Variance-based Regularization with Convex Objectives

Hongseok Namkoong, John Duchi
(NIPS 2017 Best Paper)

马佳明 SA23229004
中国科学技术大学 人工智能与数据科学学院
2024.6.13

中国科学技术大学
University of Science and Technology of China

Content

- Intuition motivation
- Optimization theoretical analysis
- Experiments
- Summary

1. Intuition motivation

Stochastic optimization problem

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

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

We want to solve the population (risk) problem

$$\min R(\theta) := \mathbb{E}_{P_0}[\ell(\theta; X)] \ ,$$

$$\text{s.t.}\ \theta \in \Theta.$$

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

Empirical Risk Optimization

Goal:

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

Empirical risk minimization:

$$\hat{\theta}^{\text{erm}} = \underset{\theta \in \Theta}{\text{argmin}}\, \mathbb{E}_{\hat{P}_n}[\ell(\theta; X)] = \frac{1}{n}\sum_{i=1}^{n} \ell(\theta; X_i),$$

Solve empirical risk minimization problem

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

Bias & variance tradeoff

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

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

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

Bias & variance tradeoff

- From empirical Bernstein's inequality, with probability $1 - \delta$,

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

- **Variance Regularization:** trade off bias-variance optimally by solving

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

Optimizing for bias & variance

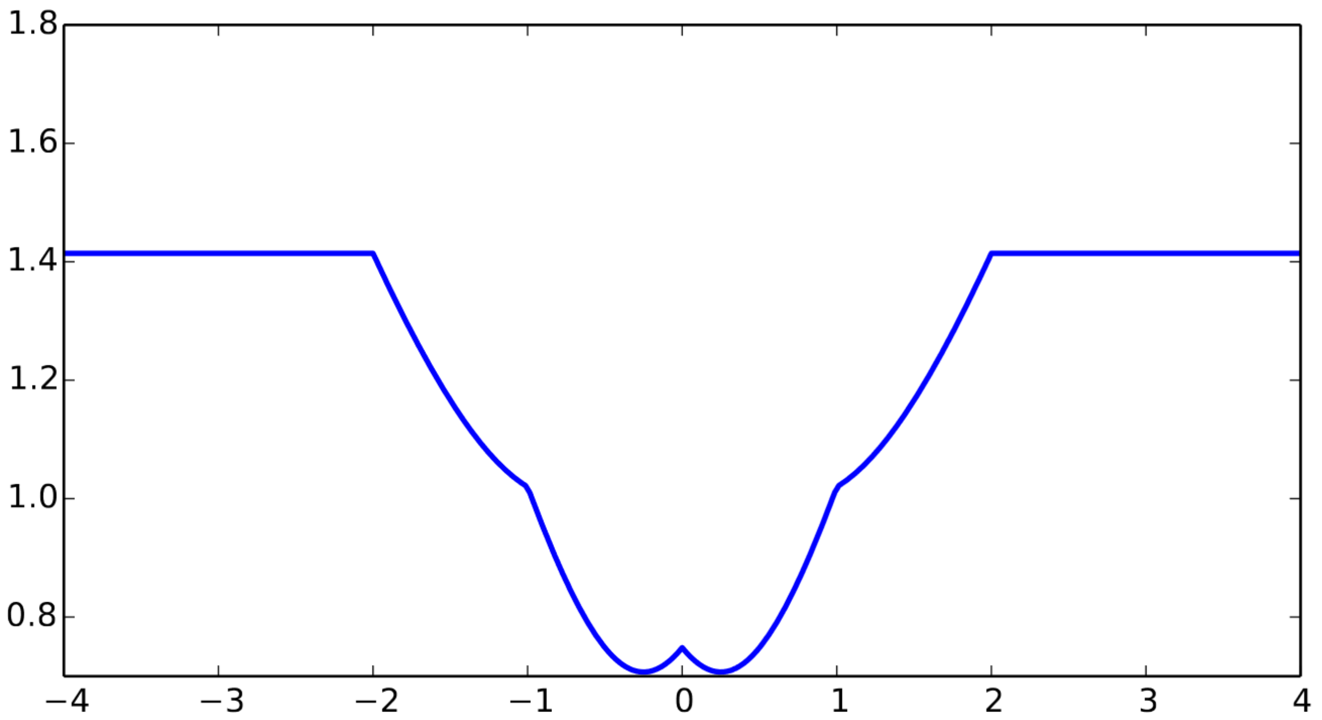
$$\hat{\theta}^{\mathrm{var}} \in \operatorname*{argmin}_{\theta\in\Theta} \left\{ \underbrace{\mathbb{E}_{\hat{P}_n}[\ell(\theta : X)]}_{\text{bias}} + \underbrace{\sqrt{\frac{C\,\mathrm{Var}_{\hat{P}_n}[\ell(\theta : X)]}{n}}}_{\text{variance}} \right\}.$$

Issue: Variance is non-convex.

