

coherence defines a class of "sensible" disutility functions.

Def $\mathcal{R}$ is a coherent risk measure if

1) (Convexity) $\quad \mathcal{R}(\lambda W + (1-\lambda)W') \leq \lambda \mathcal{R}(W) + (1-\lambda)\mathcal{R}(W') \quad \forall \lambda \in [0,1]$

Diminishing marginal returns ↓ or caring more about high losses.

2) (Monotonicity) $\quad W \leq W'$ P-a.s. $\Rightarrow \mathcal{R}(W) \leq \mathcal{R}(W')$

3) (Translation equivariance) $\quad \mathcal{R}(W+c) = \mathcal{R}(W) + c \quad \forall c \in \mathbb{R}$

constant amt of deterioration in loss → same for risk

4) (Positive homogeneity) $\quad \mathcal{R}(\lambda W) = \lambda \mathcal{R}(W) \quad \forall \lambda > 0$.

Example (semideviations) $\quad \mathcal{R}(W) = \mathbb{E}_P W + \left(\mathbb{E}_P (W - \mathbb{E}W)_+^{k_0}\right)^{\frac{1}{k_0}}$ is coherent.

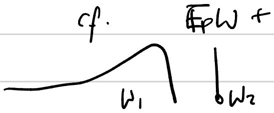
1, 3, 4 are obvious. To see 2,

$\mathcal{R}(W) \leq \mathbb{E}_P W + \left(\mathbb{E}_P (W' - \mathbb{E}W)_+^{k_0}\right)^{\frac{1}{k_0}} = \mathbb{E}_P W + \left(\mathbb{E}_P (W' - \mathbb{E}W' + \mathbb{E}W' - \mathbb{E}W)_+^{k_0}\right)^{\frac{1}{k_0}}$

$= \mathbb{E}_P W + 2\left(\mathbb{E}_P \left(\tfrac{1}{2}(W' - \mathbb{E}W') + \tfrac{1}{2}(\mathbb{E}W' - \mathbb{E}W)\right)_+^{k_0}\right)^{\frac{1}{k_0}} \leq \mathbb{E}_P W + \left(\mathbb{E}_P (W' - \mathbb{E}W')_+^{k_0}\right)^{\frac{1}{k_0}} + \mathbb{E}W' - \mathbb{E}W = \mathcal{R}(W')$.

cf. $\mathbb{E}_P W + \left(\mathbb{E}_P (W - \mathbb{E}W)^2\right)^{\frac{1}{2}}$ is NOT coherent since it also penalizes downward deviations of W.

then $\mathcal{R}(W_1) \geq \mathcal{R}(W_2)$ may hold.



Rmk Wass DRO is not coherent. Why? (no monotonicity)

Example $\mathcal{R}(W) = \inf_\eta \{ c\, (\mathbb{E}(W-\eta)_+^{k_*})^{\frac{1}{k_*}} + \eta \}$ is coherent.

In fact, ANY DRO problem over convex $\mathcal{Q} \subseteq \{Q : Q \ll P\}$, $\mathcal{R}(W) = \sup_{Q \in \mathcal{Q}} \mathbb{E}_Q W$, defines a coherent risk measure. This follows by the likelihood ratio representation

$$\mathcal{R}(W) = \sup \{ \mathbb{E}_P LW : L = \tfrac{dQ}{dP},\ Q \in \mathcal{Q} \}$$

We can show that the converse is also true. Recall the conjugate function

$$\mathcal{R}^*(L) := \sup_{W \in \mathcal{L}^{k_*}(P)} \{ \mathbb{E}_P LW - \mathcal{R}(W) \}.$$

<u>Theorem</u> For any coherent risk measure, $\exists\, \mathcal{L} \subset \mathcal{L}^k(P)$ s.t. $\mathcal{R}(W) = \sup_{L \in \mathcal{L}} \mathbb{E}_P LW$.

<u>Pf</u> First note that since any finite-valued convex function is continuous, $\mathcal{R}^{**} = \mathcal{R}$ by Fenchel-Moreau.

So $\mathcal{R}(W) = \sup_{L \in \mathcal{L}^k(P)} \{ \mathbb{E}_P LW - \mathcal{R}^*(L) \} = \sup_{L \in \mathrm{dom}(\mathcal{R}^*)} \{ \mathbb{E}_P LW - \mathcal{R}^*(L) \} \quad \cdots (*)$

We know proceed in three parts.

① Property 2 $\iff \forall L \in \mathrm{dom}(\mathcal{R}^*)$, $L \geq 0$ P-a.s..

