Project: Designing Statistical Estimators That Balance Sample Size, Risk, and Computational Cost

Reporter: Yedi Zhang

ShanghaiTech University

June 26, 2018

Outline

Section 1

Background and Authors' Work

Background

- S. Shalev-Shwartz and N. Srebro, "SVM optimization: inverse dependence on training set size," in 2008
- V. Chandrasekaran and M. I. Jordan, "Computational and statistical tradeoffs via convex relaxation," in 2013.
- D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp, "Living on the edge: A geometric theory of phase transitions in convex optimization," in 2014.
- J. J. Bruer, J. A. Tropp, V. Cevher, and S. R. Becker, "Time-Data Tradeoffs by Aggressive Smoothing," in 2014.

Background and Authors' Work

Authors' Work

- continuous sequence of relaxations
- denoising problem
- regularized linear regression: sparse vector, low-rank matrix
- both theoretically and experimentally

Theoretical Basis

Data Model

- $b = Ax^{\natural} + v$, $(A_{m\times d}, m < d)$

Geometric opportunity

- Descent Cones: $\mathcal{D}(f; x) := \bigcup_{\tau>0} \{y \in R^d : f(x+\tau y) \leq f(x)\}$
- Statistical Dimension: $\delta(\mathcal{C}) := \mathbb{E}_g[||\Pi_{\mathcal{C}}(g)||^2]$

Phase Transition

- $m < \delta$: $\max_{\sigma>0} \frac{\mathbb{E}_v[R(x^*)|A]}{\sigma^2} = 1$
- $m > \delta$: $|\max_{\sigma>0} \frac{\mathbb{E}_v[R(x^*)|A]}{\sigma^2} - \frac{\delta}{m}| \leq tm^{-1}\sqrt{d}$

Theoretical Basis

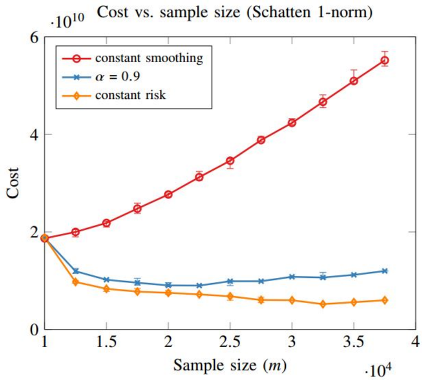
Relaxed Regularizer: f_μ

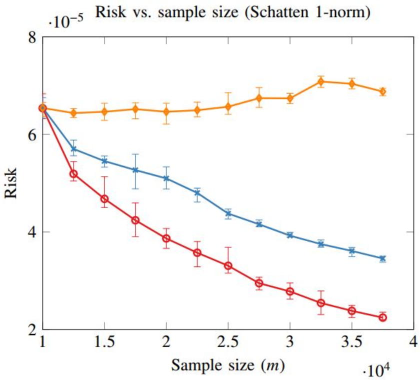
- $f_\mu(x) := f(x) + \frac{\mu}{2}||x||^2$
- Computation Opportunity: Dual-smoothing method

Choosing a Smoothing Parameter

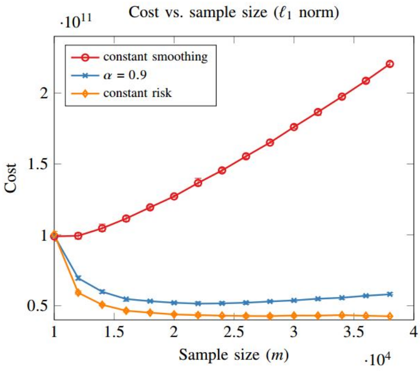
1. Constant Smoothing: fix μ.
2. Constant Risk: $\frac{\delta(\mathcal{D}(f_\mu; x^\natural))}{m} = \frac{\bar{\delta}}{\bar{m}}$
3. A Tunable Balance: $\frac{\delta(\mathcal{D}(f_\mu; x^\natural))}{m} = \frac{\bar{\delta}}{\bar{m} + (m - \bar{m})^\alpha}$

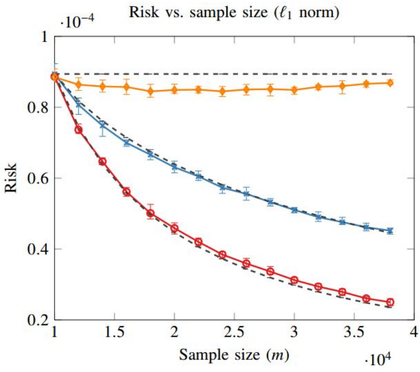
Primal: Low-Rank Matrix





Primal: Sparse Vector





Section 2

Fix some "clerical errors" in this paper......

