SIAM Conference on
Computational Science and Engineering
February 27-March 3, 2017
Hilton Atlanta, Atlanta, Georgia, USA

Sketched Ridge Regression:

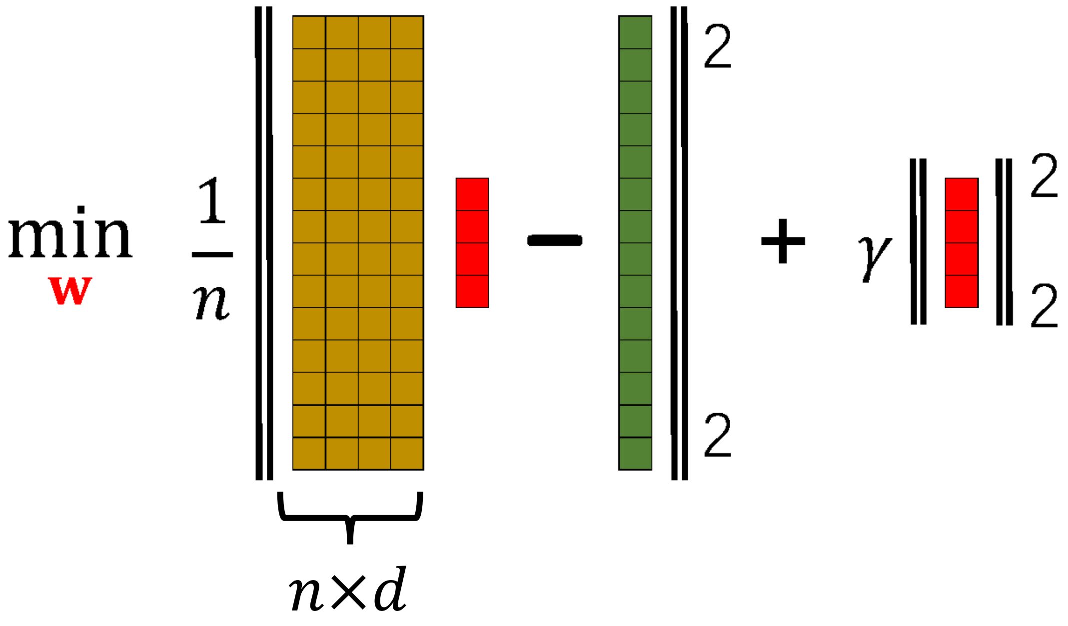
Optimization and Statistical Perspectives

Shusen Wang — UC Berkeley
Alex Gittens — RPI
Michael Mahoney — UC Berkeley

Overview

Ridge Regression

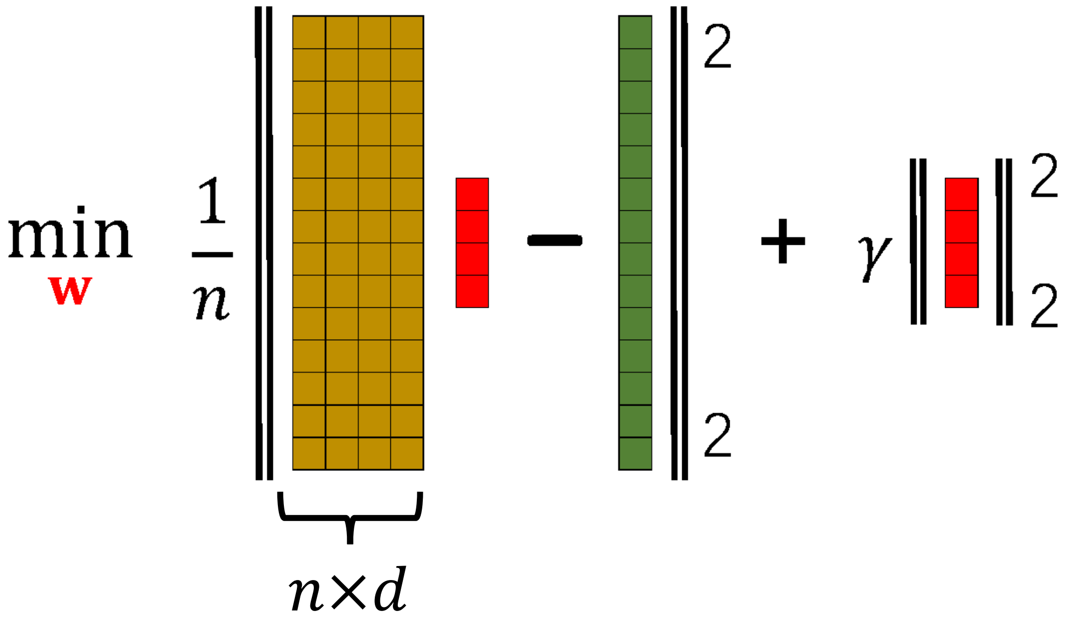
$$\min_{\mathbf{w}} \left\{ f(\mathbf{w}) = \frac{1}{n} ||\mathbf{X}\mathbf{w} - \mathbf{y}||_2^2 + \gamma ||\mathbf{w}||_2^2 \right\}$$