<u>Pf</u> '$\Rightarrow$' Assume $\exists L \in \mathrm{dom}(\mathcal{R}^*)$ st. $L < 0$ on some set S of positive measure. Fix any $W \in \mathcal{L}^{k_*}(P)$.

Then, $\mathcal{R}^*(L) \geq \sup_{\lambda \geq 0} \{ \mathbb{E}_P L(W - \lambda \mathbb{1}_S) - \mathcal{R}(W - \lambda \mathbb{1}_S) \} \geq \sup_{\lambda \geq 0} \{ \mathbb{E}_P L(W - \lambda \mathbb{1}_S) - \mathcal{R}(W) \} = \infty$. ✗

since $\mathcal{R}(W) \leq \mathcal{R}(W - \lambda \mathbb{1}_S)$ by Property 2.

'$\Leftarrow$' For any $W \leq W'$ P-a.s., $\mathbb{E}_P LW \leq \mathbb{E}_P LW'$ so $(*)$ gives the result.

② Property 3 $\iff \forall L \in \mathrm{dom}(\mathcal{R}^*)$, $\mathbb{E}_P L = 1$

<u>Pf</u> '$\Rightarrow$' $\mathcal{R}^*(L) \geq \sup_{c \in \mathbb{R}} \{ \mathbb{E}_P L(W+c) - \mathcal{R}(W+c) \} = \sup_c \{ c(\mathbb{E}_P L - 1) + \mathbb{E}_P LW - \mathcal{R}(W) \} = \infty$ if $\mathbb{E}_P L \neq 1$.

'$\Leftarrow$' Follows from $(*)$.

③ Property 4 $\iff \mathcal{R}(W) = \sup_{L \in \mathrm{dom}\, \mathcal{R}^*} \mathbb{E}_P LW$.

<u>Pf</u> '$\Rightarrow$' $\mathcal{R}^*(L) = \sup_{W \in \mathcal{L}^{k_*}(P)} \{ \mathbb{E}_P LW - \mathcal{R}(W) \} = \sup_{\lambda > 0,\, W \in \mathcal{L}^{k_*}(P)} \{ \mathbb{E}_P L(\lambda W) - \mathcal{R}(\lambda W) \}$

$= \sup_{\lambda > 0} \lambda \mathcal{R}^*(L)$. So either $\mathcal{R}^*(L) = 0$ or ∞.

'$\Leftarrow$' trivial. $\blacksquare$

So DRO $\Leftrightarrow$ risk-aversion (coherence). i.e., good perf. under distributional shifts
$\Leftrightarrow$ good tail-performance.

Qs
- Which $\mathcal{R}$ to use? Or equivalently, which $\mathcal{Q}$? $\begin{cases} ① \text{ Desired notion of risk-aversion} \\ ② \text{ statistical efficiency} \\ ③ \text{ Computational efficiency} \end{cases}$
- Risk-aversion generally means less sample. efficiency
- Open Qs:
  - Linking problem structure to a particular kind of distr shift
    - ↳ This gives rough guidelines on how to choose $\mathcal{Q}$.
  - So far, we considered shifts in full distr of Z.

What if we are interested in partial distr shifts, only w.r.t. marginal distr of $P_{X_{pro}}$.

- Outliers
- Statistics in high-dim
- Training deep nets with risk-aversion

Transition So far, we studied DRO as a population formulation.
In practice, we would formulate a empirical approximation to this problem and solve

$$\underset{\theta\in\Theta}{\text{minimize}} \quad \sup_{Q:\, D(Q,\hat{P}_n)\le\rho} \mathbb{E}_Q \ell(\theta; Z),$$

where $\hat{P}_n$ is the empirical distribution, $\frac{1}{n}$ uniform weights on each data point.
For f-divergence DRO, duality gives

$$\underset{\theta\in\Theta,\, \lambda\ge 0,\, \eta\in\mathbb{R}}{\text{minimize}} \frac{1}{n}\sum \left\{ \lambda \cdot f^*\left(\frac{\ell(\theta; Z_i) - \eta}{\lambda} \right) + \lambda\rho + \eta \right\}.$$

It turns out that by setting the radius $\rho_n \to 0$ as $n\to\infty$ above, we can view these finite sample procedures as approximations to the average-case optimization problem

i.e. $\min_{\theta\in\Theta} \sup_{q:\, \frac{1}{n}\sum f(q_i / \frac{1}{n}) \le \rho_n} \sum_{i=1}^n q_i \ell(\theta; Z_i)$ is a good approximation to $\mathbb{E}_P \ell(\theta; Z)$

if we choose ρ_n appropriately.

This is what we will show now.

