Robust Optimization as a Convex Variance Regularization

Hongseok Namkoong
(Joint work with John Duchi and Peter Glynn)

Stanford University

September 2016

Stochastic optimization problems

$$\text{minimize } \mathbb{E}_{P_0}[\ell(\theta; X)] = \int \ell(\theta; X) dP_0(X)$$
$$\text{subject to } \theta \in \Theta.$$

Stochastic optimization problems

$$\begin{aligned} &\text{minimize } \mathbb{E}_{P_0}[\ell(\theta; X)] = \int \ell(\theta; X) dP_0(X) \\ &\text{subject to } \theta \in \Theta. \end{aligned}$$

- Data/randomness is X
- Parameter space Θ is a nonempty closed set

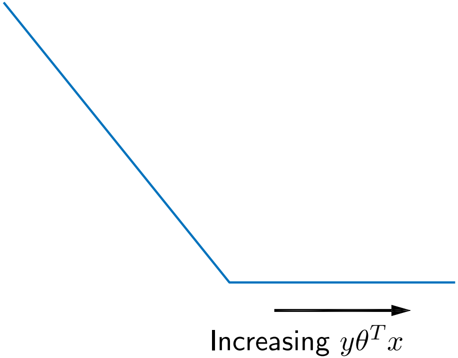
Applications

Machine learning all sorts of loss minimization problems, e.g. classification:

$$X = (x, y) \in \mathbb{R}^d \times \{-1, 1\},$$

goal is to find θ such that $\mathrm{sign}(\theta^T x) = y$ usually.

$$\ell(\theta; X) = \ell(\theta; (x, y)) = \big(1 - y\theta^T x\big)_+$$



Goal of This Talk

How do we optimize?

$$\underset{\theta\in\Theta}{\text{minimize}}\, R(\theta) = \mathbb{E}_{P_0}[\ell(\theta; X)]$$

- Expensive to compute $\mathbb{E}_{P_0}$ (simulation optimization) and P_0 often unkown (statistics, machine learning)

Goal: Given i.i.d. samples $X_1, \ldots, X_n \stackrel{\text{iid}}{\sim} P_0$, how can we say with **confidence** that our algorithm has learned something **useful**?

Empirical Risk Minimization / Sample Average Approximation

Standard approach: Solve

Empirical Risk Minimization / Sample Average Approximation

Standard approach: Solve

$$\widehat{\theta}^{\mathrm{erm}} \in \operatorname*{argmin}_{\theta \in \Theta} \widehat{R}_n(\theta) := \frac{1}{n} \sum_{i=1}^{n} \ell(\theta; X_i) \approx R(\theta).$$

Empirical Risk Minimization / Sample Average Approximation

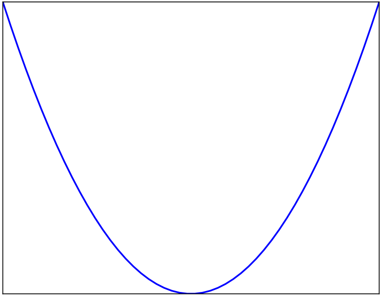
Standard approach: Solve

$$\operatorname*{argmin}_{\theta \in \Theta} \widehat{R}_n(\theta) := \frac{1}{n} \sum_{i=1}^{n} \ell(\theta; X_i) \; \underbrace{\approx R(\theta)}_{\text{Hopefully!}} \; .$$

A few asides

Why do we like convex optimization problems?

- We can solve them (algorithms)
- We can certify they are *solved* (duality)



- We want to do the same thing for stochastic problems!

Point of departure: bias/variance tradeoff

Point of departure: bias/variance tradeoff

- Any learning algorithm has *bias* (approximation error) and *variance* (estimation error)

Point of departure: bias/variance tradeoff

- Any learning algorithm has *bias* (approximation error) and *variance* (estimation error)
- From empirical Bernstein's inequality, with probability $1 - \delta$

$$R(\theta) \leq \underbrace{\widehat{R}_n(\theta)}_{\text{bias}} + \underbrace{\sqrt{\frac{2\mathrm{Var}_{\widehat{P}_n}(\ell(\theta; X))}{n}}}_{\text{variance}} + \frac{C \log \frac{1}{\delta}}{n}$$

Point of departure: bias/variance tradeoff

- Any learning algorithm has *bias* (approximation error) and *variance* (estimation error)
- From empirical Bernstein's inequality, with probability $1 - \delta$

$$R(\theta) \le \underbrace{\widehat{R}_n(\theta)}_{\text{bias}} + \underbrace{\sqrt{\frac{2\mathrm{Var}_{\widehat{P}_n}\left(\ell(\theta; X)\right)}{n}}}_{\text{variance}} + \frac{C \log \frac{1}{\delta}}{n}$$

- Can be made uniform in $\theta \in \Theta$ [Maurer & Pontil 09]

Point of departure: bias/variance tradeoff

- Any learning algorithm has *bias* (approximation error) and *variance* (estimation error)
- From empirical Bernstein's inequality, with probability $1 - \delta$

$$R(\theta) \le \underbrace{\widehat{R}_n(\theta)}_{\text{bias}} + \underbrace{\sqrt{\frac{2\mathrm{Var}_{\widehat{P}_n}\left(\ell(\theta; X)\right)}{n}}}_{\text{variance}} + \frac{C \log \frac{1}{\delta}}{n}$$

- Can be made uniform in $\theta \in \Theta$ [Maurer & Pontil 09]

Goal: Trade between these automatically and optimally by solving

$$\widehat{\theta}^{\mathrm{var}} \in \operatorname*{argmin}_{\theta \in \Theta} \left\{ \widehat{R}_n(\theta) + \sqrt{\frac{2\mathrm{Var}_{\widehat{P}_n}\left(\ell(\theta; X)\right)}{n}} \right\}.$$

Optimizing for bias and variance

Good idea: Directly minimize bias + variance, certify optimality!

Optimizing for bias and variance

Good idea: Directly minimize bias + variance, certify optimality!
Minor issue: variance is wildly non-convex

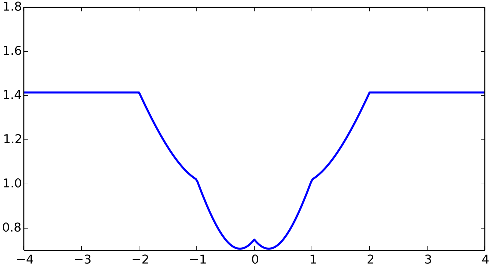


Figure: Variance of $|\theta - X|$

Robust ERM

Goal:

$$\underset{\theta \in \Theta}{\text{minimize}} \; R(\theta) = \mathbb{E}_{P_0}[\ell(\theta; X)]$$

Robust ERM

Goal:

$$\underset{\theta \in \Theta}{\text{minimize}} \;\; R(\theta) = \mathbb{E}_{P_0}[\ell(\theta; X)]$$

Solve empirical risk minimization problem

$$\underset{\theta \in \Theta}{\text{minimize}} \;\; \sum_{i=1}^{n} \frac{1}{n} \ell(\theta; X_i)$$

Robust ERM

Goal:

$$\underset{\theta \in \Theta}{\text{minimize}} \;\; R(\theta) = \mathbb{E}_{P_0}[\ell(\theta; X)]$$

Solve empirical risk minimization problem

$$\underset{\theta \in \Theta}{\text{minimize}} \;\; \sum_{i=1}^{n} \frac{1}{n} \ell(\theta; X_i)$$

Robust ERM

Goal:

$$\underset{\theta \in \Theta}{\text{minimize}} \;\; R(\theta) = \mathbb{E}_{P_0}[\ell(\theta; X)]$$

Instead, solve *distributionally robust optimization (RO) problem*

$$\underset{\theta \in \Theta}{\text{minimize}} \;\; \sup_{p \in \mathcal{P}_{n,\rho}} \sum_{i=1}^{n} p_i \ell(\theta; X_i)$$

where $\mathcal{P}_{n,\rho}$ is some appropriately chosen set of vectors

Robust ERM

Goal:

$$\underset{\theta \in \Theta}{\text{minimize}} \; R(\theta) = \mathbb{E}_{P_0}[\ell(\theta; X)]$$

Instead, solve *distributionally robust optimization (RO) problem*

$$\underset{\theta \in \Theta}{\text{minimize}} \; \sup_{p \in \mathcal{P}_{n,\rho}} \sum_{i=1}^{n} p_i \ell(\theta; X_i)$$

where $\mathcal{P}_{n,\rho}$ is some appropriately chosen set of vectors

Today: Give a principled statistical approach to choosing $\mathcal{P}_{n,\rho}$ and give stochastic optimality certificates for RO.

Empirical likelihood

Idea: Instead of using empirical distribution $\widehat{P}_n$ on sample $X_1, \ldots, X_n$, look at all distributions "near" it.

Empirical likelihood

Idea: Instead of using empirical distribution $\widehat{P}_n$ on sample $X_1, \ldots, X_n$, look at all distributions "near" it.

- The *f-divergence* between distributions P and Q is

$$D_f\left(P \| Q\right) := \int f\left(\frac{dP}{dQ}\right) dQ$$

  where f is some convex function with $f(1) = 0$.
  (w.l.o.g. can take $f'(1) = 0$ too)

Empirical likelihood

Idea: Instead of using empirical distribution $\widehat{P}_n$ on sample $X_1, \ldots, X_n$, look at all distributions "near" it.