Over-determined:
$n \gg d$

Ridge Regression

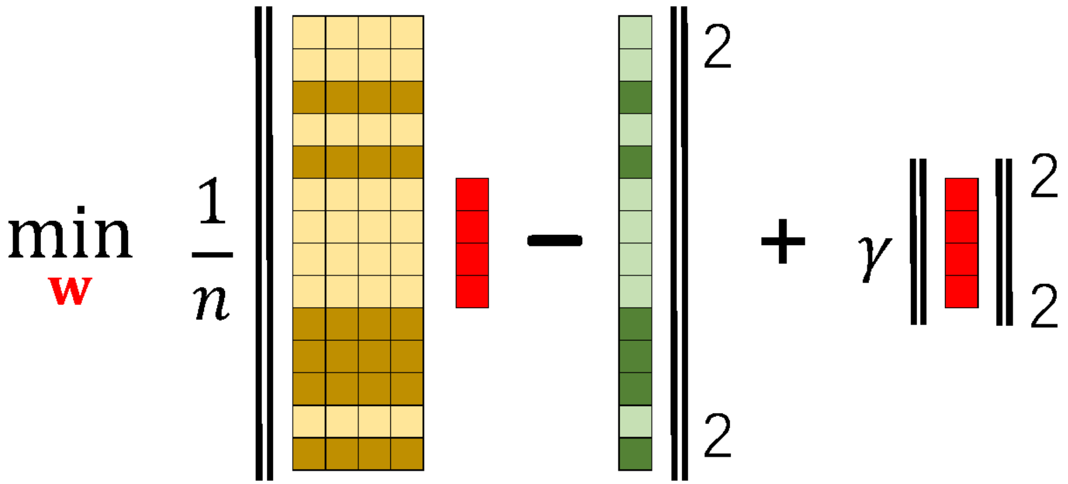
$$\min_{\mathbf{w}} \left\{ f(\mathbf{w}) = \frac{1}{n} ||\mathbf{X}\mathbf{w} - \mathbf{y}||_2^2 + \gamma ||\mathbf{w}||_2^2 \right\}$$



- Efficient and approximate solution?
- Use only part of the data?

Ridge Regression

$$\min_{\mathbf{w}} \left\{ f(\mathbf{w}) = \frac{1}{n} ||\mathbf{X}\mathbf{w} - \mathbf{y}||_2^2 + \gamma ||\mathbf{w}||_2^2 \right\}$$

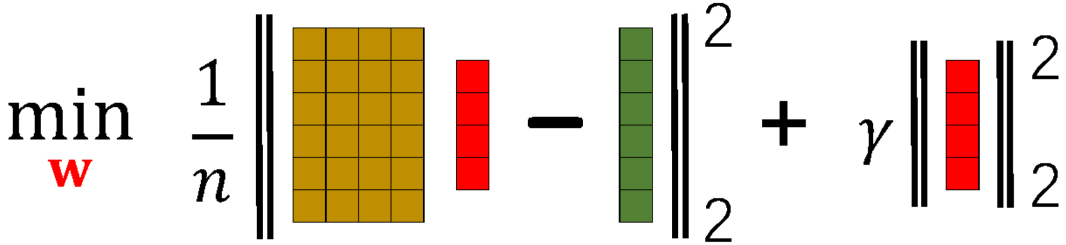


Matrix Sketching:

- Random selection
- Random projection

Approximate Ridge Regression

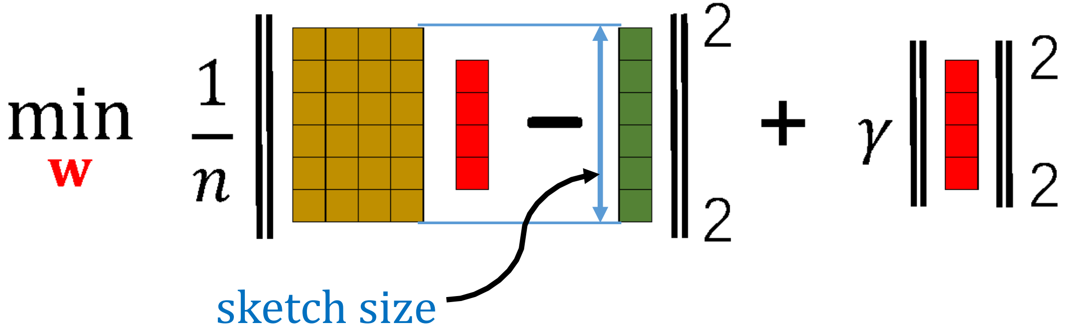
$$\min_{\mathbf{w}} \left\{ f(\mathbf{w}) = \frac{1}{n} ||\mathbf{X}\mathbf{w} - \mathbf{y}||_2^2 + \gamma ||\mathbf{w}||_2^2 \right\}$$



