Variance Regularization with Convex Objectives

Hongseok Namkoong, John Duchi

Stanford University

December 2017

Curly fries $\propto$ Intelligence?

WIRED Technology | Science | Culture | Video | Reviews | Magazine

Liking curly fries on Facebook reveals your high IQ

By **PHILIPPA WARR**
12 Mar 2013

What you Like on Facebook could reveal your race, age, IQ, sexuality and other personal data, even if you've set that information to "private".

Curly fries $\propto$ Intelligence?

WIRED | Technology | Science | Culture | Video | Reviews | Magazine

Liking curly fries on Facebook reveals your high IQ

By **PHILIPPA WARR**
12 Mar 2013

What you Like on Facebook could reveal your race, age, IQ, sexuality and other personal data, even if you've set that information to "private".

Unlikely to be robust to even small changes in the underlying data

Stochastic optimization problems

Data $X_1, \ldots, X_n$ and parameters θ to learn, with loss

$$\ell(\theta, X)$$

Stochastic optimization problems

Data $X_1, \ldots, X_n$ and parameters θ to learn, with loss

$$\ell(\theta, X)$$

We want to solve the population (risk) problem

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

Stochastic optimization problems

Data $X_1, \ldots, X_n$ and parameters θ to learn, with loss

$$\ell(\theta, X)$$

We want to solve the population (risk) problem

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

- Loss $\ell(\theta; X)$, Data/randomness X, Parameters $\theta \in \Theta$
- P_0 often unknown

Stochastic optimization problems

Data $X_1, \ldots, X_n$ and parameters θ to learn, with loss

$$\ell(\theta, X)$$

We want to solve the population (risk) problem

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

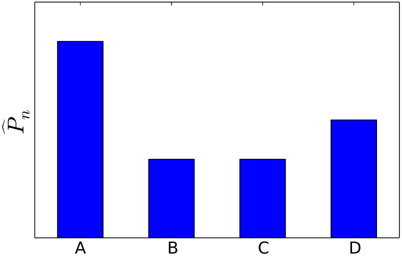
- Loss $\ell(\theta; X)$, Data/randomness X, Parameters $\theta \in \Theta$
- P_0 often unknown

Empirical risk minimization:

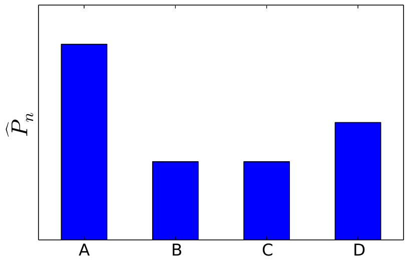
$$\widehat{\theta}^{\mathrm{erm}} = \operatorname*{argmin}_{\theta \in \Theta} \mathbb{E}_{\widehat{P}_n}[\ell(\theta; X)] = \frac{1}{n} \sum_{i=1}^{n} \ell(\theta; X_i)$$