- Example: $\theta \to \sqrt{\mathrm{Var}[\ell(\theta, X)]}$

for $\ell(\theta; X) = |\theta - X|$

where $X \sim \mathrm{Uni}(\{-2, -1, 0, 1, 2\})$



Robust Optimization ≈ Variance Regularization

Theorem (N. & Duchi 2017)

Assume that $|\ell(\theta; X)| \leq M$. *With prob. at least* $1 - exp\left(-\frac{n\,Var[\ell(\theta;X)]}{36M^2}\right)$

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

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

2. Optimization theoretical analysis

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

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

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

Remark: Do well almost all the time instead of on average. Statistically principled choice of $\mathcal{P}_{n,p}$ → optimality certificates

Generalized Empirical Likelihood

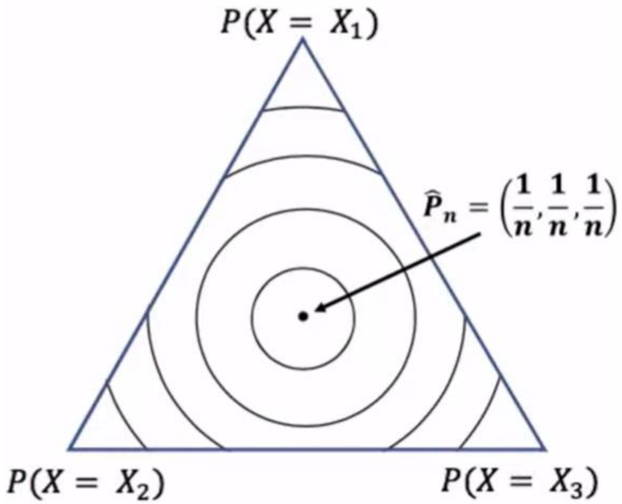
Idea: Instead of using empirical distribution $\hat{P}_n$ on sample $X_1,\ldots,X_n$, look at all distribution 'near' it.

Measures of closeness paper use: Chi-square divergence

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

Worst-case region:

$$\mathcal{P}_{n,p} = \left\{P; D_{\chi^2}\big(P\|\hat{P}_n\big) \leq \frac{\rho}{n}\right\}$$



Robust Optimization

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

Nice properties: 😊 👍

- Convex optimization problem = Computationally efficient
- Conic forms [2]
- Efficient solution methods as fast as stochastic gradient descent [3]

[2] *Ben-Tal A, Den Hertog D, De Waegenaere A, et al. Robust solutions of optimization problems affected by uncertain probabilities[J]. Management Science, 2013, 59(2): 341-357.*
[3] *Duchi J C, Glynn P W, Namkoong H. Statistics of robust optimization: A generalized empirical likelihood approach[J]. Mathematics of Operations Research, 2021, 46(3): 946-969.*

Robust Optimization ≈ Variance Regularization

Theorem (N. & Duchi 2017)

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

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

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

Optimal bias & variance tradeoff

Let $\mathcal{C}_n(\Theta)$ be complexity of $\{\ell(\theta;\cdot);\theta \in \Theta\}$ and

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

Theorem (N. & Duchi 2017)

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

$$R(\hat{\theta}^{\text{rob}}) = \mathbb{E}[\ell(\hat{\theta}^{\text{rob}};X)] \le \underbrace{\min_{\theta\in\Theta}\left\{R(\theta) + 2\sqrt{\frac{2\rho\operatorname{Var}[\ell(\theta:X)]}{n}}\right\}}_{\text{optimal trafeoff}} + \frac{CM\rho}{n}.$$

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

Fast rates from optimal tradeoff

- **ERM:** For $R(\theta^*) \coloneqq \inf_{\theta \in \Theta} R(\theta)$, with high prob.,

$$R\left(\hat{\theta}^{\mathrm{erm}}\right) \le R(\theta^*) + \sqrt{\frac{2\rho M R(\theta^*)}{n}} + \frac{CM\rho}{n}.$$

