Minimax Rates for Memory-Constrained Sparse Linear Regression

Jacob Steinhardt   John Duchi

Stanford University

{*jsteinha,jduchi*}*@stanford.edu*

July 6, 2015

Resource-Constrained Learning

How do we solve statistical problems with limited resources?

Resource-Constrained Learning

How do we solve statistical problems with limited resources?

- computation (Natarajan, 1995; Berthet & Rigollet, 2013; Zhang et al., 2014; Foster et al., 2015)

Resource-Constrained Learning

How do we solve statistical problems with limited resources?

- computation (Natarajan, 1995; Berthet & Rigollet, 2013; Zhang et al., 2014; Foster et al., 2015)
- privacy (Kasiviswanathan et al., 2011; Duchi et al., 2013)

Resource-Constrained Learning

How do we solve statistical problems with limited resources?

- computation (Natarajan, 1995; Berthet & Rigollet, 2013; Zhang et al., 2014; Foster et al., 2015)
- privacy (Kasiviswanathan et al., 2011; Duchi et al., 2013)
- communication / memory (Zhang et al., 2013; Shamir, 2014; Garg et al., 2014; Braverman et al., 2015)

Setting

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

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

Setting

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

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

Memory constraint:

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

Setting

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

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

Memory constraint:

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

w^*

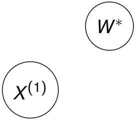
Setting

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

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

Memory constraint:

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



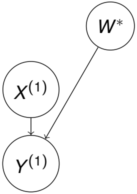
Setting

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

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

Memory constraint:

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



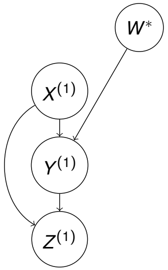
Setting

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

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

Memory constraint:

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



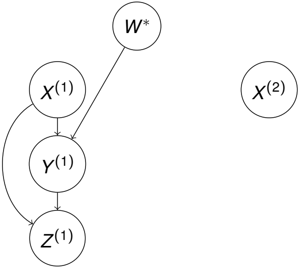
Setting

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

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

Memory constraint:

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



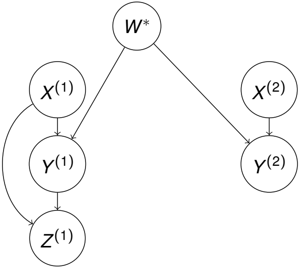
Setting

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

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

Memory constraint:

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



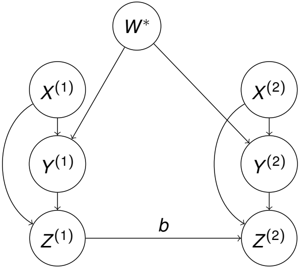
Setting

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

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

Memory constraint:

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



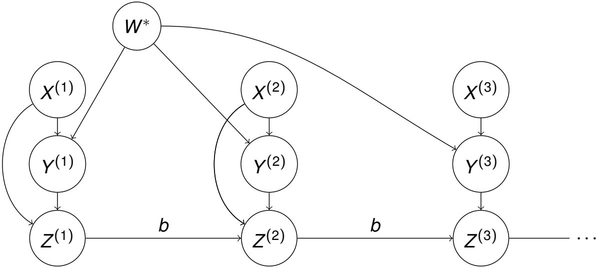
Setting

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

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

Memory constraint:

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



Motivating Question

If we have enough memory to ***represent*** *the answer, can we also efficiently* ***learn*** *the answer?*

Problem Statement

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

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

Problem Statement

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

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

Classical case (no memory constraint):

Theorem (Wainwright, 2009)

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

Problem Statement

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

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

Classical case (no memory constraint):

Theorem (Wainwright, 2009)

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

Achievable with $\tilde{\mathcal{O}}(d)$ memory (Agarwal et al., 2012; S., Wager, & Liang, 2015).

Problem Statement

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

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

Classical case (no memory constraint):

Theorem (Wainwright, 2009)

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

With memory constraints b:

Theorem (S. & Duchi, 2015)

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

Problem Statement

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

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

Classical case (no memory constraint):

Theorem (Wainwright, 2009)

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

With memory constraints b:

Theorem (S. & Duchi, 2015)

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

Exponential increase if $b \ll d$!