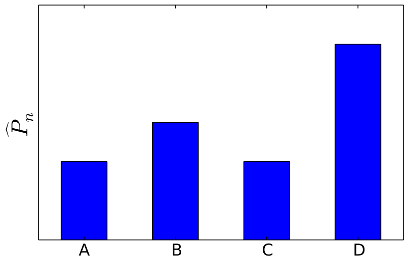
Typical learning problems

Typical learning problems

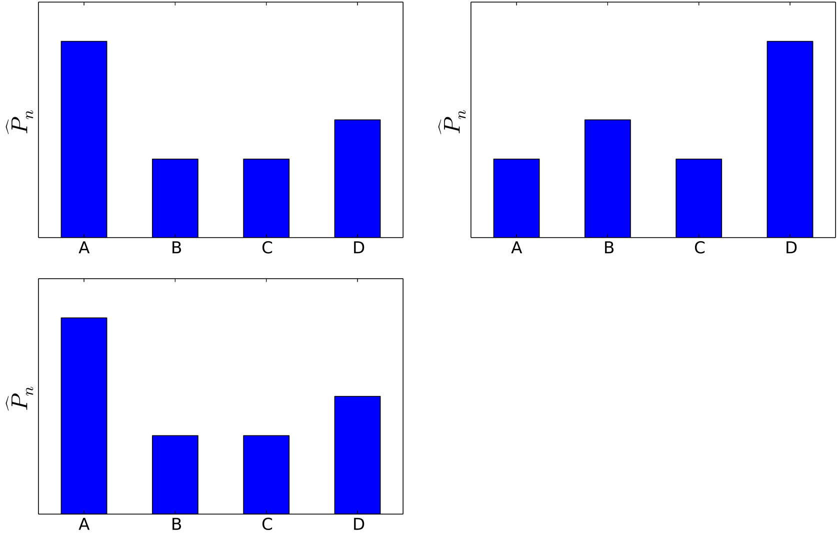


Typical learning problems

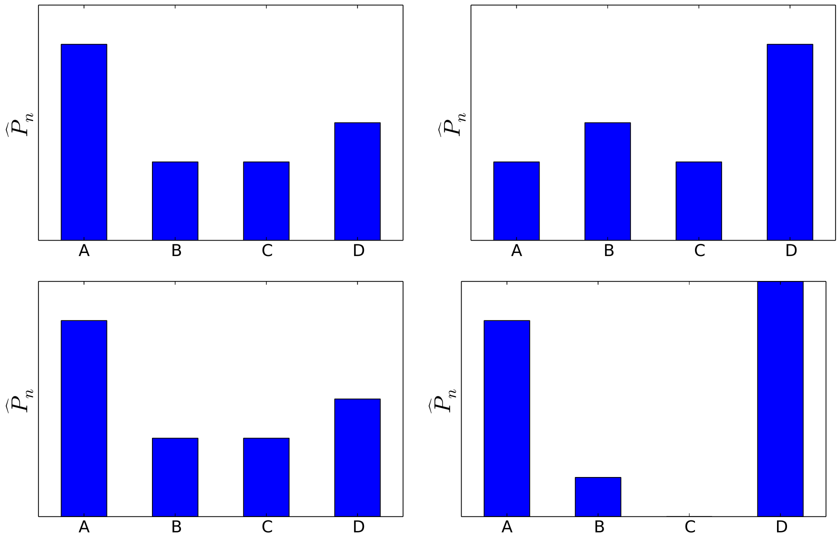




Typical learning problems

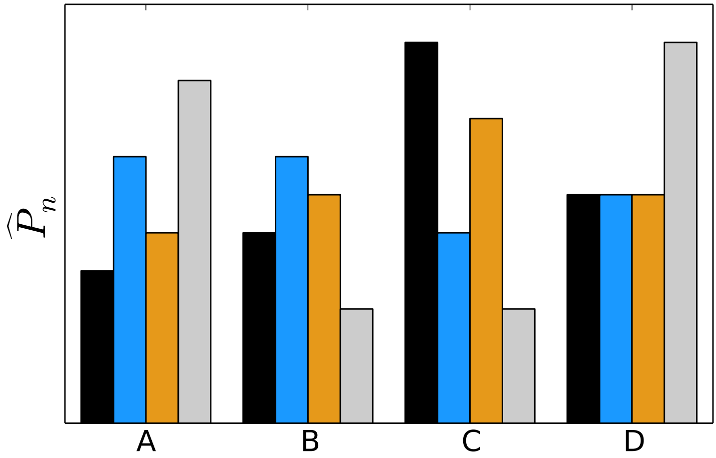


Typical learning problems



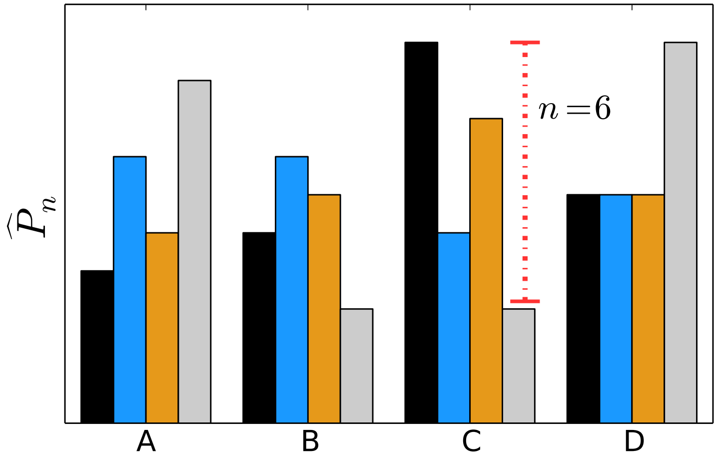
Typical learning problems

- Want to be robust to small perturbations in $\widehat{P}_n$



Typical learning problems

- Want to be robust to small perturbations in $\widehat{P}_n$



Distributionally Robust Optimization

Goal:

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

Distributionally Robust Optimization

Goal:

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

Solve empirical risk minimization problem

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

Distributionally Robust Optimization

Goal:

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

Solve empirical risk minimization problem

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

Distributionally Robust Optimization

Goal:

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

Instead, solve *distributionally robust optimization problem*