- The f-*divergence* between distributions P and Q is

$$D_f(P\|Q) := \int f\left(\frac{dP}{dQ}\right) dQ$$

where f is some convex function with $f(1) = 0$.
(w.l.o.g. can take $f'(1) = 0$ too)

- Measures of closeness we use: $f(t) = \frac{1}{2}(t-1)^2$

$$D_{\chi^2}(P\|Q) = \frac{1}{2}\sum_x \frac{(p(x) - q(x))^2}{q(x)} \qquad \text{Chi-square}$$

(Owen (1990): original empirical likelihood $f(t) = -\log t$)

Empirical likelihood

$$E_n(\rho) := \left\{ \sum_{i=1}^n p_i Z_i : D_{\chi^2}\left(p \| \mathbb{1}/n\right) \leq \frac{\rho}{n} \right\}$$

then *independently of distribution* on $Z \in \mathbb{R}^k$

$$\mathbb{P}(\mathbb{E}[Z] \in E_n(\rho)) \to \mathbb{P}(\chi_k^2 \leq \rho).$$

ellipse [Owen 90, Baggerly 98, Newey and Smith 01, Imbens 02]

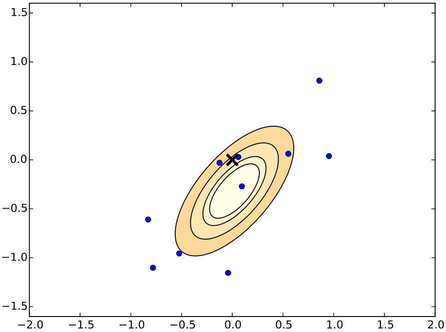
Empirical likelihood

$$E_n(\rho) := \left\{ \sum_{i=1}^n p_i Z_i : D_{\chi^2}\left(p \| \mathbb{1}/n\right) \le \frac{\rho}{n} \right\}$$

then *independently of distribution* on $Z \in \mathbb{R}^k$

$$\mathbb{P}(\mathbb{E}[Z] \in E_n(\rho)) \to \mathbb{P}(\chi_k^2 \le \rho).$$

ellipse [Owen 90, Baggerly 98, Newey and Smith 01, Imbens 02]



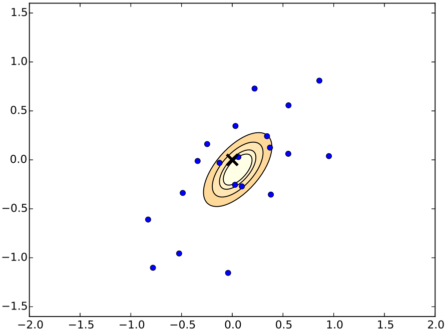
Empirical likelihood

$$E_n(\rho) := \left\{ \sum_{i=1}^{n} p_i Z_i : D_{\chi^2}\left(p \| \mathbb{1}/n\right) \le \frac{\rho}{n} \right\}$$

then *independently of distribution* on $Z \in \mathbb{R}^k$

$$\mathbb{P}(\mathbb{E}[Z] \in E_n(\rho)) \to \mathbb{P}(\chi_k^2 \le \rho).$$

ellipse [Owen 90, Baggerly 98, Newey and Smith 01, Imbens 02]



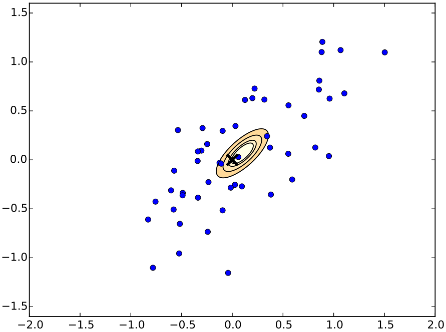
Empirical likelihood

$$E_n(\rho) := \left\{ \sum_{i=1}^{n} p_i Z_i : D_{\chi^2}\left(p \| \mathbb{1}/n\right) \le \frac{\rho}{n} \right\}$$

then *independently of distribution* on $Z \in \mathbb{R}^k$

$$\mathbb{P}(\mathbb{E}[Z] \in E_n(\rho)) \to \mathbb{P}(\chi_k^2 \le \rho).$$

ellipse [Owen 90, Baggerly 98, Newey and Smith 01, Imbens 02]



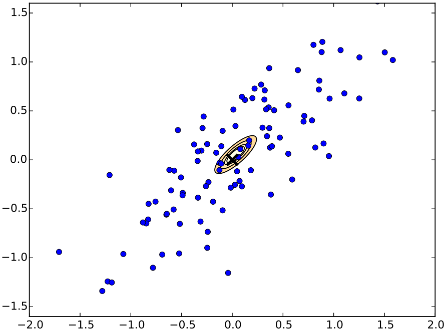
Empirical likelihood

$$E_n(\rho) := \left\{ \sum_{i=1}^n p_i Z_i : D_{\chi^2}\left(p \| \mathbb{1}/n\right) \leq \frac{\rho}{n} \right\}$$

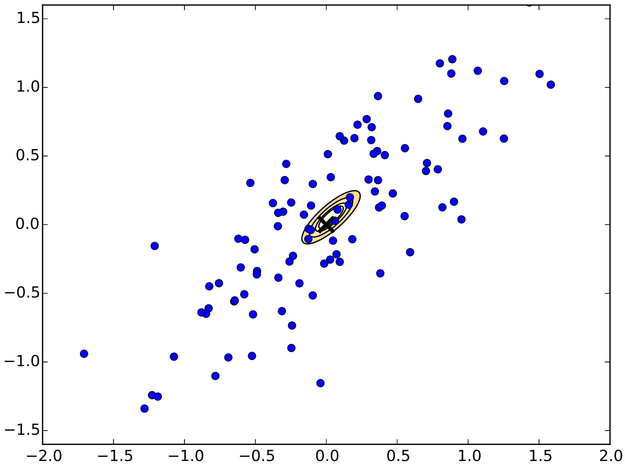
then *independently of distribution* on $Z \in \mathbb{R}^k$

$$\mathbb{P}(\mathbb{E}[Z] \in E_n(\rho)) \to \mathbb{P}(\chi^2_k \leq \rho).$$

ellipse [Owen 90, Baggerly 98, Newey and Smith 01, Imbens 02]



Empirical likelihood



Idea: Leverage this in robust and stochastic optimization

Robust Optimization

Idea: Optimize over *uncertainty set* of possible distributions,

$$\mathcal{P}_{n,\rho} := \left\{ \text{Distributions } P \text{ such that } D_{\chi^2}\left(P \| \widehat{P}_n\right) \leq \frac{\rho}{n} \right\}$$

for some $\rho > 0$.

Robust Optimization

Idea: Optimize over *uncertainty set* of possible distributions,

$$\mathcal{P}_{n,\rho} := \left\{ \text{Distributions } P \text{ such that } D_{\chi^2}\left(P \| \widehat{P}_n\right) \le \frac{\rho}{n} \right\}$$

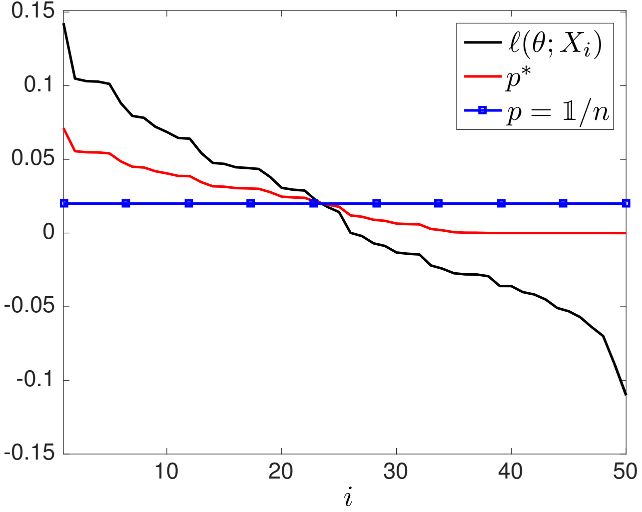
for some $\rho > 0$.

Define (and optimize) *empirical likelihood upper confidence bound*

$$R_n(\theta, \mathcal{P}_{n,\rho}) := \max_{P : D_{\chi^2}\left(P \| \widehat{P}_n\right) \le \frac{\rho}{n}} \mathbb{E}_P[\ell(\theta; X)] = \max_{p : D_{\chi^2}\left(P \| \widehat{P}_n\right) \le \frac{\rho}{n}} \sum_{i=1}^n p_i \ell(\theta; X_i)$$

[Ben-Tal et al. 13, Bertsimas et al. 16, Lam & Zhou 16]

Visualization of worst-case



Robust Optimization

Solve

$$\widehat{\theta}^{\mathrm{rob}} := \operatorname*{argmin}_{\theta \in \Theta} \left\{ R_n(\theta, \mathcal{P}_{n,\rho}) := \max_{P : D_{\chi^2}\left(P \| \widehat{P}_n\right) \le \frac{\rho}{n}} \mathbb{E}_P[\ell(\theta; X)] \right\}.$$

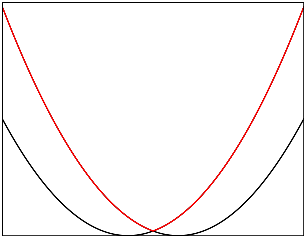
Robust Optimization

Solve

$$\widehat{\theta}^{\mathrm{rob}} := \operatorname*{argmin}_{\theta \in \Theta} \left\{ R_n(\theta, \mathcal{P}_{n,\rho}) := \max_{P : D_{\chi^2}\left(P \| \widehat{P}_n\right) \le \frac{\rho}{n}} \mathbb{E}_P[\ell(\theta; X)] \right\}.$$

Nice properties:

- Convex optimization problem.
- Solve dual reformulation using interior point methods [Ben-Tal et al. 13]
- For large n and d, efficient solution methods as fast as SGD [N. & Duchi, 16] 1



Robust Optimization $\approx$ Variance Regularization

Theorem (Duchi & N. 2016)

Assume that $\ell(\theta; X) \le M$*. Let* $\sigma^2(\theta) := \mathrm{Var}(\ell(\theta; X))$.

$$R_n(\theta; \mathcal{P}_{n,\rho}) = \widehat{R}_n(\theta) + \sqrt{\frac{2\rho \mathrm{Var}_{\widehat{P}_n}(\ell(\theta; X))}{n}} + \mathit{Rem}_n(\theta).$$