Subexponential RVs & Bernstein bounds

So far, we studied tail-bounds for subGaussian RVs X satisfying $\mathbb{E}e^{\lambda(X-\mathbb{E}X)} \le e^{\frac{\lambda^2\sigma^2}{2}}$ $\forall \lambda \in \mathbb{R}$.
This puts a restrictive condition on the tails of X, so it's natural to consider relaxations.

Def A RV X is sub-exponential with parameters (ν, α) nonnegative if
$\mathbb{E}e^{\lambda(X-\mathbb{E}X)} \le e^{\frac{\lambda^2\nu^2}{2}}$ for all λ s.t. $|\lambda| < \frac{1}{\alpha}$.

Any subGaussian RV is sub-exponential, but the converse is not true.
Ex Let $X \sim N(0,1)$. Then, $\mathbb{E}e^{\lambda(X^2-1)} = \frac{1}{\sqrt{1-2\lambda}} e^{-\lambda} \le e^{2\lambda^2}$ $\forall |\lambda| \le \frac{1}{4}$

By the Chernoff bound, we can again derive a tail-bound for sub-exponential RVs
$\mathbb{P}(X-\mathbb{E}X \ge t) \le e^{-\lambda t}\mathbb{E}e^{\lambda(X-\mathbb{E}X)} \le \exp\left(-\lambda t + \frac{\lambda^2\nu^2}{2}\right)$ if $0 < \lambda < \frac{1}{\alpha}$.

Unconstrained min of $-\lambda t + \frac{\lambda^2 \nu^2}{2}$: $\lambda^* = \frac{t}{\nu^2}$, this gives $RHS = -\frac{t^2}{2\nu^2}$ if $\frac{t}{\nu^2} < \frac{1}{\alpha}$.
If $t/\nu^2 \ge \frac{1}{\alpha}$, $\lambda \mapsto -\lambda t + \frac{\lambda^2\nu^2}{2}$ is monotone, so $\lambda^* = \frac{1}{\alpha}$ gives $RHS = -\frac{t}{\alpha} + \frac{1}{2\alpha}\frac{\nu^2}{\alpha} \le -\frac{t}{\alpha} + \frac{1}{2\alpha}t = -\frac{t}{2\alpha}$.
Combining, $\mathbb{P}(X-\mathbb{E}X \ge t) \le \begin{cases} \exp(-t^2/2\nu^2) & \text{if } 0 \le t \le \frac{\nu^2}{\alpha} \\ \exp(-t/2\alpha) & \text{if } t > \nu^2/\alpha \end{cases}$
(Similarly, left tail can be controlled)

(Bernstein's condition) Let $\sigma^2 = \mathrm{Var}\,X$. $\forall k \ge 2$, $|\mathbb{E}(X-\mathbb{E}X)^k| \le \frac{1}{2}k!\sigma^2 b^{k-2}$
e.g. If $|X-\mu| \le b$, B's condition holds.

If X satisfies B's condition, it is sub-exponential.
$\mathbb{E}e^{\lambda(X-\mathbb{E}X)} = \mathbb{E}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}(X-\mu)^k = 1 + \frac{\lambda^2\sigma^2}{2} + \sum_{k=3}^{\infty}\lambda^k\frac{1}{k!}\mathbb{E}(X-\mathbb{E}X)^k \le 1 + \frac{\lambda^2\sigma^2}{2} + \sum_{k=3}^{\infty}\frac{\lambda^2\sigma^2}{2}\cdot(|\lambda|\cdot b)^{k-2}$
For any $|\lambda| < 1/b$, $RHS = 1 + \frac{\lambda^2\sigma^2}{2}\cdot\frac{1}{1-b|\lambda|} \le \exp\left(\frac{\lambda^2\sigma^2}{2}\cdot\frac{1}{1-b|\lambda|}\right)$ by $1+t \le e^t$. $\cdots (*)$
So for $|\lambda| < 1/2b$, $\mathbb{E}e^{\lambda(X-\mathbb{E}X)} \le \exp(\lambda^2\sigma^2)$.
i.e. B's condition $\Rightarrow$ sub-exponential with $(\sqrt{2}\sigma, 2b)$.

Directly applying condition (*) in the Chernoff bound,

$$\mathbb{P}(X-\mathbb{E}X \ge t) \le e^{-\lambda t}\,\mathbb{E}e^{\lambda(X-\mathbb{E}X)} \le \exp\left(-\lambda t + \frac{1}{1-b|\lambda|}\cdot\frac{\lambda^2\sigma^2}{2}\right).$$ Let $\lambda = \frac{t}{bt+\sigma^2} \in [0, b^{-1})$,

$$\text{RHS} = -\frac{t^2}{bt+\sigma^2} + \frac{1}{1-bt/(bt+\sigma^2)}\cdot\frac{\sigma^2}{2}\cdot\frac{t^2}{(bt+\sigma^2)^2}$$

$$= -\frac{t^2}{bt+\sigma^2} + \frac{bt+\sigma^2}{\sigma^2}\cdot\frac{\sigma^2}{2}\cdot\frac{t^2}{(bt+\sigma^2)^2} = -\frac{t^2}{2(bt+\sigma^2)}$$

We have the following lemma.