- Sketched solution: $\mathbf{w}^{\mathrm{s}}$

Approximate Ridge Regression

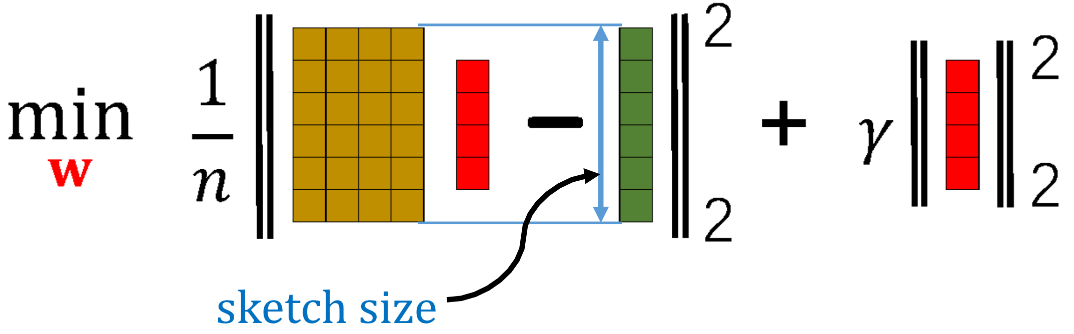
$$\min_{\mathbf{w}} \left\{ f(\mathbf{w}) = \frac{1}{n} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 + \gamma \|\mathbf{w}\|_2^2 \right\}$$



- **Sketched solution: $\mathbf{w}^s$**
- Sketch size $\tilde{O}\left(\frac{d}{\epsilon}\right)$

Approximate Ridge Regression

$$\min_{\mathbf{w}} \left\{ f(\mathbf{w}) = \frac{1}{n} ||\mathbf{X}\mathbf{w} - \mathbf{y}||_2^2 + \gamma ||\mathbf{w}||_2^2 \right\}$$

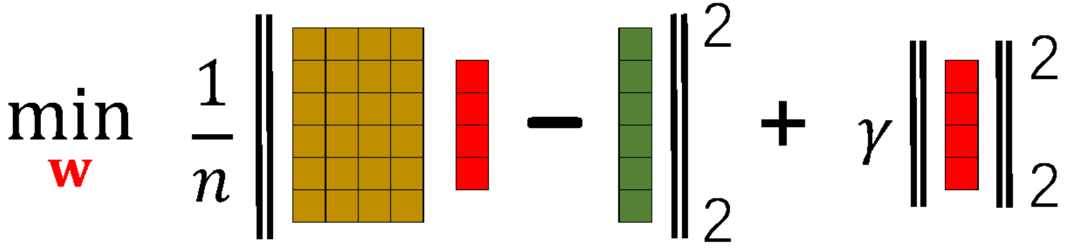


- **Sketched solution: $\mathbf{w}^s$**
- Sketch size $\tilde{O}\left(\frac{d}{\epsilon}\right)$
- $f(\mathbf{w}^s) \leq (1+\epsilon) \min_{\mathbf{w}} f(\mathbf{w})$

Optimization Perspective

Approximate Ridge Regression

$$\min_{\mathbf{w}} \left\{ f(\mathbf{w}) = \frac{1}{n} ||\mathbf{X}\mathbf{w} - \mathbf{y}||_2^2 + \gamma ||\mathbf{w}||_2^2 \right\}$$