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

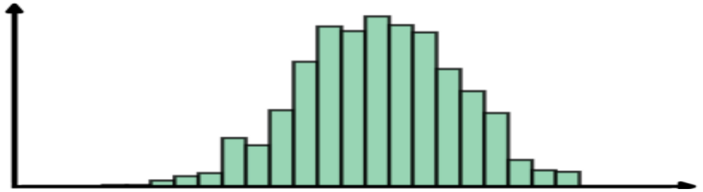
$$R\left(\hat{\theta}^{\mathrm{rob}}\right) \le R(\theta^*) + \frac{C_1}{n} \quad \text{but} \quad R\left(\hat{\theta}^{\mathrm{erm}}\right) \ge R(\theta^*) + \frac{C_2}{\sqrt{n}}$$

Algorithm

Play a two-player stochastic game

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

Adversary: Sample p



Sample

reweight

Player: Run SGD

$$\theta^{t+1} = \theta^{t} - \alpha \nabla \ell(\theta^{t}, X_i)$$

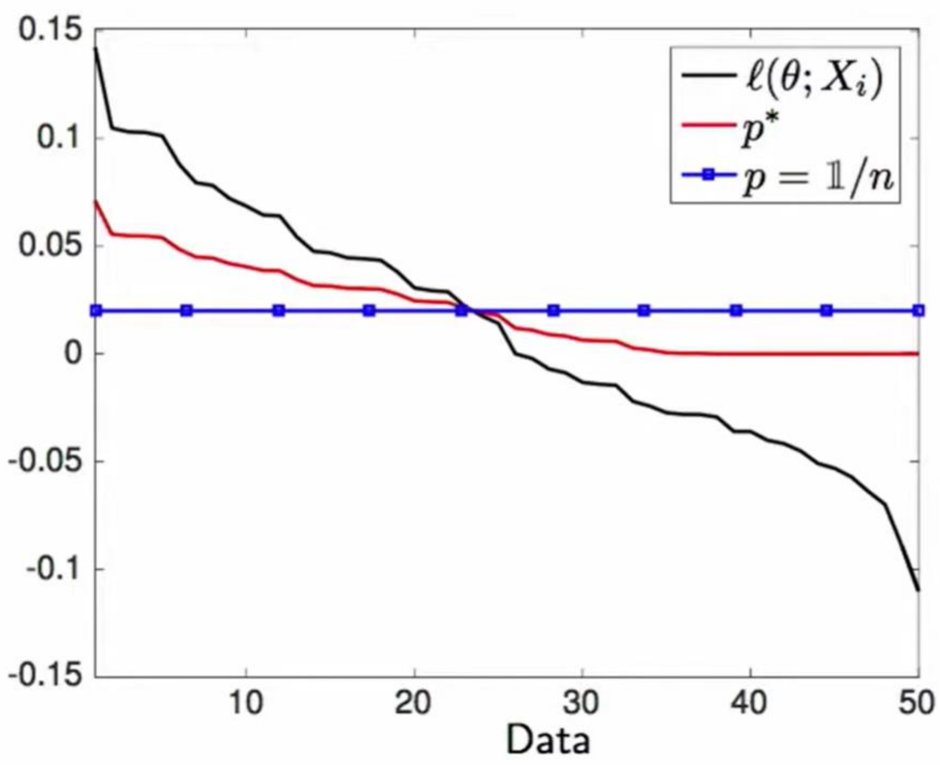
3. Experiments

Upweighting Harder Examples

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

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

on **rare** and **hard** expmples



Reuters Corpus (路透社语料库)

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

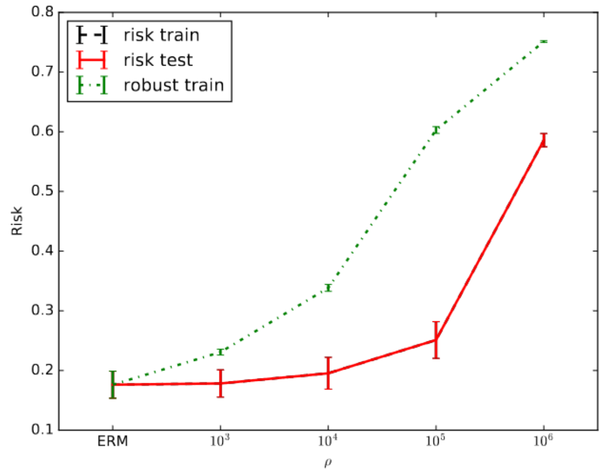
{Corporate, Economics, Government, Markets}