Thanks to:

J. J. Bruer, J. A. Tropp, V. Cevher, and S. R. Becker, "**Time-Data Tradeoffs by Aggressive Smoothing**," in 2014.

Auslender-Teboulle Algorithm

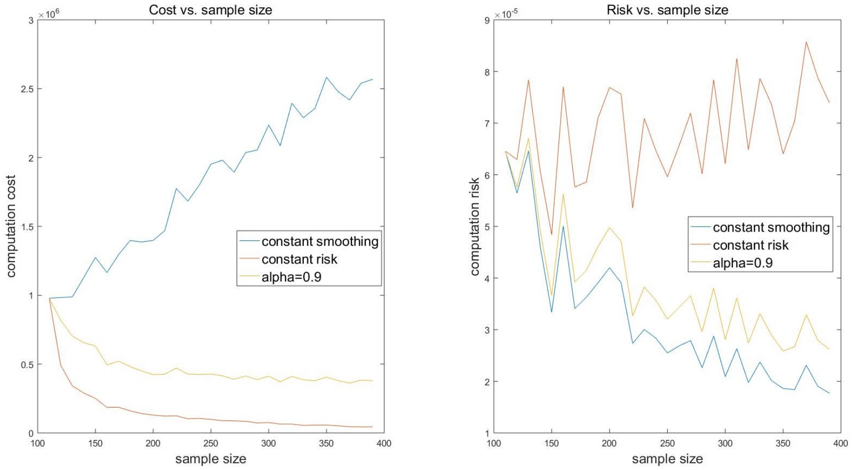
1. $x_k \leftarrow \mu \cdot \text{SoftTresh}(A^T y_k, 1)$
   $\Rightarrow x_k \leftarrow \mu^{-1} \cdot \text{SoftTresh}(A^T z_k, 1)$
2. $\bar{z}_k \leftarrow \text{Shrink}(\bar{z}_k - (b - Ax_k)/(L_\mu \cdot \theta_k), \epsilon/(L_\mu \cdot \theta))$
   $\Rightarrow \bar{z}_k \leftarrow \text{Shrink}(\bar{z}_k + (b - Ax_k)/(L_\mu \cdot \theta_k), \epsilon/(L_\mu \cdot \theta))$

Adjust Experiment Scale and Parameter

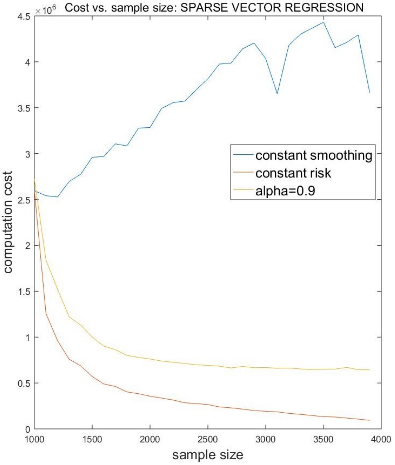
Table: scale and parameter adjustment

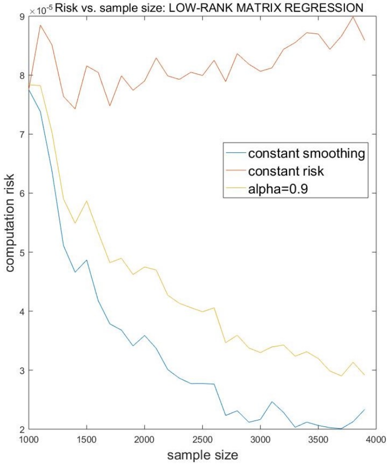
	scale	$\bar{\delta}$	$\bar{\mu}$	$\bar{m}$	σ	samples
vec	40000	98000	0.1	10000	0.01	10000$\sim$38000
mat	200$\times$200	98000	0.1	10000	0.01	10000$\sim$37500
vec$_{adj}$	400	90	0.1	110	0.01	110$\sim$390
vec$_{adj}$	4000	894	0.1	957	0.01	1000$\sim$3900
mat$_{adj}$	20$\times$20	87	0.1	107	0.01	110$\sim$390
mat$_{adj}$	40$\times$40	350	0.1	390	0.01	400$\sim$1500