Statistical Perspective

- Bias
- Variance

Related Work

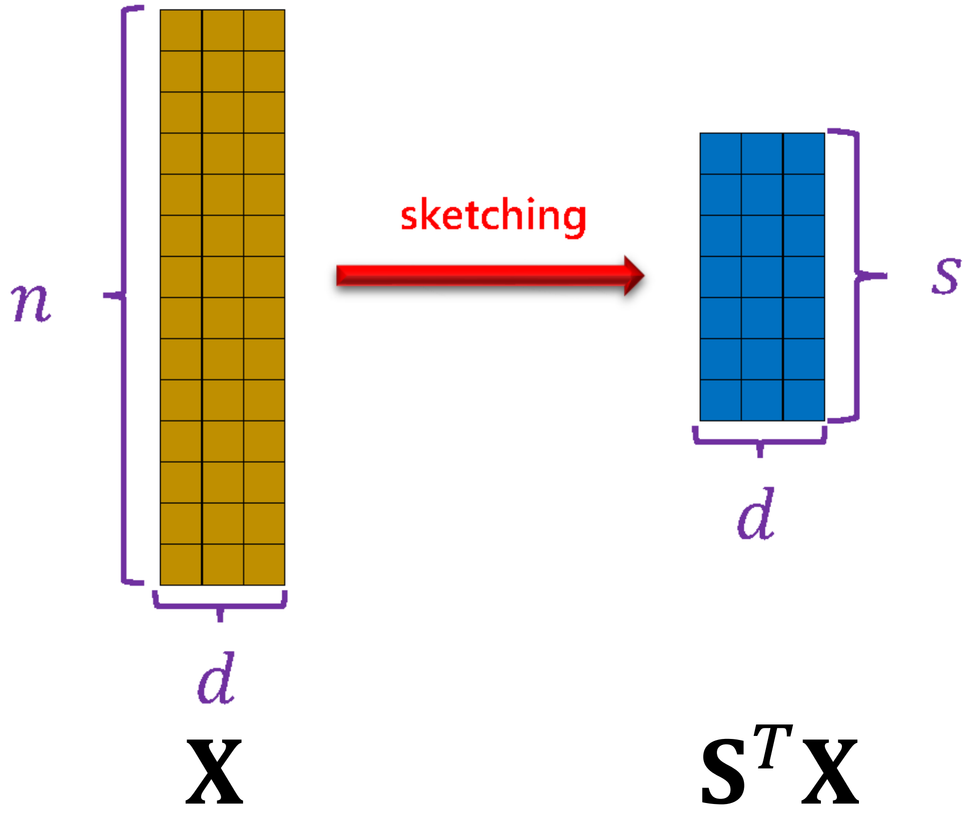
- Least squares regression: $\min_{\mathbf{w}} ||\mathbf{X}\mathbf{w} - \mathbf{y}||_2^2$

Reference

- Drineas, Mahoney, and Muthukrishnan: Sampling algorithms for l2 regression and applications. In *SODA,* 2006.
- Drineas, Mahoney, Muthukrishnan, and Sarlos: Faster least squares approximation. *Numerische Mathematik,* 2011.
- Clarkson and Woodruff: Low rank approximation and regression in input sparsity time. In *STOC*, 2013.
- Ma, Mahoney, and Yu: A statistical perspective on algorithmic leveraging. *Journal of Machine Learning Research*, 2015.
- Pilanci and Wainwright: Iterative Hessian sketch: fast and accurate solution approximation for constrained least squares. *Journal of Machine Learning Research*, 2015.
- Raskutti and Mahoney: A statistical perspective on randomized sketching for ordinary least-squares. *Journal of Machine Learning Research*, 2016.
- Etc …

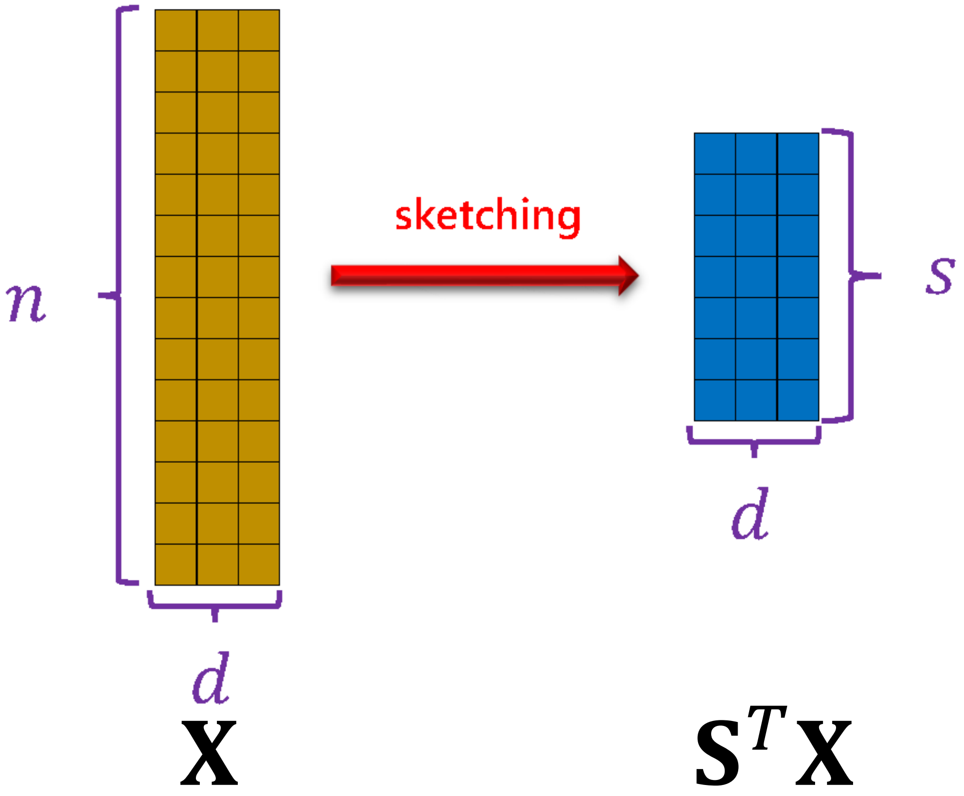
Sketched Ridge Regression

Matrix Sketching



- Turn big matrix into smaller one.
- $\mathbf{X} \in \mathbb{R}^{n \times d}$ ➡ $\mathbf{S}^T\mathbf{X} \in \mathbb{R}^{s \times d}$.
- $\mathbf{S} \in \mathbb{R}^{n \times s}$ is called *sketching matrix, e.g.,*
  - Uniform sampling
  - Leverage score sampling
  - Gaussian projection
  - Subsampled randomized Hadamard transform (SRHT)
  - Count sketch (sparse embedding)
  - Etc.