- Data: pairs $x \in \mathbb{R}^d$ represents document, $y \in \{-1, 1\}^4$.
- Logistic loss, with $\Theta = \{\theta \in \mathbb{R}^d; \|\theta\|_1 \leq 1000\}$
- $d = 47{,}236$, $n = 804{,}414$, 10-fold cross-validation/

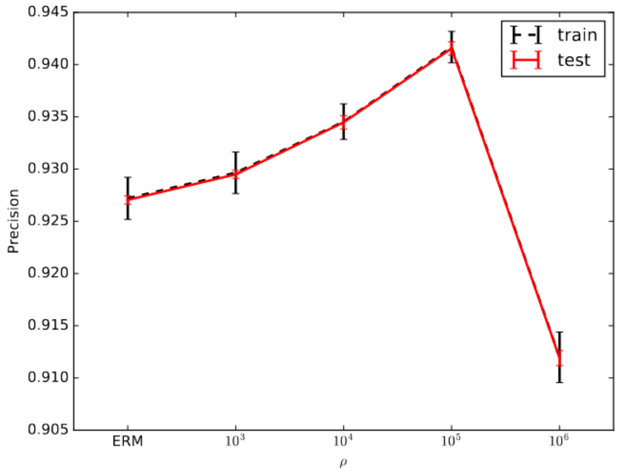
Table: Reuters Number of Examples

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

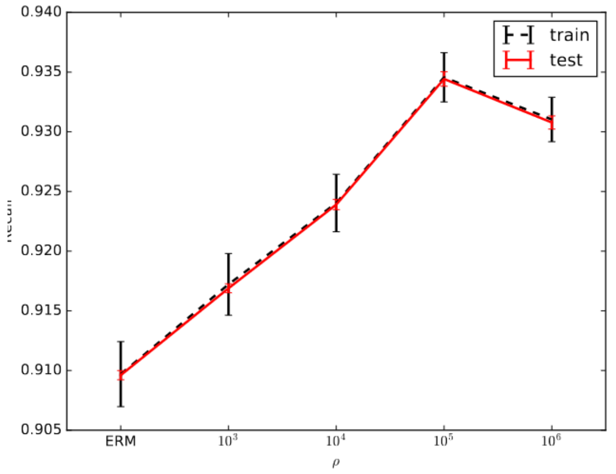
Reuters Corpus (路透社语料库)