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

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

Distributionally Robust Optimization

Goal:

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

Instead, solve *distributionally robust optimization problem*

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

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

Do well **almost all the time** instead of on average!

Today: Statistically principled choice of $\mathcal{P}_{n,\rho} \Rightarrow$ optimality certificates

Generalized Empirical likelihood

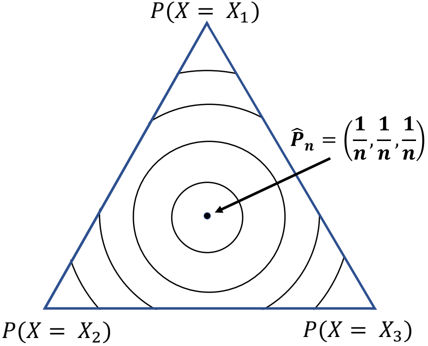
Idea: Instead of using empirical distribution $\widehat{P}_n$ on sample $X_1, \ldots, X_n$, look at all distributions "near" it.

Measures of closeness we use:
Chi-square divergence

$$D_{\chi^2}(P\|Q) = \frac{1}{2} \sum_{x:q(x)>0} \frac{(p(x) - q(x))^2}{q(x)}$$

Worst-case region:

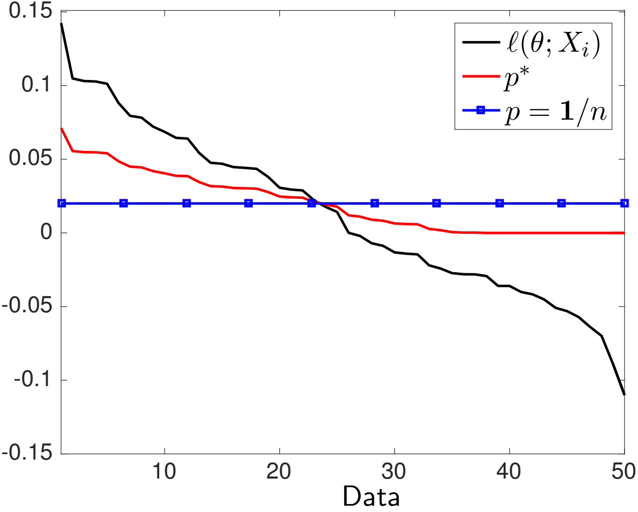
$$\mathcal{P}_{n,\rho} := \left\{ P : D_{\chi^2}\left(P\|\widehat{P}_n\right) \le \frac{\rho}{n} \right\}$$



[Ben-Tal et al. 13, Bertsimas et al. 16, Lam & Zhou 17, Duchi, Glynn & N. 17, Lam17]