Matrix Sketching



- Some matrix sketching methods are efficient.
  - Time cost is $o(nds)$ — lower than multiplication.
- Examples:
  - Leverage score sampling: $O(nd \log n)$ time
  - SRHT: $O(nd \log s)$ time

Ridge Regression

- Objective function:

$$f(\mathbf{w}) = \frac{1}{n} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 + \gamma \|\mathbf{w}\|_2^2$$

- Optimal solution:

$$\mathbf{w}^\star = \underset{\mathbf{w}}{\operatorname{argmin}} f(\mathbf{w})$$
$$= (\mathbf{X}^T\mathbf{X} + n\gamma\mathbf{I}_d)^\dagger (\mathbf{X}^T\mathbf{y})$$

- Time cost: $O(nd^2)$ or $O(ndt)$

Sketched Ridge Regression

- Goal: *efficiently* and *approximately* solve

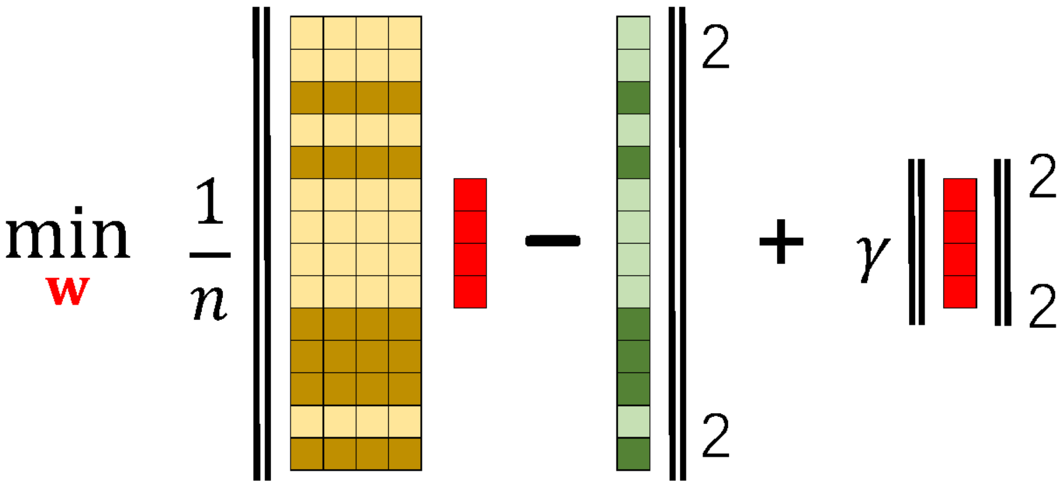
$$\underset{\mathbf{w}}{\operatorname{argmin}} \left\{ f(\mathbf{w}) = \frac{1}{n} ||\mathbf{X}\mathbf{w} - \mathbf{y}||_2^2 + \gamma ||\mathbf{w}||_2^2 \right\}.$$

Sketched Ridge Regression

- Goal: *efficiently* and *approximately* solve

$$\operatorname*{argmin}_{\mathbf{w}} \left\{ f(\mathbf{w}) = \frac{1}{n} ||\mathbf{X}\mathbf{w} - \mathbf{y}||_2^2 + \gamma ||\mathbf{w}||_2^2 \right\}.$$

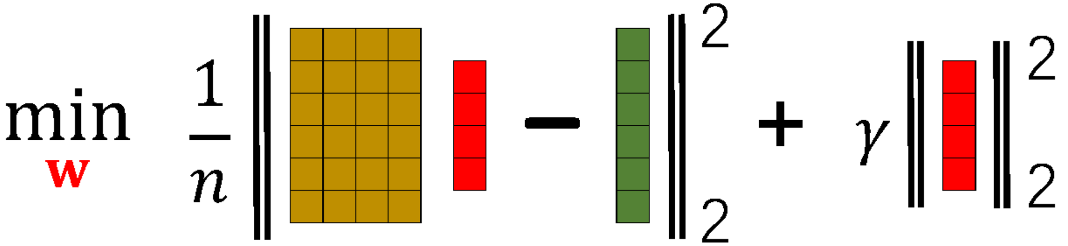
- Approach: reduce the size of $\mathbf{X}$ and $\mathbf{y}$ by matrix sketching.