Lemma If $\mathbb{E}\exp(\lambda(X-\mathbb{E}X)) \le \exp\left(\frac{\lambda^2\sigma^2}{2}\cdot\frac{1}{1-b|\lambda|}\right)$ for $|\lambda|<1/b$, then

$$P(X-\mathbb{E}X \ge t) \le \exp\left(-\frac{t^2}{2(bt+\sigma^2)}\right).$$

Example Recalling that if $|X-\mathbb{E}X|\le b$ then $X-\mathbb{E}X$ is subG with param b^2 the tail-bound for subG RVs gives $\mathbb{P}(X-\mathbb{E}X \ge t) \le \exp\left(-\frac{t^2}{2b^2}\right)$

$\equiv \quad X-\mathbb{E}X \le b\sqrt{2s} \quad$ w.p. $\ge 1-e^{-s}$

Let us try using the above Lemma. Set $s = \frac{t^2}{2(bt+\sigma^2)}$. Solving for t, $t = bs+\sqrt{b^2s^2+2\sigma^2 s}$.

So $X-\mathbb{E}X \le \sqrt{2\sigma^2 s}+2bs$ w.p. $\ge 1-e^{-s}$, where we used $\sqrt{a+b}\le\sqrt{a}+\sqrt{b}$ $\forall a,b>0$.

When $\sigma^2 \ll b^2$ which is often true when X occasionally take large values, the second bound is better.

Example Consider i.i.d. $X_i \in [-M, M]$, so that (*) holds with $b=M$.

$$\mathbb{E}\exp\left(\lambda\left(\tfrac{1}{n}\textstyle\sum X_i-\mathbb{E}X\right)\right) \le \prod_{i=1}^n \mathbb{E}\exp\left(\tfrac{\lambda}{n}(X_i-\mathbb{E}X)\right) \le \prod_{i=1}^n \exp\left(\frac{\lambda^2\sigma^2}{2n^2}\,\frac{1}{1-M\cdot|\lambda|/n}\right) \quad \text{for } |\lambda|<\frac{n}{M}$$

$$= \exp\left(\frac{\lambda^2\sigma^2}{2n}\cdot\frac{1}{1-M|\lambda|/n}\right).$$

So Lemma gives $\mathbb{P}\left(\frac{1}{n}\sum X_i - \mathbb{E}X \le -t\right) \le \exp\left(-\frac{nt^2}{2(\sigma^2+Mt)}\right)$.

We arrive at the following bound:

$$\mathbb{E}X \le \frac{1}{n}\sum X_i + \sqrt{\frac{2s\operatorname{Var}X}{n}} + 2M\cdot\frac{s}{n} \quad \text{w.p.} \ge 1-e^{-s}.$$

If $\operatorname{Var}X$ concentrates around the sample variance $\hat{\sigma}_n^2$, then

$$\mathbb{E}X \le \frac{1}{n}\sum X_i + \sqrt{2s\cdot\hat{\sigma}_n^2/n} + CM\,s/n \quad \text{w.h.p.}$$

Variance regularization

sample variance of $\ell(\theta;z)$

Given Bernstein's inequality we just derived, we have

$$\mathbb{E}_P \ell(\theta;Z) \le \frac{1}{n}\sum \ell(\theta;z_i) + \sqrt{\frac{Cs}{n}\widehat{\mathrm{Var}}(\ell(\theta;z))} + \frac{CMs}{n} \quad \text{w.p.} \ge 1-e^{-s}$$

for some numerical constant $c>0$.

Trade-off bias & variance optimally

So given this upper bound on the population loss, it is natural to optimize

$$\underset{\theta\in\Theta}{\text{minimize}} \left\{ \frac{1}{n}\sum \ell(\theta;z_i) + C\sqrt{\frac{\widehat{\mathrm{Var}}(\ell(\theta;z_i))}{n}} \right\}$$

↳ We call this the variance regularized problem.

See plot

<u>Problem</u> $\theta \mapsto \sqrt{\widehat{\mathrm{Var}}\,\ell(\theta;z)}$ is nonconvex even when $\theta\mapsto\ell(\theta;z)$ is convex.