Upweighting Harder Examples

$$p^* = \mathrm{argmax}_{p \in \mathcal{P}_{n,\rho}} \sum_{i=1}^{n} p_i \ell(\theta; X_i)$$



Robust Optimization

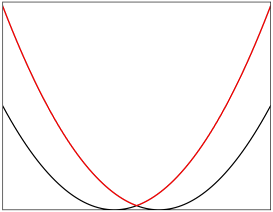
$$\widehat{\theta}^{\mathrm{rob}} := \operatorname*{argmin}_{\theta \in \Theta} \max_{P : D_{\chi^2}\left(P \| \widehat{P}_n\right) \le \frac{\rho}{n}} \mathbb{E}_P[\ell(\theta; X)].$$

Robust Optimization

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

Nice properties:

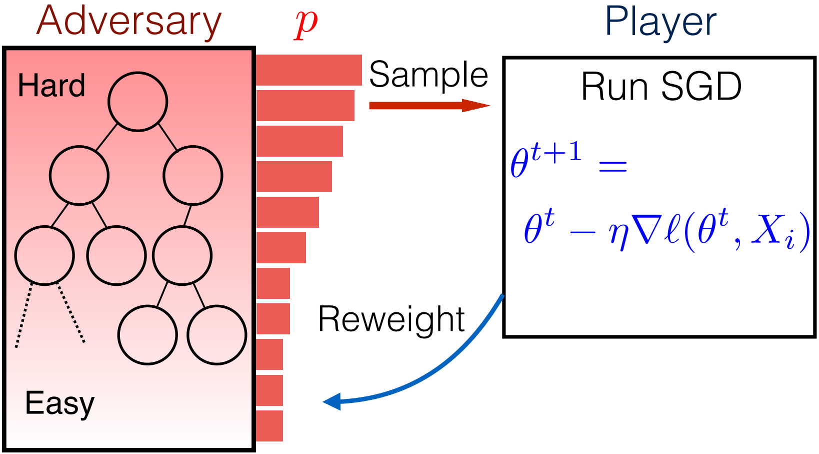
- Convex optimization problem = Computationally efficient
- Conic forms [Ben-Tal et al. 13]
- Efficient solution methods as fast as stochastic gradient descent [N. & Duchi, 16]



Algorithm

Play a **two-player stochastic game** [N. & Duchi 16]

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



Understanding Performance: bias/variance tradeoff

Understanding Performance: bias/variance tradeoff

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

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

- Can be made uniform in $\theta \in \Theta$

Understanding Performance: bias/variance tradeoff

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

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

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

  Trade off bias-variance optimally by solving

$$\widehat{\theta}^{\mathrm{var}} \in \operatorname*{argmin}_{\theta \in \Theta} \left\{ \underbrace{\mathbb{E}_{\widehat{P}_n}[\ell(\theta; X)]}_{\text{bias}} + \underbrace{\sqrt{\frac{C\mathrm{Var}_{\widehat{P}_n}(\ell(\theta; X))}{n}}}_{\text{variance}} \right\}.$$

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 non-convex

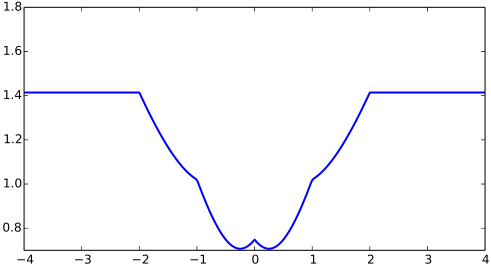


Figure: Variance of $|\theta - X|$

Robust Optimization $\approx$ Variance Regularization

Theorem (N. & Duchi 2017)

Assume that $|\ell(\theta; X)| \le M$. *With prob at least* $1 - \exp(-\frac{n\mathrm{Var}(\ell(\theta;X))}{36M^2})$

$$\underbrace{\max_{P : D_{\chi^2}(P \| \widehat{P}_n) \le \frac{\rho}{n}} \mathbb{E}_P[\ell(\theta; X)]}_{\text{Robust}} = \underbrace{\mathbb{E}_{\widehat{P}_n}[\ell(\theta; X)] + \sqrt{\frac{2\rho \mathrm{Var}_{\widehat{P}_n}(\ell(\theta; X))}{n}}}_{\text{Bias+Variance}}$$