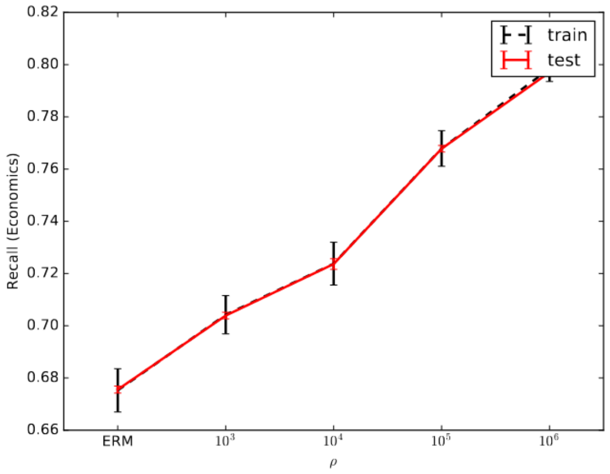
(a) Logistic risk and confidence bound



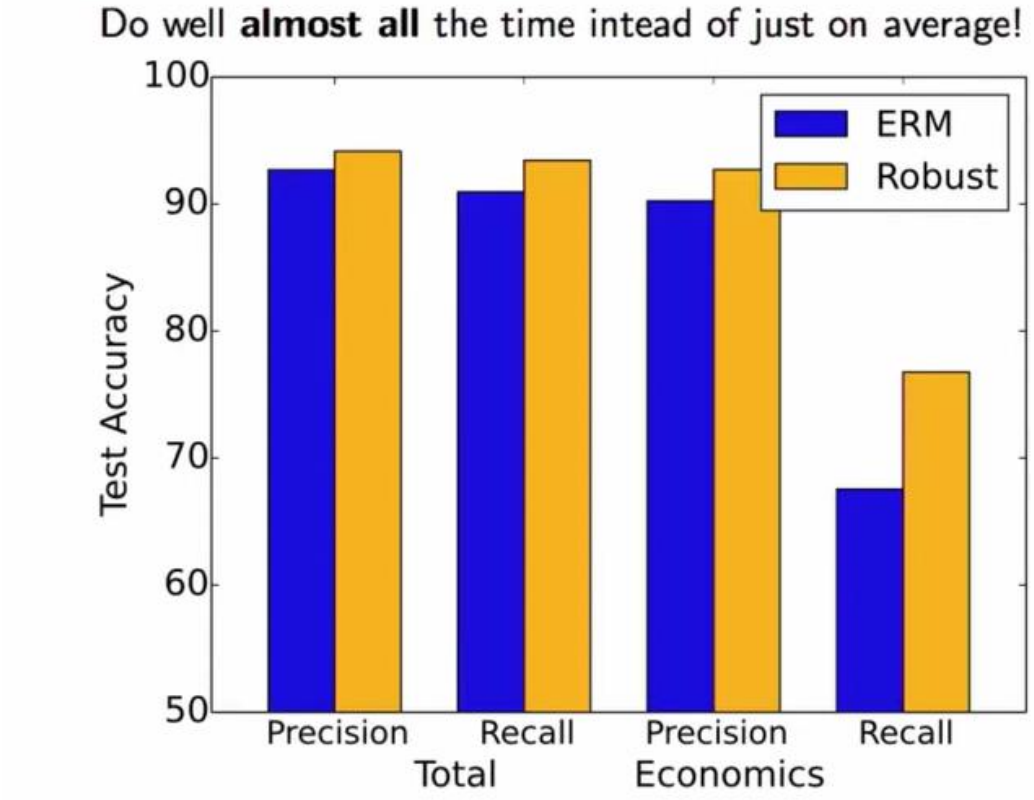
(b) Precision



(c) Recall



(d) Recall on rare category (Economics)



Reuters Corpus (路透社语料库)

Table 5: Reuters Corpus Precision (%)

	Precision		Corporate		Economics		Government		Markets	
ρ	train	test	train	test	train	test	train	test	train	test
erm	92.72	92.7	93.55	93.55	89.02	89	94.1	94.12	92.88	92.94
1E3	92.97	92.95	93.31	93.33	87.84	87.81	93.73	93.76	92.56	92.62
1E4	93.45	93.45	93.58	93.61	87.6	87.58	93.77	93.8	92.71	92.75
1E5	94.17	94.16	94.18	94.19	86.55	86.56	94.07	94.09	93.16	93.24
1E6	91.2	91.19	92	92.02	74.81	74.8	91.19	91.25	89.98	90.18

Table 6: Reuters Corpus Recall (%)

	Recall		Corporate		Economics		Government		Markets	
ρ	train	test	train	test	train	test	train	test	train	test
erm	90.97	90.96	90.20	90.25	67.53	67.56	90.49	90.49	88.77	88.78
1E3	91.72	91.69	90.83	90.86	70.42	70.39	91.26	91.23	89.62	89.58
1E4	92.40	92.39	91.47	91.54	72.38	72.36	91.76	91.76	90.48	90.45
1E5	93.46	93.44	92.65	92.71	76.79	76.78	92.26	92.21	91.46	91.47
1E6	93.10	93.08	92.00	92.04	79.84	79.71	91.89	91.90	92.00	91.97

4. Summary

Summary

Optimization and statistical theory for robust optimization

1. **Convex procedure** for variance regularization.

2. Generalization guarantees for **optimal tradeoff between bias & variance**.

3. Improves performance on **hard instances** empirically.

Reference

[1] Duchi J, Namkoong H. Variance-based regularization with convex objectives[J]. Journal of Machine Learning Research, 2019, 20(68): 1-55.

[2] Ben-Tal A, Den Hertog D, De Waegenaere A, et al. Robust solutions of optimization problems affected by uncertain probabilities[J]. Management Science, 2013, 59(2): 341-357.

[3] Duchi J C, Glynn P W, Namkoong H. Statistics of robust optimization: A generalized empirical likelihood approach[J]. Mathematics of Operations Research, 2021, 46(3): 946-969.

Thx :)