- $\mathit{Rem}_n(\theta) \le \frac{\sqrt{12}\rho M}{n}$
- $\mathit{Rem}_n(\theta) = 0$ *with probability at least* $1 - \exp(-\frac{n\sigma^2(\theta)}{36M^2})$ proof

Robust Optimization $\approx$ Variance Regularization

Theorem (Duchi & N. 2016)

Assume that $\ell(\theta; X) \le M$. *Let* $\sigma^2(\theta) := \mathrm{Var}(\ell(\theta; X))$.

$$R_n(\theta; \mathcal{P}_{n,\rho}) = \widehat{R}_n(\theta) + \sqrt{\frac{2\rho \mathrm{Var}_{\widehat{P}_n}(\ell(\theta; X))}{n}} + \mathit{Rem}_n(\theta).$$

- $\mathit{Rem}_n(\theta) \le \frac{\sqrt{12}\rho M}{n}$
- $\mathit{Rem}_n(\theta) = 0$ *with probability at least* $1 - \exp(-\frac{n\sigma^2(\theta)}{36M^2})$ proof
- *Let* $N(\mathcal{F}, \tau, \|\cdot\|_{L^\infty})$ *be the* τ*-covering number with respect to the supremum norm.*

$$\mathbb{P}\left(\mathit{Rem}_n(\theta) = 0 \text{ for all } \theta \in \Theta \text{ s.t. } \sigma^2(\theta) \ge \tau^2\right)$$
$$\ge 1 - cN(\mathcal{F}, \tau, \|\cdot\|_{L^\infty}) \exp(-\frac{n\tau^2}{M^2}).$$

Robust Optimization $\approx$ Variance Regularization

Theorem (Duchi, Glynn & N. 2016)

For general f-divergences,

$$R_n(\theta; \mathcal{P}_{n,\rho}) = \widehat{R}_n(\theta) + \sqrt{\frac{2\rho \mathrm{Var}_{\widehat{P}_n}(\ell(\theta; X))}{n}} + \mathit{Rem}_n(\theta).$$

- *If $\sigma^2(\theta) < \infty$, then $\sqrt{n}\mathit{Rem}_n(\theta) \stackrel{P^*}{\to} 0$*
- *If $\{\ell(\theta; \cdot) : \theta \in \Theta\}$ is P_0-Donsker, then $\sqrt{n} \sup_{\theta \in \Theta} \mathit{Rem}_n(\theta) \stackrel{P^*}{\to} 0$*

Robust Optimization $\approx$ Variance Regularization

Theorem (Duchi, Glynn & N. 2016)

For general f-divergences,

$$R_n(\theta; \mathcal{P}_{n,\rho}) = \widehat{R}_n(\theta) + \sqrt{\frac{2\rho \mathrm{Var}_{\widehat{P}_n}(\ell(\theta; X))}{n}} + \mathit{Rem}_n(\theta).$$

- *If* $\sigma^2(\theta) < \infty$*, then* $\sqrt{n}\mathit{Rem}_n(\theta) \stackrel{P^*}{\to} 0$
- *If* $\{\ell(\theta; \cdot) : \theta \in \Theta\}$ *is* P_0*-Donsker, then* $\sqrt{n} \sup_{\theta \in \Theta} \mathit{Rem}_n(\theta) \stackrel{P^*}{\to} 0$
- *[Lam 13] showed non-statistical, pointwise version for KL-divergence*
- *[Gotoh et al. 15] showed similar pointwise results with the objective penalty version*

Robust Optimization $\approx$ Variance Regularization

With high probability,

$$\underbrace{R_n(\theta; \mathcal{P}_{n,\rho})}_{\text{Robust}} = \underbrace{\widehat{R}_n(\theta) + \sqrt{\frac{2\rho \mathrm{Var}_{\widehat{P}_n}\left(\ell(\theta; X)\right)}{n}}}_{\text{VarReg}}$$

Robust Optimization $\approx$ Variance Regularization

With high probability,

$$\underbrace{R_n(\theta; \mathcal{P}_{n,\rho})}_{\text{Robust}} = \underbrace{\widehat{R}_n(\theta) + \sqrt{\frac{2\rho \mathrm{Var}_{\widehat{P}_n}(\ell(\theta; X))}{n}}}_{\text{VarReg}}$$

- Robust is empirical likelihood UCB and VarReg is normal UCB

Robust Optimization $\approx$ Variance Regularization

With high probability,

$$\underbrace{R_n(\theta; \mathcal{P}_{n,\rho})}_{\text{Robust}} = \underbrace{\widehat{R}_n(\theta) + \sqrt{\frac{2\rho \mathrm{Var}_{\widehat{P}_n}\left(\ell(\theta; X)\right)}{n}}}_{\text{VarReg}}$$

- Robust is empirical likelihood UCB and VarReg is normal UCB
- Robust is convex, VarReg is non-convex

Robust Optimization $\approx$ Variance Regularization

With high probability,

$$\underbrace{R_n(\theta; \mathcal{P}_{n,\rho})}_{\text{Robust}} = \underbrace{\widehat{R}_n(\theta) + \sqrt{\frac{2\rho \mathrm{Var}_{\widehat{P}_n}(\ell(\theta; X))}{n}}}_{\text{VarReg}}$$

- Robust is empirical likelihood UCB and VarReg is normal UCB
- Robust is convex, VarReg is non-convex
- Robust **only** penalizes upward (bad) deviations in the loss whereas VarReg penalizes downward (good) deviations along with the upward (bad) deviations

Robust Optimization $\approx$ Variance Regularization

With high probability,

$$\underbrace{R_n(\theta; \mathcal{P}_{n,\rho})}_{\text{Robust}} = \underbrace{\widehat{R}_n(\theta) + \sqrt{\frac{2\rho \mathrm{Var}_{\widehat{P}_n}(\ell(\theta; X))}{n}}}_{\text{VarReg}}$$

- Robust is empirical likelihood UCB and VarReg is normal UCB
- Robust is convex, VarReg is non-convex
- Robust **only** penalizes upward (bad) deviations in the loss whereas VarReg penalizes downward (good) deviations along with the upward (bad) deviations
- Robust is a coherent risk measure (i.e. it is a sensible negative utility)

Empirical likehood for stochastic optimization

Solve

$$\widehat{\theta}^{\mathrm{rob}} := \operatorname*{argmin}_{\theta \in \Theta} \left\{ R_n(\theta, \mathcal{P}_{n,\rho}) := \max_{P : D_{\chi^2}\left(P \| \widehat{P}_n\right) \le \frac{\rho}{n}} \mathbb{E}_P[\ell(\theta; X)] \right\}.$$

Empirical likehood for stochastic optimization

Solve

$$\widehat{\theta}^{\mathrm{rob}} := \operatorname*{argmin}_{\theta \in \Theta} \left\{ R_n(\theta, \mathcal{P}_{n,\rho}) := \max_{P : D_{\chi^2}\left(P \| \widehat{P}_n\right) \le \frac{\rho}{n}} \mathbb{E}_P[\ell(\theta; X)] \right\}.$$

Assume that $\{\ell(\theta; \cdot) : \theta \in \Theta\}$ is P_0-Donsker

e.g. $\Theta \subset \mathbb{R}^d$ compact and $\ell(\cdot; X)$ is $M(X)$-Lipschitz with $\mathbb{E}M(X)^2 < \infty$.

Empirical likehood for stochastic optimization

Solve

$$\widehat{\theta}^{\text{rob}} := \operatorname*{argmin}_{\theta \in \Theta} \left\{ R_n(\theta, \mathcal{P}_{n,\rho}) := \max_{P : D_{\chi^2}\left(P \| \widehat{P}_n\right) \le \frac{\rho}{n}} \mathbb{E}_P[\ell(\theta; X)] \right\}.$$

Assume that $\{\ell(\theta; \cdot) : \theta \in \Theta\}$ is P_0-Donsker

e.g. $\Theta \subset \mathbb{R}^d$ compact and $\ell(\cdot; X)$ is $M(X)$-Lipschitz with $\mathbb{E}M(X)^2 < \infty$.

Theorem (Duchi, Glynn & N. 16 [1])

If $\theta^\star := \operatorname{argmin}_{\theta \in \Theta} R(\theta)$ *is unique, then*

$$\lim_{n \to \infty} \mathbb{P}\left(\inf_{\theta \in \Theta} R(\theta) \le R_n(\widehat{\theta}^{\text{rob}}, \mathcal{P}_{n,\rho}) \right) = \mathbb{P}\left(N(0,1) \ge -\sqrt{2\rho} \right).$$

Optimal bias variance tradeoff

Solve

$$\widehat{\theta}^{\mathrm{rob}} := \operatorname*{argmin}_{\theta \in \Theta} \left\{ R_n(\theta, \mathcal{P}_{n,\rho}) := \max_{P : D_{\chi^2}\left(P \| \widehat{P}_n\right) \le \frac{\rho}{n}} \mathbb{E}_P[\ell(\theta; X)] \right\}.$$

Optimal bias variance tradeoff

Solve

$$\widehat{\theta}^{\mathrm{rob}} := \operatorname*{argmin}_{\theta \in \Theta} \left\{ R_n(\theta, \mathcal{P}_{n,\rho}) := \max_{P : D_{\chi^2}\left(P \| \widehat{P}_n\right) \le \frac{\rho}{n}} \mathbb{E}_P[\ell(\theta; X)] \right\}.$$

Let $\ell(\cdot; X)$ is M-Lipschitz and $\mathrm{diam}(\Theta) = r$

Optimal bias variance tradeoff

Solve

$$\widehat{\theta}^{\text{rob}} := \underset{\theta\in\Theta}{\operatorname{argmin}} \left\{ R_n(\theta, \mathcal{P}_{n,\rho}) := \max_{P : D_{\chi^2}\left(P \| \widehat{P}_n\right) \le \frac{\rho}{n}} \mathbb{E}_P[\ell(\theta; X)] \right\}.$$

Let $\ell(\cdot; X)$ is M-Lipschitz and $\operatorname{diam}(\Theta) = r$

Theorem (Duchi & N. 2016)

Let $\rho = \log \frac{1}{\delta} + d \log n$. *Then with probability at least* $1 - \delta$,

$$\begin{aligned} R(\widehat{\theta}^{\text{rob}}) &\le \underbrace{R_n(\widehat{\theta}^{\text{rob}}, \mathcal{P}_{n,\rho})}_{\text{optimality certificate}} + \frac{crM}{n}\rho \\ &\le \underbrace{\min_{\theta\in\Theta} \left\{ R(\theta) + 2\sqrt{\frac{2\rho \operatorname{Var}(\ell(\theta, \xi))}{n}} \right\}}_{\text{optimal tradeoff}} + \frac{crM}{n}\rho \end{aligned}$$

for some universal constant $0 < c \le 30$.