- Can be made uniform over $\theta \in \Theta$
- Robust is convex, $\text{Bias} + \text{Variance}$ is (generally) non-convex

Optimal bias variance tradeoff

Let $\mathfrak{Comp}_n(\Theta)$ denote complexity of $\{\ell(\theta; \cdot) : \theta \in \Theta\}$ and let

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

Optimal bias variance tradeoff

Let $\mathfrak{Comp}_n(\Theta)$ denote complexity of $\{\ell(\theta;\cdot) : \theta \in \Theta\}$ and let

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

Theorem (N. & Duchi 2017)

Let $\rho = \log\frac{1}{\delta} + \mathfrak{Comp}_n(\Theta)$. *If* $\ell(\theta;X) \in [0, M]$, *then with prob* $1-\delta$,

$$R(\widehat{\theta}^{\text{rob}}) = \mathbb{E}[\ell(\widehat{\theta}^{\text{rob}};X)] \le \underbrace{\min_{\theta\in\Theta}\left\{R(\theta) + 2\sqrt{\frac{2\rho\mathrm{Var}(\ell(\theta,\xi))}{n}}\right\}}_{\text{optimal tradeoff}} + \frac{CM\rho}{n}$$

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

Similar result holds with **localized Rademacher complexities**.

Fast rates from optimal tradeoff

- Let $\rho \approx \mathfrak{Comp}_n(\Theta)$. If $\ell(\theta; X) \in [0, M]$, then with high prob,

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

Fast rates from optimal tradeoff

- Let $\rho \approx \mathfrak{Comp}_n(\Theta)$. If $\ell(\theta; X) \in [0, M]$, then with high prob,

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

- ERM: For $R(\theta^\star) = \inf_{\theta \in \Theta} R(\theta)$, with high probability,

$$R(\widehat{\theta}^{\mathrm{erm}}) \le R(\theta^\star) + \sqrt{\frac{2\rho M R(\theta^\star)}{n}} + \frac{CM\rho}{n}$$

Fast rates from optimal tradeoff

- Let $\rho \approx \mathfrak{Comp}_n(\Theta)$. If $\ell(\theta; X) \in [0, M]$, then with high prob,

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

- ERM: For $R(\theta^\star) = \inf_{\theta \in \Theta} R(\theta)$, with high probability,

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

- If $\mathrm{Var}(\ell(\theta^\star; X)) \ll MR(\theta^\star)$, first bound is **tighter**
  - See paper for an **explicit example** where

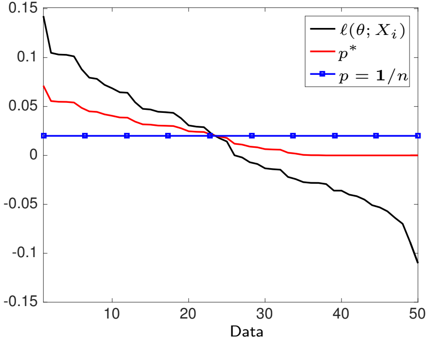
$$R(\widehat{\theta}^{\text{rob}}) \le R(\theta^\star) + \frac{C_1}{n} \quad \text{but} \quad R(\widehat{\theta}^{\text{erm}}) \ge R(\theta^\star) + \frac{C_2}{\sqrt{n}}$$

Experiments

Upweighting Harder Examples

$$\underset{\theta\in\Theta}{\text{minimize}} \max_{P: D_{\chi^2}(P \| \widehat{P}_n) \le \frac{\rho}{n}} \mathbb{E}_P[\ell(\theta; X)].$$

- Upweights **hard (high loss)** examples when learning
- Often, **rare** examples are **hard**
- Expect improvements on **rare** and **hard** examples