Consider the empirical f-divergence DRO formulation with $f(t)=\frac{1}{2}(t-1)^2$, and $\rho_n>0$.

$$\max_Q \left\{ \mathbb{E}_Q \ell(\theta;Z) : D_f(Q,\hat{P}_n) = \mathbb{E}_{\hat{P}_n}\frac{1}{2}\left(\frac{dQ}{d\hat{P}_n}-1\right)^2 = \frac{1}{n}\sum_{i=1}^n \frac{1}{2}\left(\frac{q_i}{1/n}-1\right)^2 \le \rho_n \right\}$$

$$= \max_{q\ge 0} \left\{ \sum_{i=1}^n q_i \ell(\theta;z_i) : \frac{1}{2n}\sum_i^n (nq_i-1)^2 \le \rho_n,\ q^T\mathbb{1}=1 \right\}.$$

Taking $\rho_n = \frac{c}{n}$, we show that is actually equivalent to Variance Reg.

Fix $\theta\in\Theta$, and let $W=\ell(\theta;Z)$, $\hat{\mu}_n := \frac{1}{n}\sum_{i=1}^n W_i$, $\hat{\sigma}_n^2 := \frac{1}{n}\sum_1^n W_i^2 - \left(\frac{1}{n}\sum W_i\right)^2$.

Define $Q_n(c) := \left\{ q\in\mathbb{R}^n_+ : q^T\mathbb{1}=1,\ \frac{1}{2}\sum(nq_i-1)^2 = \frac{1}{2}\|nq-\mathbb{1}\|_2^2 \le c \right\}$

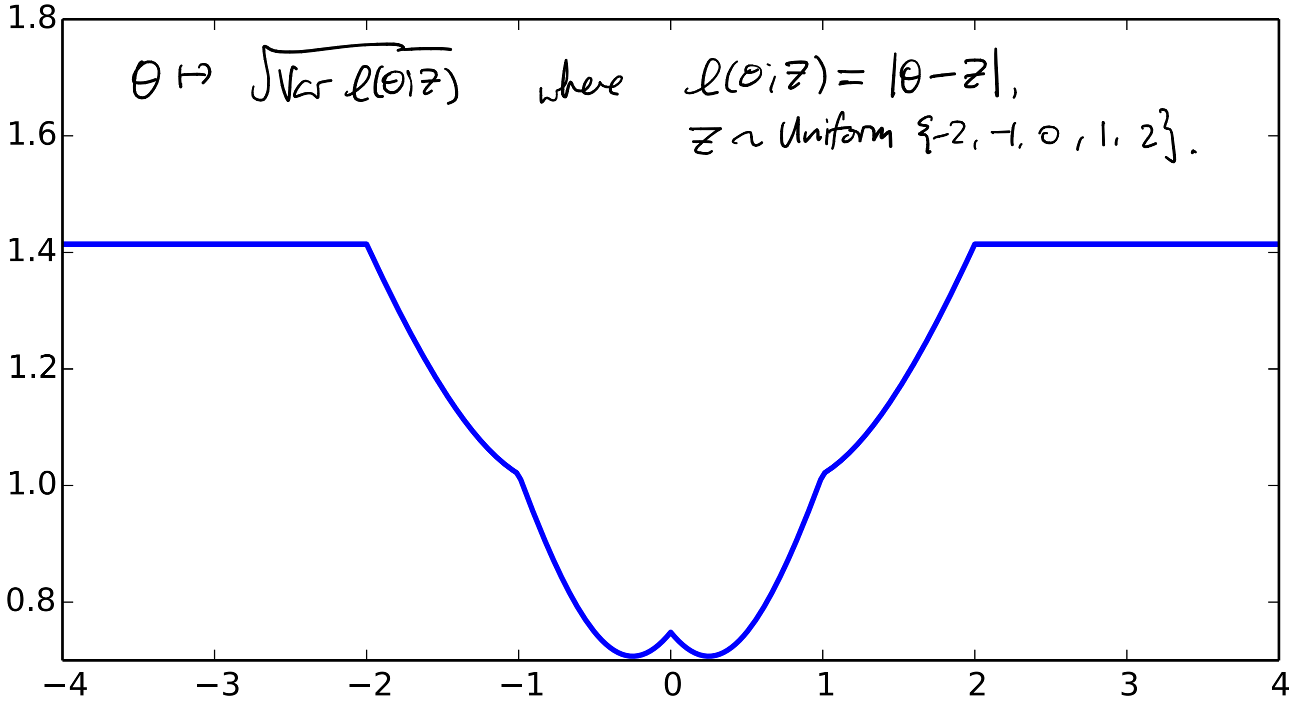
Doing a change of variables $u = q-\frac{1}{n}\mathbb{1}$, and denoting $\vec{W}_n := [W_1,\dots,W_n]\in\mathbb{R}^n$,

$$\sup_{q\in Q_n} q^T\vec{W}_n = \hat{\mu}_n + \sup\left\{ u^T(\vec{W}_n-\hat{\mu}_n\mathbb{1}) : u\ge -\frac{1}{n}\mathbb{1},\ u^T\mathbb{1}=0,\ \|u\|_2^2 \le 2c/n^2 \right\}.$$