Problem Statement

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

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

Classical case (no memory constraint):

Theorem (Wainwright, 2009)

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

With memory constraints b:

Theorem (S. & Duchi, 2015)

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

[Note: up to log factors; assumes $k\log(d) \ll b \leq d$]

Proof Overview

- Lower bound:

Proof Overview

- Lower bound:
  - information-theoretic

Proof Overview

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

$$\begin{tikzcd} W^* \arrow[r] \arrow[rr, bend right, dashed, "1"'] & X, Y \arrow[r, "d"] & Z \end{tikzcd}$$

Proof Overview

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

$$\begin{tikzcd} W^* \arrow[r] \arrow[rr, dashed, bend right, "\cancel{1}\ \frac{b}{d}"'] & X, Y \arrow[r, "\cancel{d}\ b"] & Z \end{tikzcd}$$

Proof Overview

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

$$\begin{tikzcd} W^* \arrow[r] \arrow[rr, dashed, bend right, "\cancel{1}\,\frac{b}{d}"'] & X, Y \arrow[r, "\cancel{d}\,b"] & Z \end{tikzcd}$$

  - main challenge: dependence between X, Y

Proof Overview

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

$$\begin{tikzcd} W^* \arrow[r] \arrow[rr, dashed, bend right, "\cancel{1}\,\frac{b}{d}"'] & X, Y \arrow[r, "\cancel{d}\,b"] & Z \end{tikzcd}$$

  - main challenge: dependence between X, Y
- Upper bound:

Proof Overview

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

$$\begin{tikzcd} W^* \arrow[r] \arrow[rr, dashed, bend right, "\cancel{1}\ \frac{b}{d}"'] & X, Y \arrow[r, "\cancel{d}\, b"] & Z \end{tikzcd}$$

  - main challenge: dependence between X, Y
- Upper bound:
  - count-min sketch + ℓ^1-regularized dual averaging

Proof Overview

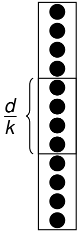
- Lower bound:
  - information-theoretic
  - strong data-processing inequality

$$\begin{tikzcd} W^* \arrow[r] \arrow[rr, dashed, bend right, "\cancel{1}\,\frac{b}{d}"'] & X, Y \arrow[r, "\cancel{d}\,b"] & Z \end{tikzcd}$$

  - main challenge: dependence between X, Y
- Upper bound:
  - count-min sketch + ℓ^1-regularized dual averaging
  - more regularization $\rightarrow$ easier sketching problem

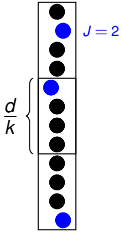
Lower Bound Construction

- Split coordinates into k blocks of size d/k



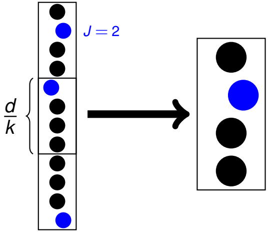
Lower Bound Construction

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



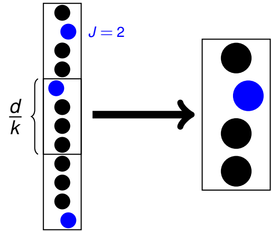
Lower Bound Construction

- 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 $k = 1$



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 $k = 1$

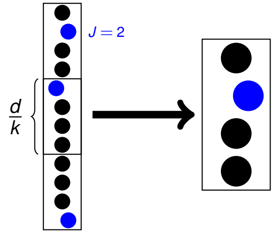


- Estimation to testing:

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

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:

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

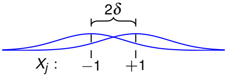
Looking ahead: bound KL between P_j and base distribution P_0

Some Information Theory

- Let $X \sim \text{Uniform}(\{\pm 1\}^d)$

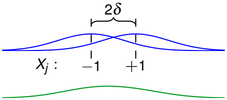
Some Information Theory

- Let $X \sim \text{Uniform}(\{\pm 1\}^d)$
- Let $P_j(Z^{(1:n)})$ be distribution conditioned on $J = j$



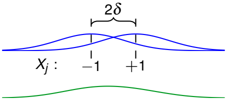
Some Information Theory

- Let $X \sim \text{Uniform}(\{\pm 1\}^d)$
- 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