Experiment: 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.
- Logistic loss, with $\Theta = \left\{\theta \in \mathbb{R}^d : \|\theta\|_1 \leq 1000\right\}$
- $d = 47,236$, $n = 804,414$. 10-fold cross-validation.
- Use precision and recall to evaluate performance

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

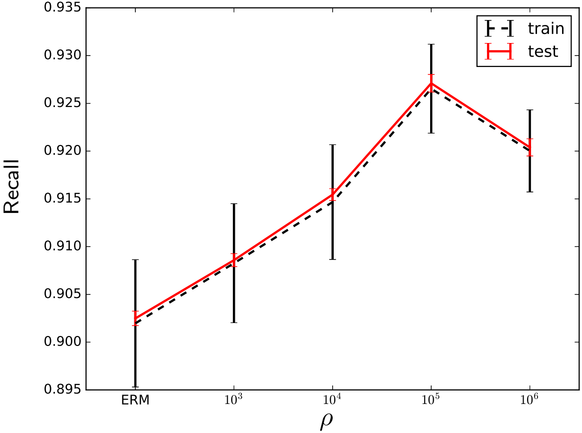
Experiment: Reuters Corpus (multi-label)

Table: Reuters Number of Examples

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

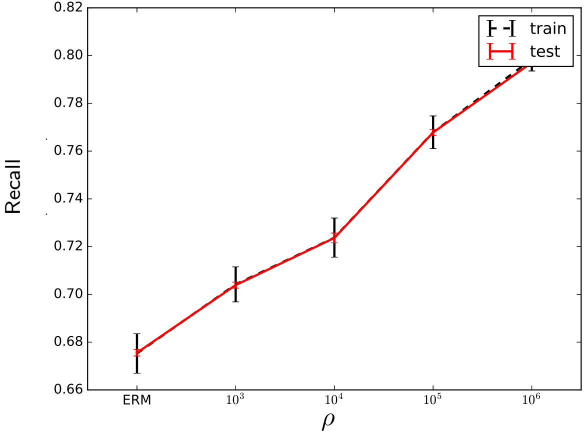
Experiment: Reuters Corpus (multi-label)

Figure: Recall on common category (Corporate)



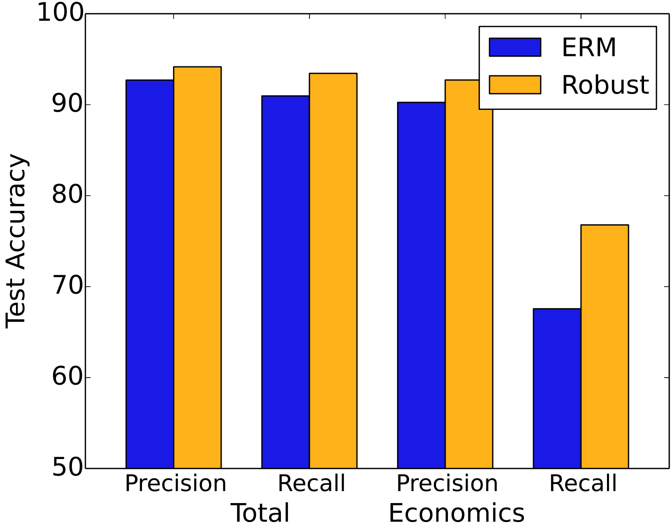
Experiment: Reuters Corpus (multi-label)

Figure: Recall on rare category (Economics)



Experiment: Reuters Corpus (multi-label)

Do well **almost all** the time intead of just on average!



Summary

Statistical theory for robust optimization

1. **Convex procedure** for variance regularization
2. Generalization guarantees for **optimal tradeoff between bias vs variance**
3. Improves performance on **hard instances** empirically

Summary

Statistical theory for robust optimization

1. **Convex procedure** for variance regularization
2. Generalization guarantees for **optimal tradeoff between bias vs variance**
3. Improves performance on **hard instances** empirically

Thanks! (Poster 212)

Long version arXiv:1610.02581