From Cauchy-Schwarz, $u^T(\vec{W}_n-\hat{\mu}_n\mathbb{1}) \le \|u\|_2\|\vec{W}_n-\hat{\mu}_n\mathbb{1}\|_2 \le \frac{\sqrt{2c}}{n}\|\vec{W}_n-\hat{\mu}_n\mathbb{1}\|_2 = \sqrt{\frac{2c\hat{\sigma}_n^2}{n}}$

Equality is attained iff $u\propto \vec{W}_n-\hat{\mu}_n\mathbb{1}$, or $u_i^* = \frac{\sqrt{2c}\,(W_i-\hat{\mu}_n)}{n\|\vec{W}_n-\hat{\mu}_n\mathbb{1}\|_2} = \frac{\sqrt{2c}\,(W_i-\hat{\mu}_n)}{n\sqrt{n\hat{\sigma}_n^2}}$.

Such choice is possible if $u_i^*\ge -\frac{1}{n}\ \forall i$.



$\theta \mapsto \sqrt{\mathrm{Var}\, \ell(\theta; Z)}$ where $\ell(\theta; Z) = |\theta - Z|$,
$Z \sim \mathrm{Uniform}\{-2, -1, 0, 1, 2\}$.
1.8
1.6
1.4
1.2
1.0
0.8
−4
−3
−2
−1
0
1
2
3
4

So as long as $\sqrt{2c}\min_{1\le i\le n}(W_i-\hat{\mu}_n) \ge -\sqrt{n\hat{\sigma}_n^2} \equiv \hat{\sigma}_n \ge \sqrt{\frac{2c}{n}}\cdot\max_{1\le i\le n}(\hat{\mu}_n-W_i)$,

$$\sup_{q\in Q_n(c)} q^T\vec{W}_n = \hat{\mu}_n + \sqrt{\frac{2c}{n}\hat{\sigma}_n^2}.$$

By arguing that $\hat{\sigma}_n^2$ is large enough with high probability, we arrive at the following

<u>Theorem</u> Let $W\in[0,M]$ a.s., and $\sigma^2 := \mathrm{Var}\,W > 0$

$$\left(\sqrt{\frac{2c}{n}\hat{\sigma}_n^2}-\frac{2Mc}{n}\right)_+ \le \sup_{q\in Q_n(c)} q^T W - \hat{\mu}_n \le \sqrt{\frac{2c}{n}\hat{\sigma}_n^2},$$

and for $n \ge \max\left(2, \frac{M^2}{\sigma^2}\max(8c,44)\right)$, w.p. $\ge 1-\exp\left(-\frac{3n\sigma^2}{5M^2}\right)$

$$\sup_{q\in Q_n(c)} q^T W = \hat{\mu}_n + \sqrt{\frac{2c}{n}\hat{\sigma}_n^2}.$$

That is, we have shown

$$\frac{1}{n}\sum \ell(\theta;z_i) + c\cdot\sqrt{\frac{\hat{\mathrm{Var}}(\ell(\theta;z_i))}{n}} \overset{\star}{=} \sup_{q\in Q_n(c)} \sum_1^n q_i\,\ell(\theta;z_i) \quad \text{w.h.p.}$$

LHS : nonconvex, not coherent, but a natural quantity from a learning theoretic perspective

RHS : convex, coherent, "robust" w.r.t. reweighting.

A computationally tractable way of regularizing by variance.

Deep connections to EL

Remark 1) The guarantee $\star$ can be made uniform in $\theta\in\Theta$

2) A variant can be shown for any smooth f-divergence

3) An asymptotic version can be shown for any fast mixing seq of RV

So what does this give us?

upweight harder examples

$$\hat{\theta}_n^{rob} = \arg\min_{\theta\in\Theta} \sup_{q\in Q_n(c)} \sum_1^n q_i\,\ell(\theta;z_i)$$

$$\hat{\theta}_n^{ERM} = \arg\min_{\theta\in\Theta} \frac{1}{n}\sum_1^n \ell(\theta;z_i)$$

Optimal bias-variance trade-off

<u>Theorem</u> Let $c = s + C \cdot \mathfrak{R}_n\{\ell(\theta;\cdot) : \theta \in \Theta\}$. If $\ell(\theta; Z) \in [0, M]$ a.s.,

$$\mathbb{E}_{Z\sim P}\, \ell(\hat{\theta}_n^{\text{rob}}; Z) \le \inf_{\theta \in \Theta} \left\{ \mathbb{E}_P\, \ell(\theta; Z) + 2\sqrt{\frac{2c}{n} \operatorname{Var} \ell(\theta; Z)} \right\} + \frac{cM}{n} \cdot a$$

w.p. $\ge 1 - e^{-s}$, for some numerical const $a > 0$.