Fast rates from optimal tradeoff

Theorem (Duchi & N. 2016)

Let $\rho = \log \frac{1}{\delta} + d \log n$. *Then with probability at least* $1 - \delta$,

$$R(\widehat{\theta}^{\mathrm{rob}}) \le \min_{\theta \in \Theta} \left\{ R(\theta) + 2\sqrt{\frac{2\rho \mathrm{Var}(\ell(\theta, \xi))}{n}} \right\} + \frac{cMR}{n}\rho$$

for some $0 < c \le 30$.

Fast rates from optimal tradeoff

Theorem (Duchi & N. 2016)

Let $\rho = \log \frac{1}{\delta} + d \log n$. *Then with probability at least* $1 - \delta$*,*

$$R(\widehat{\theta}^{\text{rob}}) \le \min_{\theta \in \Theta} \left\{ R(\theta) + 2\sqrt{\frac{2\rho \mathrm{Var}(\ell(\theta, \xi))}{n}} \right\} + \frac{cMR}{n}\rho$$

for some $0 < c \le 30$*.*

Compare with the ERM: If $\mathrm{Var}(\ell(\theta; X)) \le R(\theta)$ (e.g. $\ell(\theta; X) \in [0, 1]$), then with probability $1 - \delta$,

$$R(\widehat{\theta}^{\text{erm}}) \le R(\theta^\star) + \sqrt{\frac{2\rho R(\theta^\star)}{n}} + \frac{cMR}{n}\rho$$

where $R(\theta^\star) = \inf_{\theta \in \Theta} R(\theta)$. [Vapnik & Chervonenkis 71, 74, Mammen & Tsybakov 99, Bartlett et al. 06]

Robust solution can't be too bad

Theorem (Duchi, Glynn & N. 2016)

Let $S := \operatorname{argmin}_{\theta \in \Theta} R(\theta)$.

Robust solution can't be too bad

Theorem (Duchi, Glynn & N. 2016)

Let $S := \operatorname{argmin}_{\theta \in \Theta} R(\theta)$.

- *Consistency: Under essentially same conditions as* ERM*,* $\operatorname{dist}(\widehat{\theta}^{\mathrm{rob}}, S) \stackrel{P^*}{\to} 0$

Robust solution can't be too bad

Theorem (Duchi, Glynn & N. 2016)

Let $S := \mathrm{argmin}_{\theta \in \Theta} R(\theta)$.

- *Consistency: Under essentially same conditions as* ERM*,* $\mathrm{dist}(\widehat{\theta}^{\mathrm{rob}}, S) \stackrel{P^*}{\to} 0$
- *Fast rates under growth conditions: Assume* $\ell(\cdot; X)$ *is convex,* $M(X)$*-Lipschitz with* $\mathbb{E} \exp\left(\frac{M^2(X)}{M^2}\right) \leq \exp(1)$.

  If $R(\theta) \geq \inf_{\theta^* \in \Theta} R(\theta^*) + \mathrm{dist}(\theta, S)^\gamma$ *and* $\rho = \log \frac{1}{\delta} + d \log n$*, then with probability at least* $1 - \delta$*,*

$$R(\widehat{\theta}^{\mathrm{rob}}) \leq \inf_{\theta^* \in \Theta} R(\theta^*) + c \left(\frac{\rho M^2}{\lambda^{\frac{2}{\gamma}} n} \right)^{\frac{\gamma}{2(\gamma - 1)}} .$$

Robust solution can't be too bad

Theorem (Duchi & N. 2016)

- *Efficiency loss: Define* $\theta^\star = \operatorname{argmin}_{\theta \in \Theta} R(\theta)$ *and let*

$$b(\theta^\star) := \nabla \sqrt{\mathrm{Var}(\ell(\theta^\star; X))} \quad \textit{and}$$
$$\Sigma(\theta^\star) = \left(\nabla^2 R(\theta^\star)\right)^{-1} \mathrm{Cov}(\nabla \ell(\theta^\star, \xi)) \left(\nabla^2 R(\theta^\star)\right)^{-1}.$$

If $\nabla^2 R(\theta^\star) \succ 0$*, then*

$$\sqrt{n}(\widehat{\theta}^{\mathrm{rob}} - \theta^\star) \overset{d}{\rightsquigarrow} N(-\sqrt{2\rho}\, b(\theta^\star), \Sigma(\theta^\star)).$$

Experiment 1: Amino Acid Cleavage

Problem: Amino acid strings are given, and we wish to predict whether HIV-1 will cleave in central position

Experiment 1: Amino Acid Cleavage

Problem: Amino acid strings are given, and we wish to predict whether HIV-1 will cleave in central position

- Data: pairs $x \in \mathbb{R}^d$ represents amino acid, $y \in \{-1, 1\}$ is 1 if HIV-1 cleaves

Experiment 1: Amino Acid Cleavage

Problem: Amino acid strings are given, and we wish to predict whether HIV-1 will cleave in central position

- Data: pairs $x \in \mathbb{R}^d$ represents amino acid, $y \in \{-1, 1\}$ is 1 if HIV-1 cleaves
- Use logistic loss as a convex surrogate for 0-1 error $\ell(\theta, (x, y)) = \log(1 + e^{-yx^\top \theta})$.

Experiment 1: Amino Acid Cleavage

Problem: Amino acid strings are given, and we wish to predict whether HIV-1 will cleave in central position

- Data: pairs $x \in \mathbb{R}^d$ represents amino acid, $y \in \{-1, 1\}$ is 1 if HIV-1 cleaves
- Use logistic loss as a convex surrogate for 0-1 error $\ell(\theta, (x, y)) = \log(1 + e^{-yx^\top \theta})$.
- $d = 50,960$, $n = 6590$ ($y = +1 :\ 1360$ v $y = -1 : 5230$)

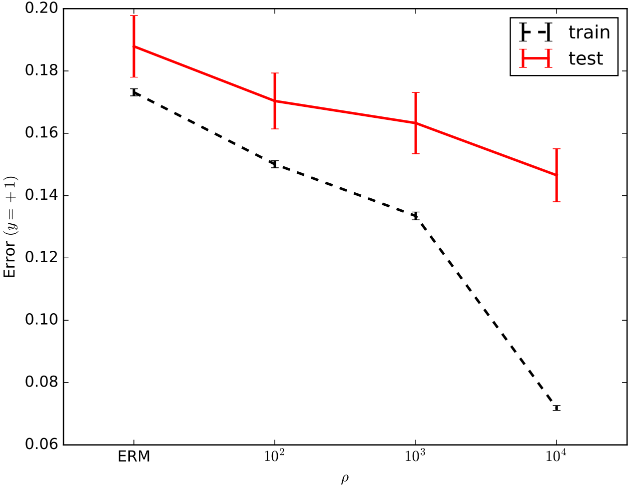
Experiment 1: Amino Acid Cleavage

Problem: Amino acid strings are given, and we wish to predict whether HIV-1 will cleave in central position

- Data: pairs $x \in \mathbb{R}^d$ represents amino acid, $y \in \{-1, 1\}$ is 1 if HIV-1 cleaves
- Use logistic loss as a convex surrogate for 0-1 error $\ell(\theta, (x, y)) = \log(1 + e^{-yx^\top \theta})$.
- $d = 50,960$, $n = 6590$ ($y = +1 :\ 1360$ v $y = -1 : 5230$)
- Subsample $9/10$ of data for training and evaluate on $1/10$, repeating 50 times for validation.