Some Information Theory

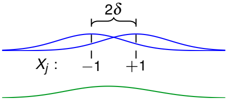
- Let $X \sim \mathrm{Uniform}(\{\pm 1\}^d)$
- Let $P_j(Z^{(1:n)})$ be distribution conditioned on $J = j$
- Let $P_0(Z^{(1:n)})$ be distribution with Y independent of X
- Assouad's method:



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

Some Information Theory

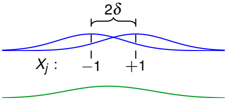
- Let $X \sim \text{Uniform}(\{\pm 1\}^d)$
- Let $P_j(Z^{(1:n)})$ be distribution conditioned on $J = j$
- Let $P_0(Z^{(1:n)})$ be distribution with Y independent of X
- Assouad's method:



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

Some Information Theory

- Let $X \sim \text{Uniform}(\{\pm 1\}^d)$
- 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:

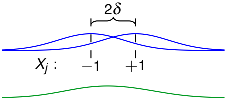


$$\mathbb{P}[J \neq \hat{J}] \geq \frac{1}{2} - \sqrt{\frac{1}{d}\sum_{j=1}^{d} D_{\mathrm{kl}}\left(P_0(Z^{(1:n)}) \,\|\, P_j(Z^{(1:n)})\right)}$$

- Key fact: (Y, X_j) independent of $X_{\neg j}$ under P_j

Some Information Theory

- Let $X \sim \text{Uniform}(\{\pm 1\}^d)$
- 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:



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

- Key fact: (Y, X_j) independent of $X_{\neg j}$ under P_j
  - Intuition: $D_{\rm kl}(P_0 \,\|\, P_j)$ small unless Z stores info about X_j; need to store majority of X_j to make average $D_{\rm kl}$ small.

Strong Data-Processing Inequality

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

Strong Data-Processing Inequality

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

Proposition

For any $\hat{z}$,

$$D_{\rm kl}\left(P_0(Z \mid \hat{z}) \,\|\, P_j(Z \mid \hat{z})\right) \le 4\delta^2 \underbrace{I(X_j; Z \mid Y, \hat{Z} = \hat{z})}_{\text{mutual information}}$$

Strong Data-Processing Inequality

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

Proposition

For any $\hat{z}$,

$$\begin{aligned} D_{\mathrm{kl}}\left(P_0(Z \mid \hat{z}) \,\|\, P_j(Z \mid \hat{z})\right) &\leq 4\delta^2 I(X_j; Z \mid Y, \hat{Z} = \hat{z}) \\ &\leq 4\delta^2 I(X_j; Z, Y \mid \hat{Z} = \hat{z}) \end{aligned}$$

Strong Data-Processing Inequality

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

Proposition

For any $\hat{z}$,

$$\begin{aligned} D_{\mathrm{kl}}\left(P_0(Z \mid \hat{z}) \,\|\, P_j(Z \mid \hat{z})\right) &\le 4\delta^2 I(X_j; Z \mid Y, \hat{Z} = \hat{z}) \\ &\le 4\delta^2 I(X_j; Z, Y \mid \hat{Z} = \hat{z}) \end{aligned}$$

Plug into Assouad:

$$\frac{1}{d}\sum_{j=1}^{d} D_{\mathrm{kl}}\left(P_0 \,\|\, P_j\right)$$

Strong Data-Processing Inequality

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

Proposition

For any $\hat{z}$,

$$\begin{aligned} D_{\mathrm{kl}}\left(P_0(Z \mid \hat{z}) \,\|\, P_j(Z \mid \hat{z})\right) &\le 4\delta^2 I(X_j; Z \mid Y, \hat{Z} = \hat{z}) \\ &\le 4\delta^2 I(X_j; Z, Y \mid \hat{Z} = \hat{z}) \end{aligned}$$

Plug into Assouad:

$$\frac{1}{d}\sum_{j=1}^{d} D_{\mathrm{kl}}\left(P_0 \,\|\, P_j\right) \le \frac{4\delta^2}{d}\sum_{j=1}^{d} I(X_j; Z, Y \mid \hat{Z})$$