Using a uniform version of Bernstein's inequality, w.p. $\ge 1 - e^{-s}$,

$$\mathbb{E}_{Z\sim P}\, \ell(\hat{\theta}_n^{\text{ERM}}; Z) \le \inf_{\theta\in\Theta} \mathbb{E}_P\, \ell(\theta; Z) + \sqrt{\frac{2cM}{n} \inf_{\theta \in \Theta} \mathbb{E}_P\, \ell(\theta; Z)} + \frac{aMc}{n}.$$

If $\operatorname{Var} \ell(\theta^*; Z) \ll M \cdot \mathbb{E}_P\, \ell(\theta^*; Z)$, then bound for $\hat{\theta}_n^{\text{rob}}$ is tighter.

cf. We can also construct an explicit (contrived) example where
$\mathbb{E}_{Z\sim P}\, \ell(\hat{\theta}_n^{\text{rob}}; Z) \le \inf_{\theta\in\Theta} \mathbb{E}_P\, \ell(\theta; Z) + \frac{a}{n}$ but $\mathbb{E}_{Z\sim P}\, \ell(\hat{\theta}_n^{\text{ERM}}; Z) \ge \inf_{\theta\in\Theta} \mathbb{E}_P\, \ell(\theta; Z) + \frac{a'}{\sqrt{n}}$.

Wasserstein DRO & Regularization

By choosing certain cost functions, we can show that Wasserstein DRO is equivalent to classical regularizers.

<u>Proposition</u> (Regression) Consider $c((x,y),(x',y')) = \|(x,y) - (x',y')\|_k^2$ for some $k \in (1, \infty]$.

$$\sup_{Q:\, W_c(Q, \hat{P}_n) \le \rho} \mathbb{E}_Q (Y - \theta^T X)^2 = \left(\left(\frac{1}{n} \sum_{i}^{n} (Y_i - \theta^T X_i)^2 \right)^{\frac{1}{2}} + \sqrt{\rho}\, \|[\theta, 1]\|_{k_*} \right)^2$$

where $k_* = k/(k-1)$ so $\frac{1}{k} + \frac{1}{k_*} = 1$.

<u>Proof of main proposition</u> To ease notation, define $Z = (X, Y)$, $\bar{\theta} = [\theta, -1] \in \mathbb{R}^{d+1}$.
From the duality result for Wasserstein DRO,

$$\sup_{Q:\, W_c(Q, \hat{P}_n) \le \rho} \mathbb{E}_Q (Y - \theta^T X)^2 = \inf_{\lambda \ge 0} \left\{ \lambda \rho + \mathbb{E}_{\hat{P}_n} \sup_{z'} \left\{ (\bar{\theta}^T z')^2 - \lambda \|Z - z'\|_k^2 \right\} \right\}$$

We first simplify the robust surrogate loss $\sup_{z'} \left\{ (\bar{\theta}^T z')^2 - \lambda \|Z - z'\|_k^2 \right\} = \phi_\lambda(\bar{\theta}; Z)$

Doing a change of variable $\Delta = z - z'$. since sup should be attained when signs match.

$$\phi_\lambda(\bar\theta; z) = \sup_\Delta \left\{ (\bar\theta^T z + \bar\theta^T \Delta)^2 - \lambda \|\Delta\|_k^2 \right\} = \sup_\Delta \left\{ (\bar\theta^T z + \mathrm{sgn}(\bar\theta^T z)\cdot \bar\theta^T \Delta)^2 - \lambda \|\Delta\|_k^2 \right\}.$$

$$= \sup_\Delta \left\{ (\bar\theta^T z + \mathrm{sgn}(\bar\theta^T z)\cdot \|\bar\theta\|_{k*} \|\Delta\|_k)^2 - \lambda \|\Delta\|_k^2 \right\}$$ since Holder's inequality is tight for some choice of Δ

$$= (\bar\theta^T z)^2 + \sup_\Delta \left\{ -(\lambda - \|\bar\theta\|_{k*}) \cdot \|\Delta\|_k^2 + 2|\bar\theta^T z| \, \|\bar\theta\|_{k*} \|\Delta\|_k \right\}$$

$$= \begin{cases} \dfrac{\gamma}{\gamma - \|\bar\theta\|_{k*}^2} (\bar\theta^T z)^2 & \text{if } \lambda > \|\bar\theta\|_{k*} \\ \infty & \text{o/w} \end{cases}$$