Implement: Sparse Vector with $m = 400$

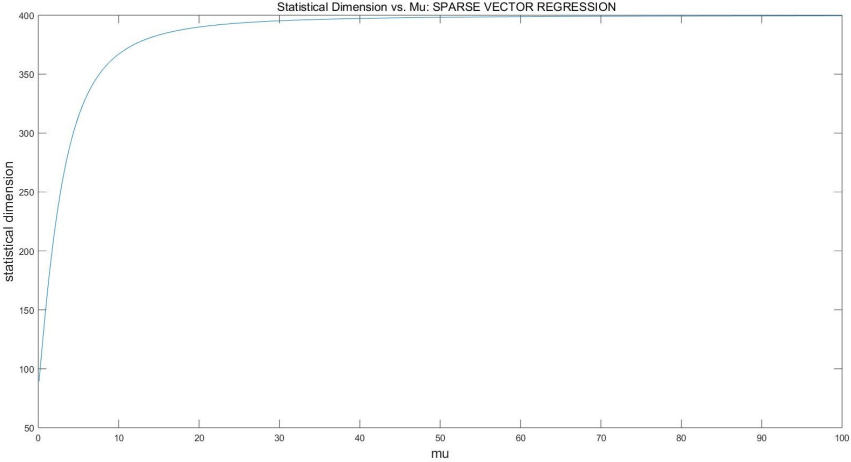


Implement: Sparse Vector with $m = 4000$

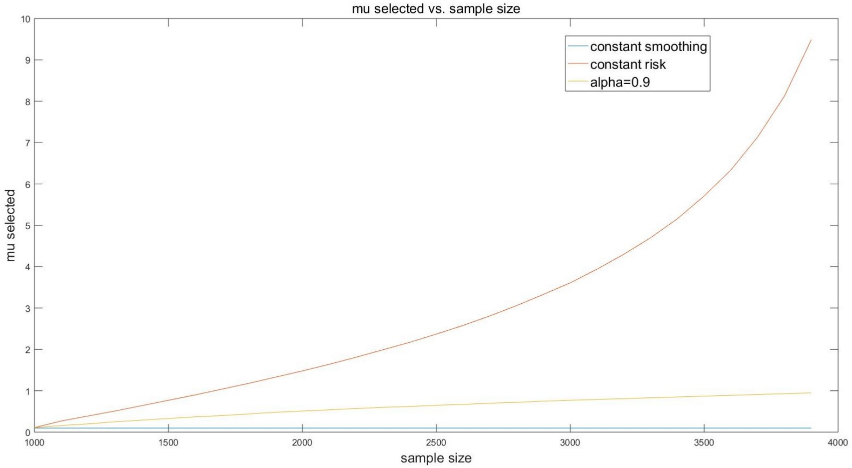




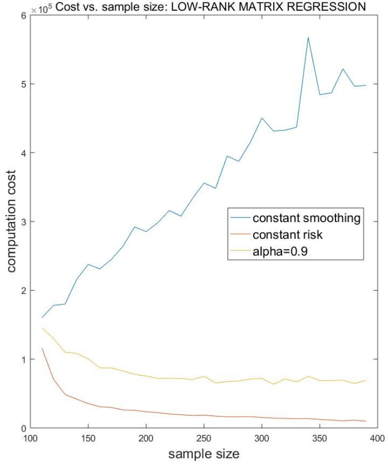
Sparse Vector: δ vs. μ

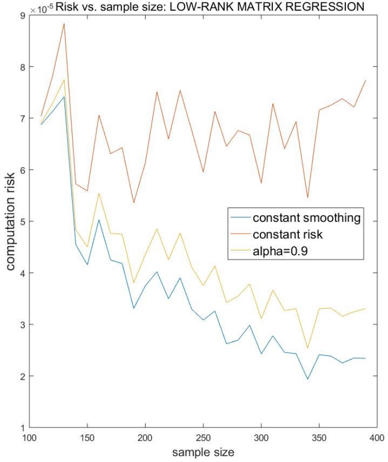


Sparse Vector: selected μ vs. m

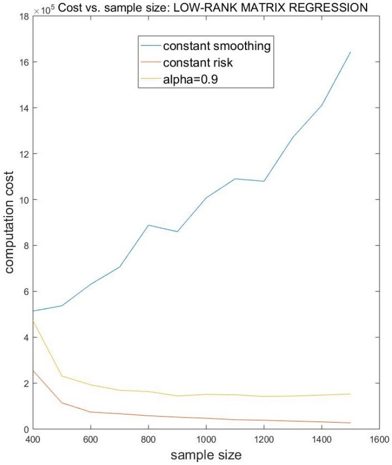


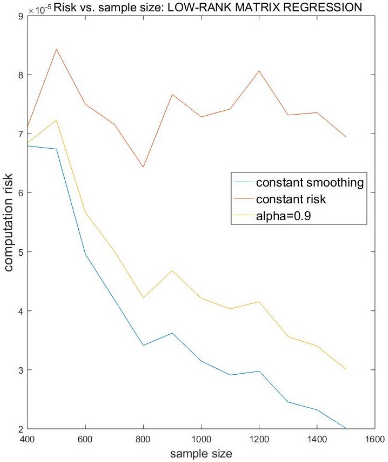
Implement: Low-Rank Matrix with $d = 20 \times 20$