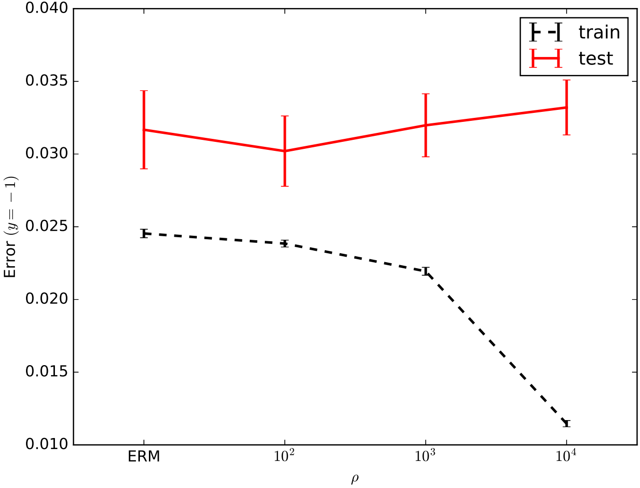
Experiment 1: Amino Acid Cleavage

Figure: Error on rare class $Y = +1$



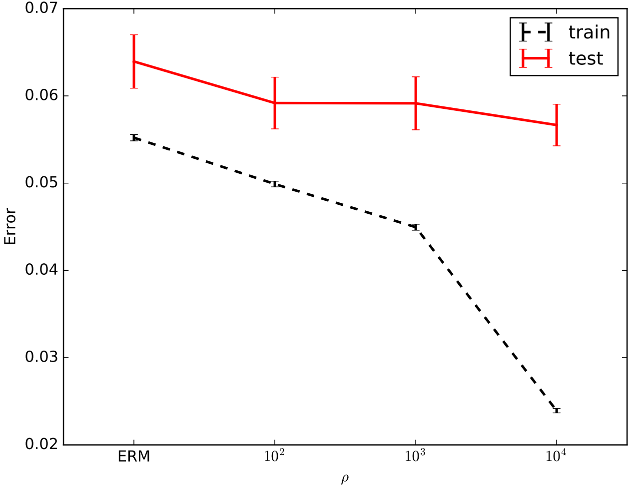
Experiment 1: Amino Acid Cleavage

Figure: Error on common class $Y = -1$



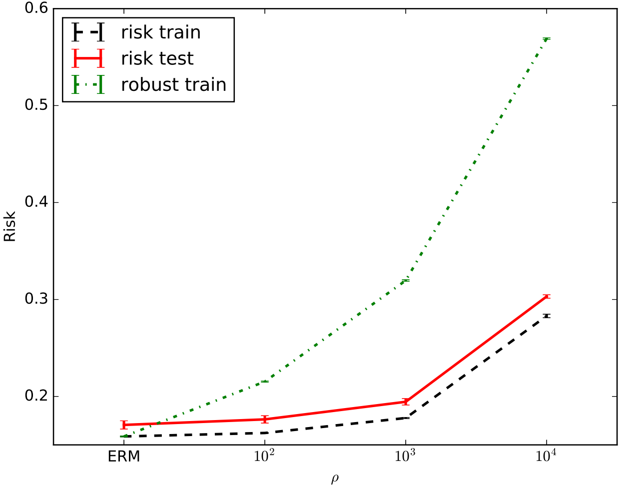
Experiment 1: Amino Acid Cleavage

Figure: Error



Experiment 1: Amino Acid Cleavage

Figure: Logistic risk and confidence bound



Experiment 1: Amino Acid Cleavage

Table: HIV-1 Cleavage Error

	risk		error (%)		$Y = +1$		$Y = -1$	
ρ	train	test	train	test	train	test	train	test
erm	0.1587	0.1706	5.52	6.39	17.32	18.79	2.45	3.17
10000	0.283	0.3031	2.39	5.67	7.18	14.65	1.15	3.32

Experiment 2: Reuters Corpus (multi-label)

Problem: Classify documents as a **subset** of the 4 categories:

$$\Big\{\text{Corporate, Economics, Government, Markets}\Big\}$$

Experiment 2: Reuters Corpus (multi-label)

Problem: Classify documents as a **subset** of the 4 categories:

$$\Big\{\text{Corporate, Economics, Government, Markets}\Big\}$$

- Data: pairs $x \in \mathbb{R}^d$ represents document, $y \in \{-1, 1\}^4$ where $y_j = 1$ indicating x belongs j-th category.

Experiment 2: Reuters Corpus (multi-label)

Problem: Classify documents as a **subset** of the 4 categories:

$$\Big\{\text{Corporate, Economics, Government, Markets}\Big\}$$

- Data: pairs $x \in \mathbb{R}^d$ represents document, $y \in \{-1, 1\}^4$ where $y_j = 1$ indicating x belongs j-th category.
- Loss $\ell(\theta_j, (x, y)) = \log(1 + e^{-yx^\top \theta_j})$ for each $j = 1, \ldots, 4$ and $\Theta = \big\{\theta \in \mathbb{R}^d : \|\theta\|_1 \leq 1000\big\}$.

Experiment 2: Reuters Corpus (multi-label)

Problem: Classify documents as a **subset** of the 4 categories:

$$\left\{ \text{Corporate, Economics, Government, Markets} \right\}$$

- Data: pairs $x \in \mathbb{R}^d$ represents document, $y \in \{-1, 1\}^4$ where $y_j = 1$ indicating x belongs j-th category.
- Loss $\ell(\theta_j, (x, y)) = \log(1 + e^{-yx^\top \theta_j})$ for each $j = 1, \ldots, 4$ and $\Theta = \left\{ \theta \in \mathbb{R}^d : \|\theta\|_1 \leq 1000 \right\}$.
- $d = 47,236$, $n = 804,414$. 10-fold cross-validation.

Experiment 2: Reuters Corpus (multi-label)

Problem: Classify documents as a **subset** of the 4 categories:

$$\left\{\text{Corporate, Economics, Government, Markets}\right\}$$

- Data: pairs $x \in \mathbb{R}^d$ represents document, $y \in \{-1, 1\}^4$ where $y_j = 1$ indicating x belongs j-th category.
- Loss $\ell(\theta_j, (x, y)) = \log(1 + e^{-yx^\top \theta_j})$ for each $j = 1, \ldots, 4$ and $\Theta = \left\{\theta \in \mathbb{R}^d : \|\theta\|_1 \leq 1000\right\}$.
- $d = 47,236$, $n = 804,414$. 10-fold cross-validation.
- Use precision and recall to evaluate performance

$$\text{Precision} = \frac{\#\ \text{Correct}}{\#\ \text{Guessed Positive}} \qquad \text{Recall} = \frac{\#\ \text{Correct}}{\#\ \text{Actually Positive}}$$

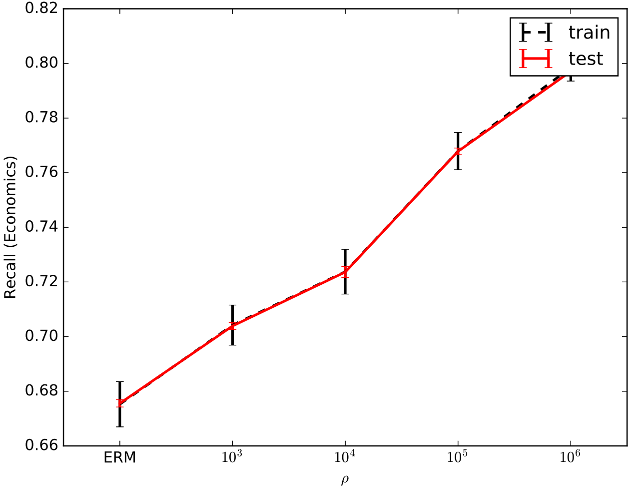
Experiment 2: Reuters Corpus (multi-label)

Table: Reuters Number of Examples

Corporate	Economics	Government	Markets
381,327	119,920	239,267	204,820

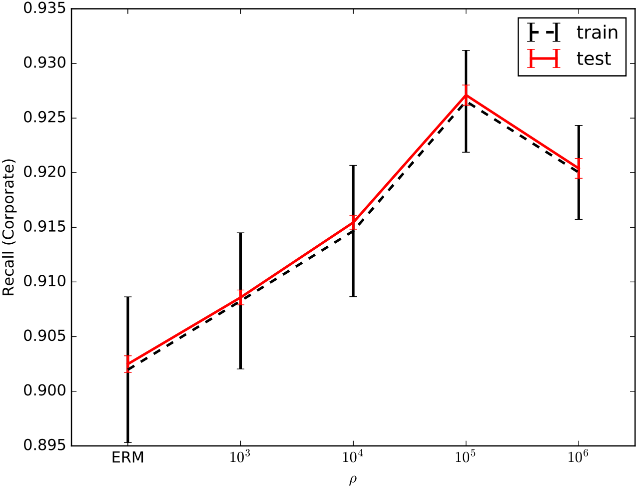
Experiment 2: Reuters Corpus (multi-label)

Figure: Recall on rare category (Economics)



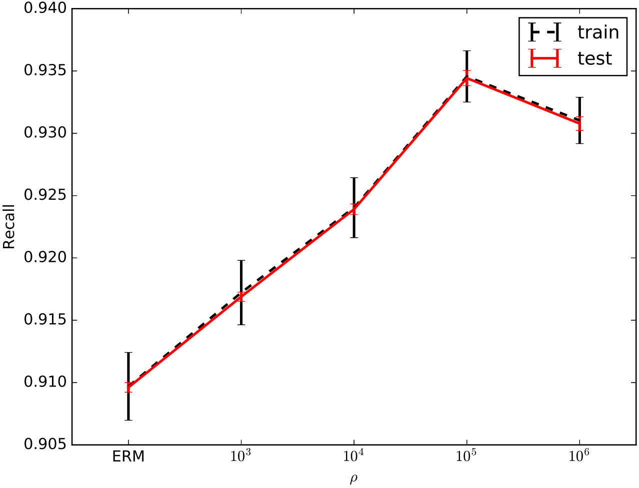
Experiment 2: Reuters Corpus (multi-label)

Figure: Recall on common category (Corporate)



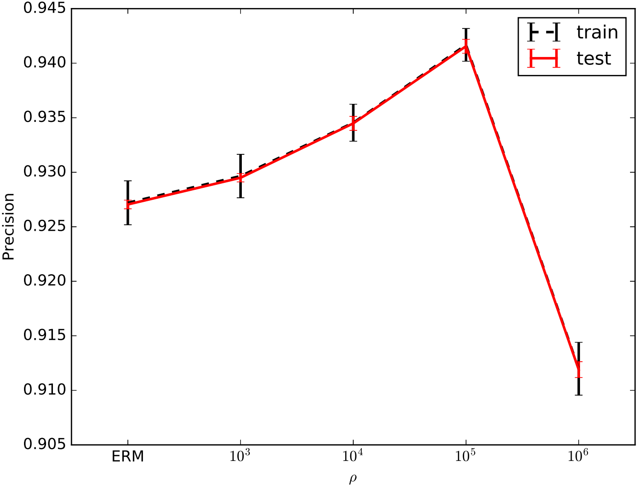
Experiment 2: Reuters Corpus (multi-label)