Sketched Ridge Regression

- Sketched solution:

$$\mathbf{w}^{s} = \operatorname*{argmin}_{\mathbf{w}} \left\{ \frac{1}{n} \left|\left| \mathbf{S}^{T}\mathbf{X}\mathbf{w} - \mathbf{S}^{T}\mathbf{y} \right|\right|_2^2 + \gamma \left|\left| \mathbf{w} \right|\right|_2^2 \right\}$$
$$= \left(\mathbf{X}^{T}\mathbf{S}\mathbf{S}^{T}\mathbf{X} + n\gamma \mathbf{I}_d\right)^{\dagger} \left(\mathbf{X}^{T}\mathbf{S}\mathbf{S}^{T}\mathbf{y}\right)$$



Sketched Ridge Regression

- Sketched solution:

$$\mathbf{w}^{\mathrm{s}} = \operatorname*{argmin}_{\mathbf{w}} \left\{ \frac{1}{n} \left|\left| \mathbf{S}^T \mathbf{X} \mathbf{w} - \mathbf{S}^T \mathbf{y} \right|\right|_2^2 + \gamma \left|\left| \mathbf{w} \right|\right|_2^2 \right\}$$

$$= \left(\mathbf{X}^T \mathbf{S} \mathbf{S}^T \mathbf{X} + n\gamma \mathbf{I}_d \right)^\dagger \left(\mathbf{X}^T \mathbf{S} \mathbf{S}^T \mathbf{y} \right)$$

- Time: $O(sd^2) + T_s$
  - T_s is the cost of sketching $\mathbf{S}^T \mathbf{X}$
  - E.g. $T_s = O(nd \log s)$ for SRHT.
  - E.g. $T_s = O(nd \log n)$ for leverage score sampling.

Theory: Optimization Perspective

Optimization Perspective

- Recall the objective function $f(\mathbf{w}) = \frac{1}{n}||\mathbf{X}\mathbf{w} - \mathbf{y}||_2^2 + \gamma||\mathbf{w}||_2^2$.
- Bound $f(\mathbf{w}^s) - f(\mathbf{w}^\star)$.
- $\frac{1}{n}||\mathbf{X}\mathbf{w}^s - \mathbf{X}\mathbf{w}^\star||_2^2 \leq f(\mathbf{w}^s) - f(\mathbf{w}^\star)$.

Optimization Perspective

For the sketching methods

- SRHT or leverage sampling with $s = \tilde{O}\left(\frac{\beta d}{\epsilon}\right)$,
- uniform sampling with $s = O\left(\frac{\mu \beta d \log d}{\epsilon}\right)$,

$f(\mathbf{w}^s) - f(\mathbf{w}^\star) \leq \epsilon f(\mathbf{w}^\star)$ holds w.p. 0.9.

- $\mathbf{X} \in \mathbb{R}^{n \times d}$: the design matrix
- γ: the regularization parameter
- $\beta = \frac{\|\mathbf{X}\|_2^2}{n\gamma + \|\mathbf{X}\|_2^2} \in (0, 1]$
- $\mu \in \left[1, \frac{n}{d}\right]$: the row coherence of $\mathbf{X}$

Optimization Perspective

For the sketching methods

- SRHT or leverage sampling with $s = \tilde{O}\left(\frac{\beta d}{\epsilon}\right)$,
- uniform sampling with $s = O\left(\frac{\mu\, \beta d \log d}{\epsilon}\right)$,

$f(\mathbf{w}^s) - f(\mathbf{w}^\star) \leq \epsilon f(\mathbf{w}^\star)$ holds w.p. 0.9.

$$\Longrightarrow \quad \frac{1}{n}\left|\left|\mathbf{X}\mathbf{w}^s - \mathbf{X}\mathbf{w}^\star\right|\right|_2^2 \leq \epsilon f(\mathbf{w}^\star).$$

- $\mathbf{X} \in \mathbb{R}^{n \times d}$: the design matrix
- γ: the regularization parameter
- $\beta = \frac{||\mathbf{X}||_2^2}{n\gamma + ||\mathbf{X}||_2^2} \in (0, 1]$
- $\mu \in \left[1, \frac{n}{d}\right]$: the row coherence of $\mathbf{X}$

Theory: Statistical Perspective

Statistical Model

- $\mathbf{X} \in \mathbb{R}^{n \times d}$: fixed design matrix
- $\mathbf{w}_0 \in \mathbb{R}^d$: the *true* and *unknown* model
- $\mathbf{y} = \mathbf{X}\mathbf{w}_0 + \boldsymbol{\delta}$: observed response vector
  - $\delta_1, \cdots, \delta_n$ are random noise
  - $\mathbb{E}[\boldsymbol{\delta}] = \mathbf{0}$ and $\mathbb{E}[\boldsymbol{\delta}\boldsymbol{\delta}^T] = \xi^2 \mathbf{I}_n$

Bias-Variance Decomposition

- Risk: $R(\mathbf{w}) = \frac{1}{n}\mathbb{E}\big||\mathbf{X}\mathbf{w} - \mathbf{X}\mathbf{w}_0|\big|_2^2$
  - $\mathbb{E}$ is taken w.r.t. the random noise $\boldsymbol{\delta}$.