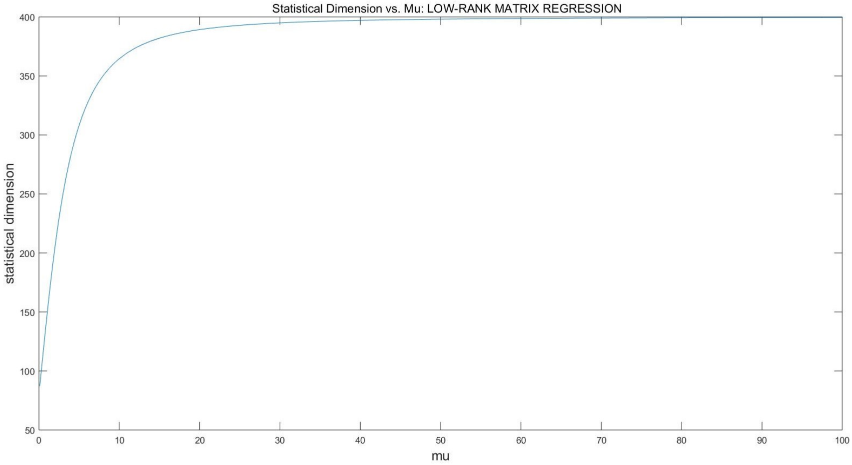


Implement: Low-Rank Matrix with $d = 40 \times 40$

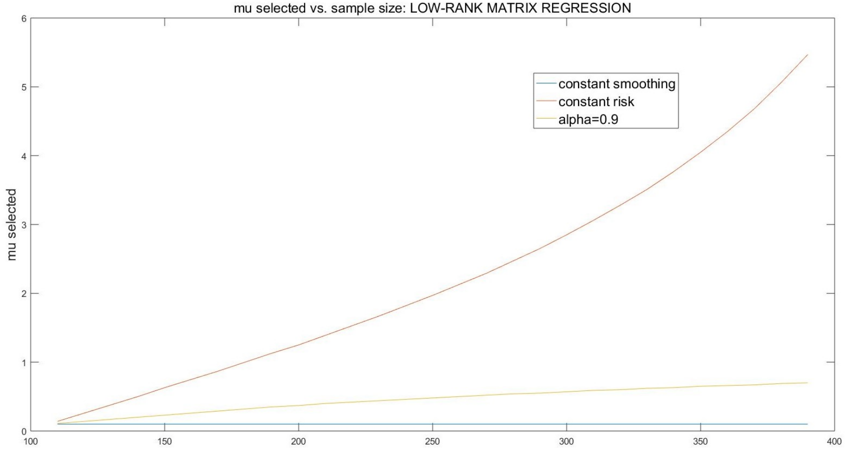




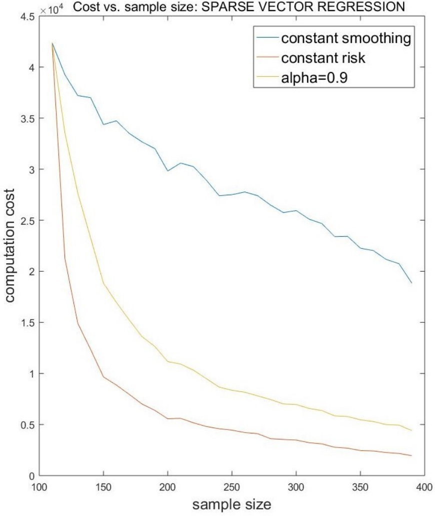
Low-Rank Matrix: δ vs. μ

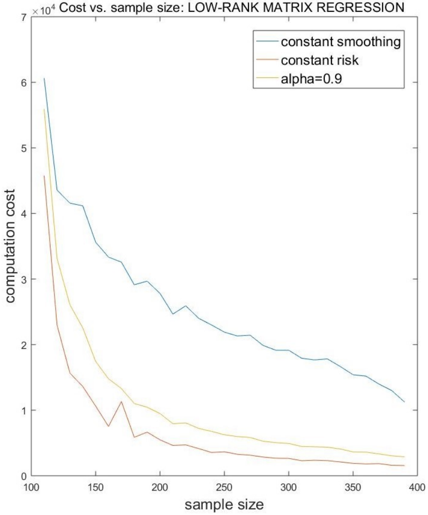


Low-Rank Matrix: selected μ vs. m



Without noise: $\sigma = 0$





Section 3

Conclusion

- When we have excess samples in the data set, we can exploit them to decrease the statistical risk of estimator, or to lower the computational cost through additional smoothing.
- When there is no noise, we can recover the unknown signal faster with more relaxation on regularized function (using "constant risk" scheme) without losing any accuracy.

Thanks!