Minimax Rates for Memory-Constrained Sparse Linear Regression

Jacob Steinhardt **John Duchi**

{jsteinha,jduchi}@stanford.edu

Resource-Constrained Learning

How do we solve statistical problems with limited resources?

- communication / memory constraints (Zhang et al., 2013; Garg et al., 2014; Shamir, 2014)
- privacy, computation constraints (Kasiviswanathan et al., 2011; Duchi et al., 2013; Berthet and Rigollet, 2013)
- NP-hardness of sparse regression (Zhang et al., 2014; Natarajan, 1995)

This work: sparse linear regression under memory constraints.

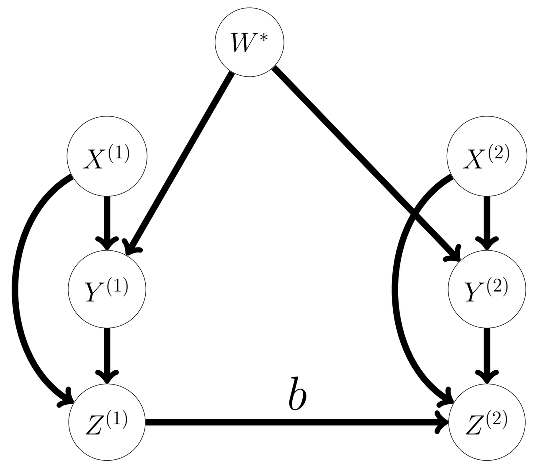
Setting

Sparse linear regression in $\mathbb{R}^d$:

- $Y^{(i)} = \langle w^*, X^{(i)} \rangle + \epsilon^{(i)}$
- $\|w^*\|_0 = k$, $k \ll d$

Memory constraint:

- $(Y^{(i)}, X^{(i)})$ observed as read-only stream
- Only keep b bits of state $Z^{(i)}$ between successive observations



Problem Statement

How much data n is needed to obtain estimator $\hat{w}$ with

$$\mathbb{E}[\|\hat{w} - w^*\|_2^2] \leq \epsilon?$$

Classical case (no memory constraint):

Theorem (Wainwright, 2009)**.**

$$\frac{k}{\epsilon}\log(d) \lesssim n \lesssim \frac{k}{\epsilon}\log(d)$$

With memory constraints b:

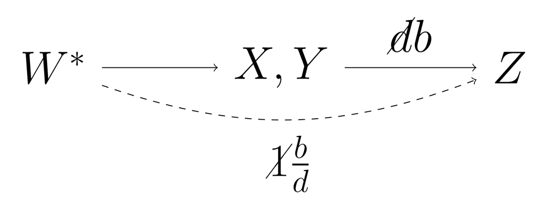
Theorem (S. & Duchi, 2015)**.**

$$\frac{k}{\epsilon}\frac{d}{b} \lesssim n \lesssim \frac{k}{\epsilon^2}\frac{d}{b}$$

Exponential increase if $b \ll d$!

Proof Overview

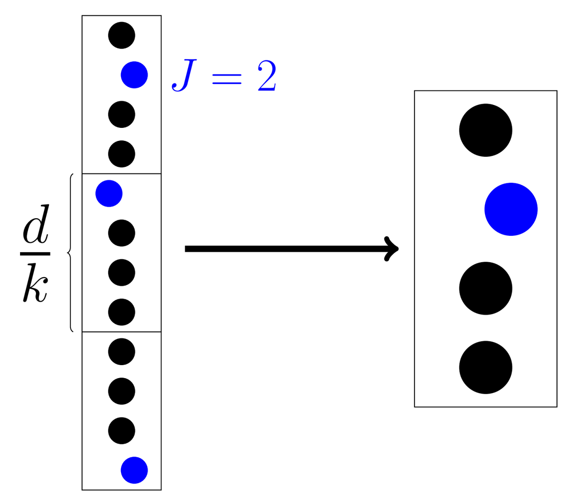
- Lower bound:
  - information-theoretic
  - strong data-processing inequality