Figure: Recall



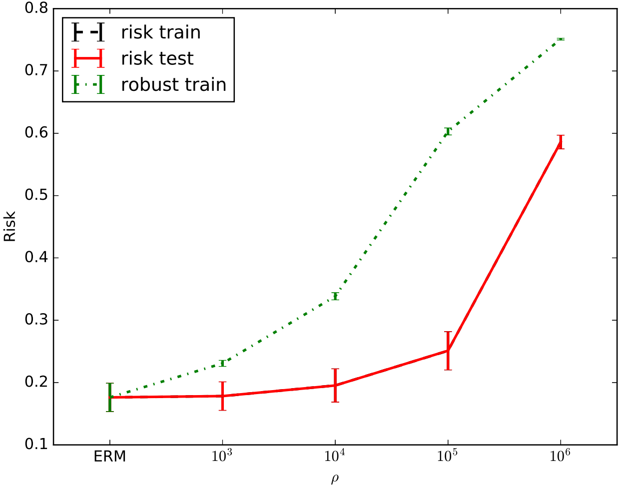
Experiment 2: Reuters Corpus (multi-label)

Figure: Precision



Experiment 3: Reuters Corpus (multi-label)

Figure: Average logistic risk and confidence bound



Experiment 2: Reuters Corpus (multi-label)

Table: Reuters Corpus (%)

	Precision		Recall		Corporate		Economics	
ρ	train	test	train	test	train	test	train	test
erm	92.72	92.7	90.97	90.96	90.2	90.25	67.53	67.56
1E5	94.17	94.16	93.46	93.44	92.65	92.71	76.79	76.78

Solving the robust optimization problem

Solve (when n and d is large)

$$\underset{\theta \in \Theta}{\text{minimize}} \quad \sup_{p \in \mathcal{P}_{n,\rho}} \sum_{i=1}^{n} p_i \ell(\theta; X_i)$$

where

$$\mathcal{P}_{n,\rho} = \left\{ p \in \mathbb{R}^n_+ \ : \ \mathbb{1}^T p = 1, \ D_f\left(p \| \mathbb{1}/n\right) \le \frac{\rho}{n} \right\}.$$

Dual reformulation

Lemma ([Ben-Tal et al. 13])

$$
\begin{aligned}
&\inf_{\theta \in \Theta} \sup_{p \in \mathcal{P}_{n,\rho}} \sum_{i=1}^{n} p_i \ell(\theta; X_i) \\
&= \inf_{\theta \in \Theta, \lambda \geq 0, \eta \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} \lambda f^* \left(\frac{\ell(\theta; X_i) - \eta}{\lambda} \right) + \frac{\rho}{n} \lambda + \eta.
\end{aligned}
$$

If $\ell(\cdot; X)$ is convex, dual problem is jointly convex in (θ, λ, η).

Approaches

$$
\begin{aligned}
&\inf_{\theta\in\Theta} \sup_{p\in\mathcal{P}_{n,\rho}} \sum_{i=1}^{n} p_i \ell(\theta; X_i) \\
&= \inf_{\theta\in\Theta,\lambda\geq 0,\eta\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^{n} \lambda f^*\left(\frac{\ell(\theta; X_i)-\eta}{\lambda}\right) + \frac{\rho}{n}\lambda + \eta.
\end{aligned}
$$

Ideas:

1. Interior point methods for the dual reformulation

Approaches

$$\begin{aligned}
&\inf_{\theta\in\Theta} \sup_{p\in\mathcal{P}_{n,\rho}} \sum_{i=1}^n p_i \ell(\theta; X_i) \\
&= \inf_{\theta\in\Theta,\lambda\geq 0,\eta\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^n \lambda f^*\left(\frac{\ell(\theta;X_i)-\eta}{\lambda}\right) + \frac{\rho}{n}\lambda + \eta.
\end{aligned}$$

Ideas:

1. Interior point methods for the dual reformulation
   $\Rightarrow$ Too slow when n large

Approaches

$$
\begin{aligned}
&\inf_{\theta \in \Theta} \sup_{p \in \mathcal{P}_{n,\rho}} \sum_{i=1}^{n} p_i \ell(\theta; X_i) \\
&= \inf_{\theta \in \Theta, \lambda \geq 0, \eta \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} \lambda f^* \left(\frac{\ell(\theta; X_i) - \eta}{\lambda} \right) + \frac{\rho}{n} \lambda + \eta.
\end{aligned}
$$

Ideas:

1. Interior point methods for the dual reformulation
   $\Rightarrow$ Too slow when n large
2. Stochastic gradient descent on the dual objective

Approaches

$$\inf_{\theta\in\Theta} \sup_{p\in\mathcal{P}_{n,\rho}} \sum_{i=1}^{n} p_i \ell(\theta; X_i)$$
$$= \inf_{\theta\in\Theta,\lambda\geq 0,\eta\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^{n} \lambda f^*\left(\frac{\ell(\theta;X_i)-\eta}{\lambda}\right) + \frac{\rho}{n}\lambda + \eta.$$

Ideas:

1. Interior point methods for the dual reformulation
   $\Rightarrow$ Too slow when n large
2. Stochastic gradient descent on the dual objective
   $\Rightarrow$ Gradient blows up as $\lambda \to 0$

Approaches

$$\begin{aligned}
&\inf_{\theta\in\Theta} \sup_{p\in\mathcal{P}_{n,\rho}} \sum_{i=1}^{n} p_i \ell(\theta; X_i) \\
&= \inf_{\theta\in\Theta, \lambda\geq 0, \eta\in\mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} \lambda f^* \left(\frac{\ell(\theta; X_i) - \eta}{\lambda} \right) + \frac{\rho}{n}\lambda + \eta.
\end{aligned}$$

Ideas:

1. Interior point methods for the dual reformulation
   $\Rightarrow$ Too slow when n large
2. Stochastic gradient descent on the dual objective
   $\Rightarrow$ Gradient blows up as $\lambda \to 0$
3. Gradient descent on primal objective

Approaches

$$
\begin{aligned}
&\inf_{\theta\in\Theta} \sup_{p\in\mathcal{P}_{n,\rho}} \sum_{i=1}^n p_i \ell(\theta; X_i) \\
&= \inf_{\theta\in\Theta, \lambda\geq 0, \eta\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^n \lambda f^*\left(\frac{\ell(\theta; X_i)-\eta}{\lambda}\right) + \frac{\rho}{n}\lambda + \eta.
\end{aligned}
$$

Ideas:

1. Interior point methods for the dual reformulation
   $\Rightarrow$ Too slow when n large
2. Stochastic gradient descent on the dual objective
   $\Rightarrow$ Gradient blows up as $\lambda \to 0$
3. Gradient descent on primal objective
   $\Rightarrow$ Still slow when n very large

Approaches

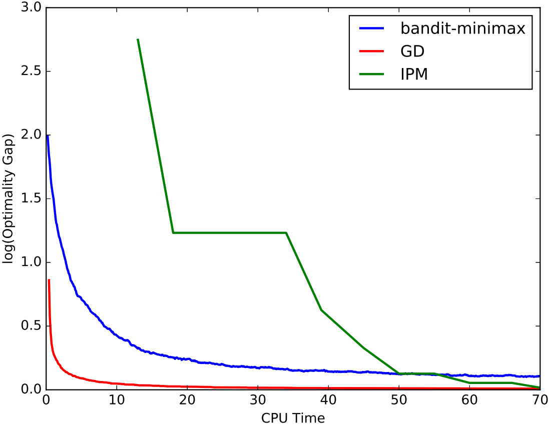
$$\inf_{\theta\in\Theta} \sup_{p\in\mathcal{P}_{n,\rho}} \sum_{i=1}^{n} p_i \ell(\theta; X_i)$$
$$= \inf_{\theta\in\Theta, \lambda\geq 0, \eta\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^{n} \lambda f^*\left(\frac{\ell(\theta; X_i) - \eta}{\lambda}\right) + \frac{\rho}{n}\lambda + \eta.$$

Ideas:

1. Interior point methods for the dual reformulation
   $\Rightarrow$ Too slow when n large
2. Stochastic gradient descent on the dual objective
   $\Rightarrow$ Gradient blows up as $\lambda \to 0$
3. Gradient descent on primal objective
   $\Rightarrow$ Still slow when n very large
4. Play a two-player stochastic game

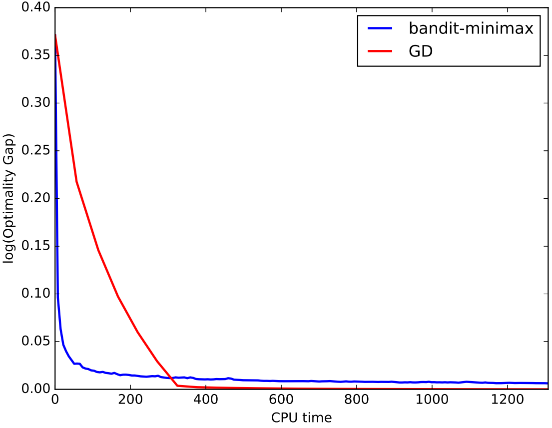
Comparison of Solvers

Figure: Small problem: $n = 2000$, $d = 500$



Comparison of Solvers

Figure: Big problem: $n = 720,000$, $d = 50,000$



Stochastic game

Ideas:

1. Play a two-player stochastic game

$$\min_{\theta} \max_{p} \sum_{i=1}^{n} p_i \ell(\theta; X_i)$$

Stochastic game

Ideas:

1. Play a two-player stochastic game (might actually work)

$$\min_{\theta} \max_{p} \sum_{i=1}^{n} p_i \ell(\theta; X_i)$$

- Player 1: Wants to minimize in $\theta \in \Theta$
- Player 2: Wants to maximize in p

Stochastic game

$$\min_{\theta} \max_{p} \sum_{i=1}^{n} p_i \ell(\theta; X_i)$$

How? Stochastic gradients for each player:

Stochastic game

$$\min_{\theta} \max_{p} \sum_{i=1}^{n} p_i \ell(\theta; X_i)$$

How? Stochastic gradients for each player:

- Player 1: For fixed $p \in \mathbb{R}_+^n$, choose index i with probability p_i and let $g^{(1)} = \nabla \ell(\theta; X_i)$. Then

$$\mathbb{E}[g^{(1)}] = \nabla_\theta \left[\sum_{i=1}^{n} p_i \ell(\theta; X_i) \right]$$