Bias-Variance Decomposition

- Risk: $R(\mathbf{w}) = \frac{1}{n} \mathbb{E} \left\| \mathbf{X}\mathbf{w} - \mathbf{X}\mathbf{w}_0 \right\|_2^2$
  - $\mathbb{E}$ is taken w.r.t. the random noise $\boldsymbol{\delta}$.
  - Risk measures prediction error.

Bias-Variance Decomposition

- Risk: $R(\mathbf{w}) = \frac{1}{n}\mathbb{E}\left|\left|\mathbf{Xw} - \mathbf{Xw}_0\right|\right|_2^2$
- $\mathrm{R}(\mathbf{w}) = \mathrm{bias}^2(\mathbf{w}) + \mathrm{var}(\mathbf{w})$

Bias-Variance Decomposition

- Risk: $R(\mathbf{w}) = \frac{1}{n}\mathbb{E}\left\|\mathbf{X}\mathbf{w} - \mathbf{X}\mathbf{w}_0\right\|_2^2$
- $\mathrm{R}(\mathbf{w}) = \mathrm{bias}^2(\mathbf{w}) + \mathrm{var}(\mathbf{w})$

Optimal Solution

- $\mathrm{bias}(\mathbf{w}^\star) = \gamma\sqrt{n}\left\|(\mathbf{\Sigma}^2 + n\gamma\mathbf{I}_d)^{-1}\mathbf{\Sigma}\mathbf{V}^T\mathbf{w}_0\right\|_2,$
- $\mathrm{var}(\mathbf{w}^\star) = \frac{\xi^2}{n}\left\|(\mathbf{I}_d + n\gamma\mathbf{\Sigma}^{-2})^{-1}\right\|_2^2,$

Sketched Solution

- $\mathrm{bias}(\mathbf{w}^s) = \gamma\sqrt{n}\left\|(\mathbf{\Sigma}\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U}\mathbf{\Sigma} + n\gamma\mathbf{I}_d)^\dagger\mathbf{\Sigma}\mathbf{V}^T\mathbf{w}_0\right\|_2,$
- $\mathrm{var}(\mathbf{w}^s) = \frac{\xi^2}{n}\left\|(\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U} + n\gamma\mathbf{\Sigma}^{-2})^\dagger\mathbf{U}^T\mathbf{S}\mathbf{S}^T\right\|_2^2,$

- Here $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ is the SVD.

Statistical Perspective

For the sketching methods

- SRHT or leverage sampling with $s = \tilde{O}\left(\frac{d}{\epsilon^2}\right)$,
- uniform sampling with $s = O\left(\frac{\mu\, d \log d}{\epsilon^2}\right)$,

- $\mathbf{X} \in \mathbb{R}^{n \times d}$: the design matrix
- $\mu \in \left[1, \frac{n}{d}\right]$: the row coherence of $\mathbf{X}$

the followings hold w.p. 0.9:

$$1 - \epsilon \leq \frac{\mathrm{bias}(\mathbf{w}^s)}{\mathrm{bias}(\mathbf{w}^\star)} \leq 1 + \epsilon,$$

Good!

$$(1 - \epsilon)\frac{n}{s} \leq \frac{\mathrm{var}(\mathbf{w}^s)}{\mathrm{var}(\mathbf{w}^\star)} \leq (1 + \epsilon)\frac{n}{s}.$$

Bad! Because $n \gg s$.

Statistical Perspective

For the sketching methods

- SRHT or leverage sampling with $s = \tilde{O}\left(\frac{d}{\epsilon^2}\right)$,
- uniform sampling with $s = O\left(\frac{\mu\, d \log d}{\epsilon^2}\right)$,

the followings hold w.p. 0.9:

$$1 - \epsilon \leq \frac{\mathrm{bias}(\mathbf{w}^s)}{\mathrm{bias}(\mathbf{w}^\star)} \leq 1 + \epsilon,$$

$$(1 - \epsilon)\frac{n}{s} \leq \frac{\mathrm{var}(\mathbf{w}^s)}{\mathrm{var}(\mathbf{w}^\star)} \leq (1 + \epsilon)\frac{n}{s}.$$

- $\mathbf{X} \in \mathbb{R}^{n \times d}$: the design matrix
- $\mu \in \left[1, \frac{n}{d}\right]$: the row coherence of $\mathbf{X}$

If $\mathbf{y}$ is noisy

$\Rightarrow$ variance dominates bias

$\Rightarrow$ $R(\mathbf{w}^s) \gg R(\mathbf{w}^\star)$.