- Upper bound:
  - count-min sketch + ℓ^1-regularized dual averaging
  - more regularization $\rightarrow$ easier sketching problem

Lower Bound Construction

- Split coordinates into k blocks of size d/k
- w^* in each block: single non-zero coordinate J, $\pm\delta$ with equal probability
- Direct sum argument: reduce to $k = 1$



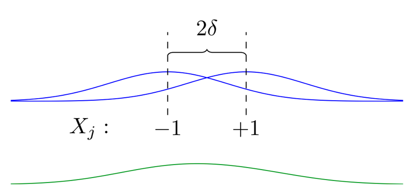
- Estimation to testing:

$$\mathbb{E}[\|w^* - \hat{w}\|_2^2] \geq \frac{\delta^2}{2}\mathbb{P}[J \neq \hat{J}]$$

Looking ahead: bound KL between P_j and base distribution P_0

Some Information Theory

- Let $X \sim \mathrm{Uniform}(\{\pm 1\}^d)$
- Let $P_j(Z^{(1:n)})$ be distribution conditioned on $J = j$
- Let $P_0(Z^{(1:n)})$ be distribution with Y independent of X
- Assouad's method:



$$\mathbb{P}[J \neq \hat{J}] \geq \frac{1}{2} - \sqrt{\frac{1}{d}\sum_{j=1}^{d} \mathrm{KL}\left(P_0(Z^{(1:n)}) \;\|\; P_j(Z^{(1:n)})\right)}$$

- Intuition: $\mathrm{KL}(P_0 \| P_j)$ small unless Z stores info about X_j

Strong Data-Processing Inequality

Focus on a single index $Z = Z^{(i)}$, with $\hat{z} = z^{(1:i-1)}$ fixed.

Proposition. *For any* $\hat{z}$*,*

$$\begin{aligned}\mathrm{KL}\left(P_0(Z \mid \hat{z}) \,\|\, P_j(Z \mid \hat{z})\right) &\leq 4\delta^2 I(X_j; Z \mid Y, \hat{Z} = \hat{z}) \\ &\leq 4\delta^2 \underbrace{I(X_j; Z, Y \mid \hat{Z} = \hat{z})}_{\text{mutual information}}\end{aligned}$$

Plug into Assouad:

$$\begin{aligned}\frac{1}{d}\sum_{j=1}^{d}\mathrm{KL}\left(P_0 \,\|\, P_j\right) &\leq \frac{4\delta^2}{d}\sum_{j=1}^{d} I(X_j; Z, Y \mid \hat{Z}) \\ &\leq \frac{4\delta^2}{d}\underbrace{I(X; Z, Y \mid \hat{Z})}_{b+O(1)}\end{aligned}$$

Only get $\frac{4\delta^2 b}{d}$ bits per round!

Upper Bound

Solve ℓ^1-regularized dual averaging problem (Xiao, 2010), $\lambda \gg 1$:

$$w^{(i)} = \operatorname*{argmin}_{w}\left\{\langle \theta^{(i)}, w\rangle + \lambda\sqrt{n}\|w\|_1\right\},$$

$$\theta^{(i)} = \sum_{i'=1}^{i-1} x^{(i')}(y^{(i')} - \langle w^{(i')}, x^{(i')}\rangle).$$

Hard part: determine support of $w^{(i)}$.

- Need to distinguish $|\theta_j| \geq \lambda\sqrt{n}$ (signal) from $|\theta_j| \approx \sqrt{n}$ (noise)
- Can use count-min sketch, memory usage $\approx \frac{d\log(d)}{\lambda^2}$
  $\implies$ computational-statistical tradeoff; seen before in ℓ^2 case (Shalev-Shwartz & Zhang, 2013; Bruer et al., 2014)

Discussion

Summary:

- Upper and lower bounds on memory-constrained regression
- Lower bound: extend data processing inequality to handle covariates
- Upper bound: use ℓ^1-regularizer to reduce to sketching

Future work:

- Close the gap ($kd/b\epsilon$ vs $kd/b\epsilon^2$)
- Weaken upper bound assumptions

The first author was supported by the Hertz foundation and by the NSF.