So we conclude $\sup_{Q:\, W_c(Q,\hat P_n)\le \rho} \mathbb{E}_Q (Y - \theta^T X)^2 = \inf_{\lambda > \|\bar\theta\|_{k*}} \left\{ \lambda\rho + \dfrac{\lambda}{\lambda - \|\bar\theta\|_{k*}^2} \dfrac{1}{n}\sum_1^n (\bar\theta^T z_i)^2 \right\}$.

Noting that the optimum is achieved at $\lambda^* = \|\bar\theta\|_{k*}^2 + \left(\dfrac{\|\bar\theta\|_{k*}^2}{\rho} \dfrac{1}{n}\sum_1^n (\bar\theta^T z_i)^2 \right)^{\frac{1}{2}}$, we have the result. $\boxtimes$

Similarly, we can consider only perturbing the feature vector by setting

$$c((x,y),(x',y')) = \begin{cases} \|x - x'\|_k^2 & \text{if } y = y' \\ \infty & \text{if } y \neq y' \end{cases}$$ "covariate shift"

$$\sup_{Q:\, W_c(Q,\hat P_n)\le \rho} \mathbb{E}_Q (Y - \theta^T X)^2 = \left(\left(\frac{1}{n}\sum_1^n (Y_i - \theta^T X_i)^2 \right)^{\frac{1}{2}} + \sqrt{\rho}\, \|\theta\|_{k*} \right)^2$$

We can show a similar equivalence for linear classification models. $Y \in \{\pm 1\}$

$$c((x,y),(x',y')) = \begin{cases} \|x - x'\|_k & \text{if } y = y' \\ \infty & \text{if } y \neq y' \end{cases}$$

$$\sup_{Q:\, W_c(Q,\hat P_n)\le \rho} \mathbb{E}_Q \log(1 + e^{-Y\theta^T X}) = \frac{1}{n}\sum_1^n \log(1 + e^{-Y_i \theta^T X_i}) + \rho \|\theta\|_{k*}$$

$$\sup_{Q:\, W_c(Q,\hat P_n)\le \rho} \mathbb{E}_Q (1 - Y\theta^T X)_+ = \frac{1}{n}\sum_1^n (1 - Y_i \theta^T X_i)_+ + \rho \|\theta\|_{k*}$$

Another basic connection

Consider the cost function $c(z,z') = \frac{1}{2}\|z-z'\|_2^2$, and the corresponding robust surrogate function $\phi_\lambda(\theta; z) = \sup_{z'} \{\ell(\theta; z') - \frac{\lambda}{2}\|z - z'\|_2^2\}$.

Plugging the first order approximation into the robust surrogate

$$* = \sup_{z'} \{\ell(\theta; z) + \nabla_z \ell(\theta; z)^T(z' - z) - \frac{\lambda}{2}\|z - z'\|_2^2\}.$$

gradient ascent step for inc. loss

The max is attained at $\nabla_z \ell(\theta; z) = \lambda(z - z') \Leftrightarrow z' = z + \frac{1}{\lambda}\nabla_z \ell(\theta; z)$,

and $* = \ell(\theta; z) + \frac{1}{2\lambda}\|\nabla_z \ell(\theta; z)\|_2^2$.

Plugging this first order approximation in to the dual, we get

$$\inf_{\lambda \geq 0} \left\{ \lambda\rho + \frac{1}{n}\sum_{i=1}^n \sup_{z'} \{\ell(\theta; z_i) + \nabla_z \ell(\theta; z_i)^T(z' - z_i) - \frac{\lambda}{2}\|z_i - z'\|_2^2\} \right\}$$

$$= \frac{1}{n}\sum^n \ell(\theta; z_i) + \inf_{\lambda \geq 0} \left\{ \lambda\rho + \frac{1}{2\lambda} \frac{1}{n}\sum_i^n \|\nabla_z \ell(\theta; z_i)\|_2^2 \right\}$$

$$= \frac{1}{n}\sum_i^n \ell(\theta; z_i) + \sqrt{\rho} \cdot \left(\mathbb{E}_{\hat{P}_n} \|\nabla_z \ell(\theta; z)\|_2^2\right)^{\frac{1}{2}}$$

↳ Regularize to make the loss more stable against data perturbations

For smooth losses, Wasserstein DRO regularizes to make $\|\nabla_z \ell(\theta; z)\|$ small, up to first order.