Stochastic game

$$\min_{\theta} \max_{p} \sum_{i=1}^{n} p_i \ell(\theta; X_i)$$

How? Stochastic gradients for each player:

- Player 1: For fixed $p \in \mathbb{R}_+^n$, choose index i with probability p_i and let $g^{(1)} = \nabla \ell(\theta; X_i)$. Then

$$\mathbb{E}[g^{(1)}] = \nabla_\theta \left[\sum_{i=1}^{n} p_i \ell(\theta; X_i) \right]$$

- Player 2: For fixed $\theta \in \Theta$, gradient

$$\nabla_p \sum_{i=1}^{n} p_i \ell(\theta; X_i) = [\ell(\theta; X_1) \ \ell(\theta; X_2) \ \cdots \ \ell(\theta; X_n)]^T.$$

  Choose index i with probability p_i, and let

$$g^{(2)} = \frac{1}{p_i} \ell(\theta; X_i) e_i \quad \text{so} \quad \mathbb{E}[g^{(2)}] = \nabla_p \sum_{i=1}^{n} p_i \ell(\theta; X_i)$$

Stochastic game

$$\min_{\theta} \max_{p} \sum_{i=1}^{n} p_i \ell(\theta; X_i)$$

Stochastic game: Repeat for $t = 1, 2, \ldots$

- Player 1: Choose index i with probability p_i and let $g^{(1)} = \nabla \ell(\theta; X_i)$. Update
$$\theta \leftarrow \text{Project}(\theta - \eta_1 g^{(1)}, \Theta)$$
- Player 2: Choose index i with probability p_i, let $g^{(2)} = \frac{1}{p_i} \ell(\theta; X_i) e_i$, and update
$$p \leftarrow \text{Project}(p + \eta_2 g^{(2)}, \mathcal{P}_{n,\rho}).$$

Stochastic game

$$\min_{\theta} \max_{p} \sum_{i=1}^{n} p_i \ell(\theta; X_i)$$

Result: After T steps of method, with probability $\geq 1 - \delta$, have near saddle pair $\widehat{\theta}_T$ and $\widehat{p}_T$ such that

$$\begin{aligned} -\frac{C\sqrt{\log \frac{1}{\delta}}}{\sqrt{T}} + \sup_{p \in \mathcal{P}_{n,\rho}} \sum_{i=1}^{n} p_i \ell(\widehat{\theta}_T; X_i) &\leq \sum_{i=1}^{n} \widehat{p}_{T,i} \ell(\widehat{\theta}_T; X_i) \\ &\leq \inf_{\theta \in \Theta} \sum_{i=1}^{n} \widehat{p}_{T,i} \ell(\theta; X_i) + \frac{C\sqrt{\log \frac{1}{\delta}}}{\sqrt{T}} \end{aligned}$$

where C is independent of n and dimension d.

Stochastic game

$$\min_{\theta} \max_{p} \sum_{i=1}^{n} p_i \ell(\theta; X_i)$$

Stochastic game: Repeat for $t = 1, 2, \ldots$

- Player 1: Choose index i with probability p_i and let $g^{(1)} = \nabla \ell(\theta; X_i)$. Update
$$\theta \leftarrow \text{Project}(\theta - \eta_1 g^{(1)}, \Theta)$$
- Player 2: Choose index i with probability p_i, let $g^{(2)} = \frac{1}{p_i} \ell(\theta; X_i) e_i$, and update
$$p \leftarrow \text{Project}(p + \eta_2 g^{(2)}, \mathcal{P}_{n,\rho}).$$

Stochastic game

$$\min_{\theta} \max_{p} \sum_{i=1}^{n} p_i \ell(\theta; X_i)$$

Stochastic game: Repeat for $t = 1, 2, \ldots$

- Player 1: Choose index i with probability p_i and let $g^{(1)} = \nabla \ell(\theta; X_i)$. Update

  $$\theta \leftarrow \text{Project}(\theta - \eta_1 g^{(1)}, \Theta)$$

  Takes time $O(\text{Time}_{\text{Update}})$
- Player 2: Choose index i with probability p_i, let $g^{(2)} = \frac{1}{p_i} \ell(\theta; X_i) e_i$, and update

  $$p \leftarrow \text{Project}(p + \eta_2 g^{(2)}, \mathcal{P}_{n,\rho}).$$

Stochastic game

$$\min_{\theta} \max_{p} \sum_{i=1}^{n} p_i \ell(\theta; X_i)$$

Stochastic game: Repeat for $t = 1, 2, \ldots$

- Player 1: Choose index i with probability p_i and let $g^{(1)} = \nabla \ell(\theta; X_i)$. Update

  $$\theta \leftarrow \text{Project}(\theta - \eta_1 g^{(1)}, \Theta)$$

  Takes time $O(\text{Time}_{\text{Update}})$
- Player 2: Choose index i with probability p_i, let $g^{(2)} = \frac{1}{p_i} \ell(\theta; X_i) e_i$, and update

  $$p \leftarrow \text{Project}(p + \eta_2 g^{(2)}, \mathcal{P}_{n,\rho}).$$

  For special sets $\mathcal{P}_{n,\rho}$ and careful algorithm, takes time $O(\log n)$

Stochastic game

$$\min_{\theta} \max_{p} \sum_{i=1}^{n} p_i \ell(\theta; X_i)$$

Stochastic game: Repeat for $t = 1, 2, \ldots$

- Player 1: Choose index i with probability p_i and let $g^{(1)} = \nabla \ell(\theta; X_i)$. Update

  $$\theta \leftarrow \mathsf{Project}(\theta - \eta_1 g^{(1)}, \Theta)$$

  Takes time $O(\mathsf{Time}_{\text{Update}})$

- Player 2: Choose index i with probability p_i, let $g^{(2)} = \frac{1}{p_i} \ell(\theta; X_i) e_i$, and update

  $$p \leftarrow \mathsf{Project}(p + \eta_2 g^{(2)}, \mathcal{P}_{n,\rho}).$$

  For special sets $\mathcal{P}_{n,\rho}$ and careful algorithm, takes time $O(\log n)$

Total time: For ϵ-solution, takes time

$$\frac{\rho \log n}{\epsilon^2} + \frac{\mathsf{Time}_{\text{Update}}}{\epsilon^2}$$

Reuters: Comparison to Gradient Descent

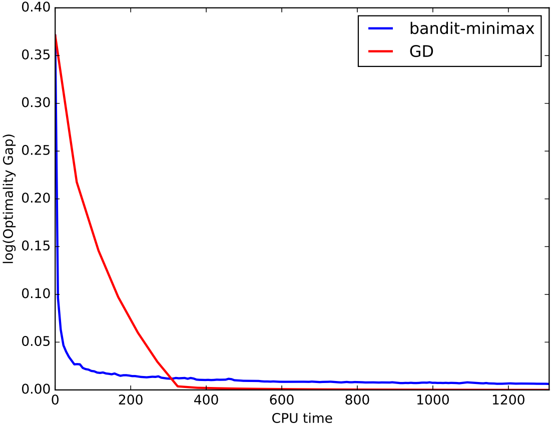
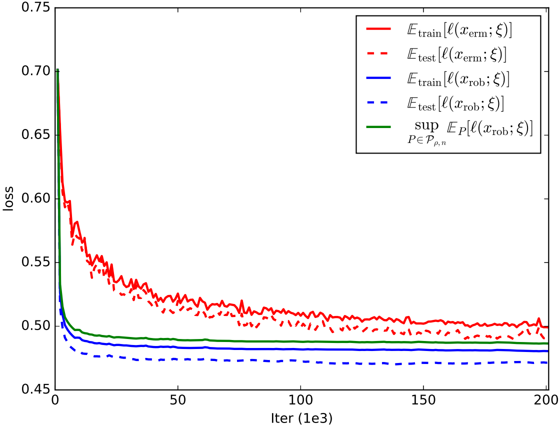


Figure: Log Optimality Ratio ($n = 720,000$, $d = 50,000$)

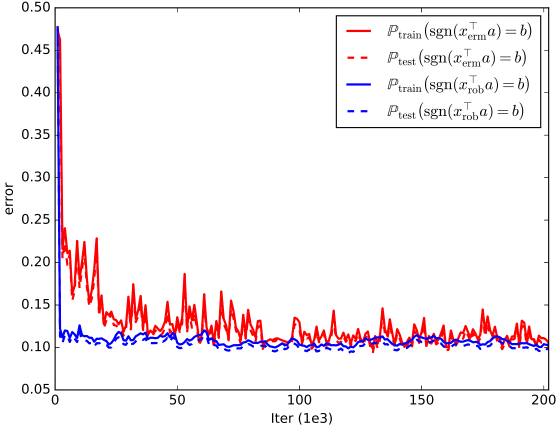
Reuters: Comparison to SGD on ERM

Figure: Logistic Objective ($n = 720,000$, $d = 50,000$)



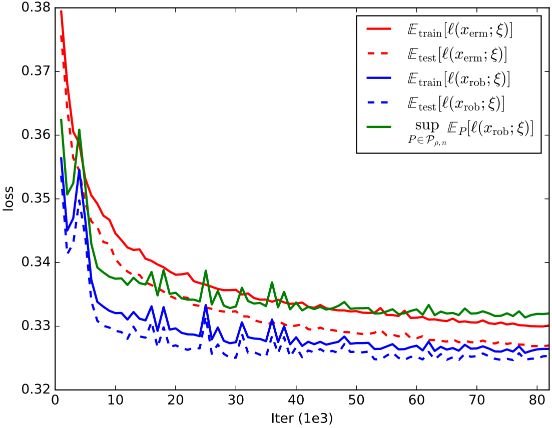
Reuters: Comparison to SGD on ERM

Figure: Classification Error ($n = 720,000$, $d = 50,000$)