Strong Data-Processing Inequality

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

Proposition

For any $\hat{z}$,

$$\begin{aligned} D_{\mathrm{kl}}\left(P_0(Z \mid \hat{z}) \,\|\, P_j(Z \mid \hat{z})\right) &\le 4\delta^2 I(X_j; Z \mid Y, \hat{Z} = \hat{z}) \\ &\le 4\delta^2 I(X_j; Z, Y \mid \hat{Z} = \hat{z}) \end{aligned}$$

Plug into Assouad:

$$\begin{aligned} \frac{1}{d} \sum_{j=1}^{d} D_{\mathrm{kl}}\left(P_0 \,\|\, P_j\right) &\le \frac{4\delta^2}{d} \sum_{j=1}^{d} I(X_j; Z, Y \mid \hat{Z}) \\ &\le \frac{4\delta^2}{d} I(X; Z, Y \mid \hat{Z}) \end{aligned}$$

Strong Data-Processing Inequality

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

Proposition

For any $\hat{z}$,

$$\begin{aligned} D_{\mathrm{kl}}\left(P_0(Z \mid \hat{z}) \,\|\, P_j(Z \mid \hat{z})\right) &\leq 4\delta^2 I(X_j; Z \mid Y, \hat{Z} = \hat{z}) \\ &\leq 4\delta^2 I(X_j; Z, Y \mid \hat{Z} = \hat{z}) \end{aligned}$$

Plug into Assouad:

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

Strong Data-Processing Inequality

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

Proposition

For any $\hat{z}$,

$$\begin{aligned} D_{\mathrm{kl}}\left(P_0(Z \mid \hat{z}) \,\|\, P_j(Z \mid \hat{z})\right) &\le 4\delta^2 I(X_j; Z \mid Y, \hat{Z} = \hat{z}) \\ &\le 4\delta^2 I(X_j; Z, Y \mid \hat{Z} = \hat{z}) \end{aligned}$$

Plug into Assouad:

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

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

Upper Bound

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

$$w^{(i)} = \operatorname{argmin}_w \left\{ \langle \theta^{(i)}, w \rangle + \lambda \sqrt{n} \|w\|_1 + \frac{1}{2\eta} \|w\|_2^2 \right\},$$

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

Upper Bound

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

$$w^{(i)} = \mathrm{argmin}_w \left\{ \langle \theta^{(i)}, w \rangle + \lambda \sqrt{n} \|w\|_1 + \frac{1}{2\eta} \|w\|_2^2 \right\},$$

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

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

Upper Bound

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

$$w^{(i)} = \mathrm{argmin}_w \left\{ \langle \theta^{(i)}, w \rangle + \lambda\sqrt{n}\|w\|_1 + \frac{1}{2\eta}\|w\|_2^2 \right\},$$

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

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

Upper Bound

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

$$w^{(i)} = \text{argmin}_w \left\{ \langle \theta^{(i)}, w \rangle + \lambda\sqrt{n} \|w\|_1 + \frac{1}{2\eta}\|w\|_2^2 \right\},$$

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

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

- Need to distinguish $|\theta_j| \geq \lambda\sqrt{n}$ (signal) from $|\theta_j| \approx \sqrt{n}$ (noise)

Upper Bound

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

$$w^{(i)} = \operatorname{argmin}_w \left\{ \langle \theta^{(i)}, w \rangle + \lambda\sqrt{n}\|w\|_1 + \frac{1}{2\eta}\|w\|_2^2 \right\},$$

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

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 $\approx \frac{d\log(d)}{\lambda^2}$

Upper Bound

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

$$w^{(i)} = \operatorname{argmin}_w \left\{ \langle \theta^{(i)}, w \rangle + \lambda\sqrt{n} \|w\|_1 + \frac{1}{2\eta} \|w\|_2^2 \right\},$$

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

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 $\approx \frac{d \log(d)}{\lambda^2}$

  $\implies$ regularization decreases computation; seen before in ℓ^2 case (Shalev-Shwartz & Zhang, 2013; Bruer et al., 2014)

Discussion

Summary:

Discussion

Summary:

- Upper and lower bounds on memory-constrained regression

Discussion

Summary:

- Upper and lower bounds on memory-constrained regression
- Lower bound: extend data processing inequality to handle covariates

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

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:

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 ($kd/b\varepsilon$ vs $kd/b\varepsilon^2$)

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 ($kd/b\varepsilon$ vs $kd/b\varepsilon^2$)
- Weaken upper bound assumptions