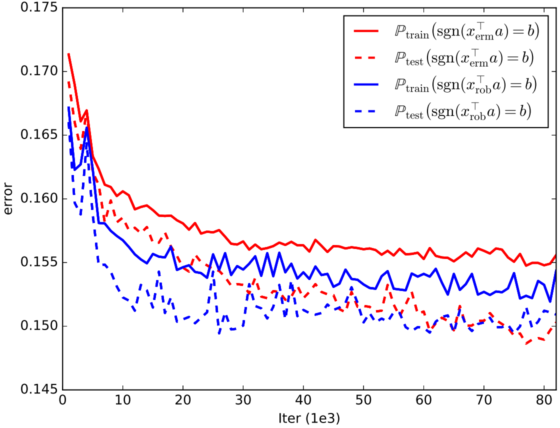
Adult: Comparison to SGD on ERM

Figure: Hinge Objective ($n = 30,000$, $d = 123$)



Adult: Comparison to SGD on ERM

Figure: Classification Error ($n = 30,000$, $d = 123$)



Summary

Statistical theory for robust optimization

1. **Convex procedure** for variance regularization
2. Guarantees of generalizability and **optimal trading off of bias v variance**
3. Comes with **statistical optimality certificates**
4. Efficient solution method as fast as **SGD**

Empirical likelihood main

The *empirical likelihood confidence region* is

Empirical likelihood main

The *empirical likelihood confidence region* is

$$E_n(\rho) := \left\{ \sum_{i=1}^{n} p_i Z_i : D_{\chi^2}\left(p \| \mathbb{1}/n\right) \le \frac{\rho}{n} \right\}.$$

[Owen 90, Baggerly 98, Newey and Smith 01, Imbens 02]

Empirical likelihood main

The *empirical likelihood confidence region* is

$$\begin{aligned} E_n(\rho) &:= \left\{ \sum_{i=1}^n p_i Z_i : D_{\chi^2}\left(p \| \mathbb{1}/n\right) \leq \frac{\rho}{n} \right\} \\ &= \left\{ \sum_{i=1}^n p_i Z_i : \frac{1}{n} \sum_{i=1}^n (np_i - 1)^2 \leq \frac{\rho}{n}, p^\top \mathbb{1} = 1, p \geq 0 \right\} \\ &= \frac{1}{n} \sum_{i=1}^n Z_i + \left\{ \sum_{i=1}^n u_i Z_i : \|u\|_2^2 \leq \frac{\rho}{n^2}, u^\top \mathbb{1} = 0, u \geq -\frac{\mathbb{1}}{n} \right\} \end{aligned}$$

by letting $u_i = p_i - \frac{1}{n}$.

Empirical likelihood main

The *empirical likelihood confidence region* is

$$
\begin{aligned}
E_n(\rho) &:= \left\{ \sum_{i=1}^n p_i Z_i : D_{\chi^2}\left(p \| \mathbb{1}/n\right) \le \frac{\rho}{n} \right\} \\
&= \left\{ \sum_{i=1}^n p_i Z_i : \frac{1}{n} \sum_{i=1}^n (np_i - 1)^2 \le \frac{\rho}{n}, p^\top \mathbb{1} = 1, p \ge 0 \right\} \\
&= \frac{1}{n} \sum_{i=1}^n Z_i + \left\{ \underbrace{\sum_{i=1}^n u_i Z_i}_{\text{Ellipse from } \textit{data}} : \|u\|_2^2 \le \frac{\rho}{n^2}, u^\top \mathbb{1} = 0, u \ge -\frac{\mathbb{1}}{n} \right\}
\end{aligned}
$$

by letting $u_i = p_i - \frac{1}{n}$.

Robust Optimization ≈ Variance Regularization main

Proof Sketch Let $z_i = \ell(\theta; X_i)$, $u_i = p_i - \frac{1}{n}$, and denote by $\bar{z}$ and s_n^2 the sample mean and variance respectively.

Robust Optimization $\approx$ Variance Regularization main

Proof Sketch Let $z_i = \ell(\theta; X_i)$, $u_i = p_i - \frac{1}{n}$, and denote by $\bar{z}$ and s_n^2 the sample mean and variance respectively.

$$R_n(\theta; \mathcal{P}_{n,\rho}) = \max_p \left\{ \langle p, z \rangle : D_{\chi^2}\left(p \| \mathbb{1}/n\right) \leq \frac{\rho}{n} \right\}$$

Robust Optimization $\approx$ Variance Regularization main

Proof Sketch Let $z_i = \ell(\theta; X_i)$, $u_i = p_i - \frac{1}{n}$, and denote by $\bar{z}$ and s_n^2 the sample mean and variance respectively.

$$R_n(\theta; \mathcal{P}_{n,\rho}) = \max_p \left\{ \langle p, z \rangle : \frac{1}{n} \sum_{i=1}^{n} (np_i - 1)^2 \leq \frac{\rho}{n}, p^\top \mathbb{1} = 1, p \geq 0 \right\}$$

Robust Optimization ≈ Variance Regularization main

Proof Sketch Let $z_i = \ell(\theta; X_i)$, $u_i = p_i - \frac{1}{n}$, and denote by $\bar{z}$ and s_n^2 the sample mean and variance respectively.

$$
\begin{aligned}
R_n(\theta; \mathcal{P}_{n,\rho}) &= \max_p \left\{ \langle p, z \rangle : \frac{1}{n} \sum_{i=1}^{n} (np_i - 1)^2 \le \frac{\rho}{n}, p^\top \mathbb{1} = 1, p \ge 0 \right\} \\
&= \bar{z} + \max_u \left\{ \langle u, z - \bar{z} \rangle : \|u\|_2^2 \le \frac{\rho}{n^2}, u^\top \mathbb{1} = 0, u \ge -\frac{\mathbb{1}}{n} \right\}
\end{aligned}
$$

Robust Optimization ≈ Variance Regularization

main

Proof Sketch Let $z_i = \ell(\theta; X_i)$, $u_i = p_i - \frac{1}{n}$, and denote by $\bar{z}$ and s_n^2 the sample mean and variance respectively.

$$\begin{aligned} R_n(\theta; \mathcal{P}_{n,\rho}) &= \max_p \left\{ \langle p, z \rangle : \frac{1}{n} \sum_{i=1}^n (np_i - 1)^2 \le \frac{\rho}{n}, p^\top \mathbb{1} = 1, p \ge 0 \right\} \\ &= \bar{z} + \max_u \left\{ \langle u, z - \bar{z} \rangle : \|u\|_2^2 \le \frac{\rho}{n^2}, u^\top \mathbb{1} = 0, u \ge -\frac{\mathbb{1}}{n} \right\} \\ &\le \bar{z} + \frac{\sqrt{2\rho}}{n} \|z - \bar{z}\|_2 = \bar{z} + \sqrt{\frac{2\rho}{n} s_n^2} \quad \text{by Cauchy-Schwarz} \end{aligned}$$

Robust Optimization $\approx$ Variance Regularization main

Proof Sketch Let $z_i = \ell(\theta; X_i)$, $u_i = p_i - \frac{1}{n}$, and denote by $\bar{z}$ and s_n^2 the sample mean and variance respectively.

$$
\begin{aligned}
R_n(\theta; \mathcal{P}_{n,\rho}) &= \max p \left\{ \langle p, z \rangle : \frac{1}{n} \sum_{i=1}^{n} (np_i - 1)^2 \leq \frac{\rho}{n}, p^\top \mathbb{1} = 1, p \geq 0 \right\} \\
&= \bar{z} + \max_u \left\{ \langle u, z - \bar{z} \rangle : \|u\|_2^2 \leq \frac{\rho}{n^2}, u^\top \mathbb{1} = 0, u \geq -\frac{\mathbb{1}}{n} \right\} \\
&\leq \bar{z} + \frac{\sqrt{2\rho}}{n} \|z - \bar{z}\|_2 = \bar{z} + \sqrt{\frac{2\rho}{n} s_n^2} \quad \text{by Cauchy-Schwartz}
\end{aligned}
$$

Last inequality is tight if for all i

$$
u_i = \frac{1}{n} \sqrt{\frac{2\rho}{n s_n^2}} (z_i - \bar{z}) \geq -\frac{1}{n}
$$

Extensions and issues main

Issue: What if $\theta^\star \in \mathbb{R}^d$ is not unique?

Extensions and issues main

Issue: What if $\theta^\star \in \mathbb{R}^d$ is not unique?
Let $S = \operatorname{argmin}_{\theta \in \Theta} R(\theta)$ and

$$\boldsymbol{r}^\star = \min_{\theta^\star \in S} \max_{\theta \in S} \|\theta - \theta^\star\|_2$$

Then [Duchi, Glynn & N. 16]

$$\begin{aligned}
&\mathbb{P}\left(\inf_{\theta \in \Theta} R(\theta) \le R_n(\widehat{\theta}^{\mathrm{rob}}, \mathcal{P}_{n,\rho})\right) \\
&\quad \ge \mathbb{P}\left(N(0,1) + \sqrt{\rho} \ge \boldsymbol{r}^\star \sqrt{\rho \mathrm{Var}(\ell(x^\star;\xi))(d+1)}\right) + O(n^{-\frac{1}{2}}).
\end{aligned}$$

Extensions and issues main

Issue: What if $\theta^\star \in \mathbb{R}^d$ is not unique?
Let $S = \mathrm{argmin}_{\theta \in \Theta} R(\theta)$ and

$$\boldsymbol{r}^\star = \min_{\theta^\star \in S} \max_{\theta \in S} \|\theta - \theta^\star\|_2$$

Then [Duchi, Glynn & N. 16]

$$\mathbb{P}\left(\inf_{\theta \in \Theta} R(\theta) \le R_n(\widehat{\theta}^{\mathrm{rob}}, \mathcal{P}_{n,\rho})\right) \\ \ge \mathbb{P}\left(N(0,1) + \sqrt{\rho} \ge \boldsymbol{r}^\star \sqrt{\rho \mathrm{Var}(\ell(x^\star;\xi))(d+1)}\right) + O(n^{-\frac{1}{2}}).$$

- If $\boldsymbol{r}^\star$ large, then lose confidence, if $\boldsymbol{r}^\star$ small, good shape