Conclusions

- Use sketched solution to initialize numerical optimization.
  - $\mathbf{X}\mathbf{w}^{\mathrm{s}}$ is close to $\mathbf{X}\mathbf{w}^{\star}$.

Optimization Perspective

Conclusions

- Use sketched solution to initialize numerical optimization.
  - $\mathbf{X}\mathbf{w}^{s}$ is close to $\mathbf{X}\mathbf{w}^{\star}$.

Optimization Perspective

- $\mathbf{w}^{(t)}$: output of the t-th iteration of CG algorithm.
- $\frac{\left\|\mathbf{X}\mathbf{w}^{(t)} - \mathbf{X}\mathbf{w}^{\star}\right\|_2^2}{\left\|\mathbf{X}\mathbf{w}^{(0)} - \mathbf{X}\mathbf{w}^{\star}\right\|_2^2} \leq 2\left(\frac{\sqrt{\kappa(\mathbf{X}^T\mathbf{X})}-1}{\sqrt{\kappa(\mathbf{X}^T\mathbf{X})}+1}\right)^t.$
- Initialization is important.

Conclusions

- Use sketched solution to initialize numerical optimization.
  - $\mathbf{X}\mathbf{w}^{s}$ is close to $\mathbf{X}\mathbf{w}^{\star}$.

Optimization Perspective

- Never use sketched solution to replace the optimal solution.
  - Much higher variance ➔ bad generalization.

Statistical Perspective

Model Averaging

Model Averaging

- Independently draw $\boldsymbol{S}_1, \cdots, \boldsymbol{S}_g$.
- Compute the sketched solutions $\mathbf{w}_1^s, \cdots, \mathbf{w}_g^s$.
- Model averaging: $\mathbf{w}^s = \frac{1}{g} \sum_{i=1}^{g} \mathbf{w}_i^s$.

Optimization Perspective

- For sufficiently large s,

$$\frac{f(\mathbf{w}_1^{s}) - f(\mathbf{w}^\star)}{f(\mathbf{w}^\star)} \leq \epsilon \quad \text{holds w.h.p.}$$

Without model averaging

Optimization Perspective

- For sufficiently large s,

$$\frac{f(\mathbf{w}_1^{\mathrm{s}})-f(\mathbf{w}^\star)}{f(\mathbf{w}^\star)} \leq \epsilon \quad \text{holds w.h.p.}$$

Without model averaging

- Using the **same** matrix sketching and **same** s,

$$\frac{f(\mathbf{w}^{\mathrm{s}})-f(\mathbf{w}^\star)}{f(\mathbf{w}^\star)} \leq \frac{\epsilon}{g} + \epsilon^2 \quad \text{holds w.h.p.}$$

With model averaging

Optimization Perspective

- For sufficiently large s,

$$\frac{f(\mathbf{w}_1^s)-f(\mathbf{w}^\star)}{f(\mathbf{w}^\star)} \leq \boxed{\epsilon} \quad \text{holds w.h.p.}$$

Without model averaging

- Using the **same** matrix sketching and **same** s,

$$\frac{f(\mathbf{w}^s)-f(\mathbf{w}^\star)}{f(\mathbf{w}^\star)} \leq \boxed{\frac{\epsilon}{g} + \epsilon^2} \quad \text{holds w.h.p.}$$

With model averaging

Optimization Perspective

- For sufficiently large s,

$$\frac{f(\mathbf{w}_1^{\mathrm{s}})-f(\mathbf{w}^\star)}{f(\mathbf{w}^\star)} \leq \boxed{\epsilon} \quad \text{holds w.h.p.}$$

Without model averaging

- Using the **same** matrix sketching and **same** s,

$$\frac{f(\mathbf{w}^{\mathrm{s}})-f(\mathbf{w}^\star)}{f(\mathbf{w}^\star)} \leq \boxed{\frac{\epsilon}{g} + \epsilon^2} \quad \text{holds w.h.p.}$$

With model averaging

If $s \gg d$ $\Longrightarrow$ ϵ^2 is very small $\Longrightarrow$ error bound $\propto \frac{\epsilon}{g}$.

Statistical Perspective

- Risk: $R(\mathbf{w}) = \frac{1}{n}\mathbb{E}\left|\left|\mathbf{Xw} - \mathbf{Xw}_0\right|\right|_2^2 = \mathrm{bias}^2(\mathbf{w}) + \mathrm{var}(\mathbf{w})$
- Model averaging :
  - $\mathrm{bias}(\mathbf{w}^s) = \gamma\sqrt{n}\left\|\frac{1}{g}\sum_{i=1}^{g}\left(\mathbf{\Sigma U}^T\mathbf{S}_i\mathbf{S}_i^T\mathbf{U\Sigma} + n\gamma\mathbf{I}_d\right)^{\dagger}\mathbf{\Sigma V}^T\mathbf{w}_0\right\|_2,$
  - $\mathrm{var}(\mathbf{w}^s) = \frac{\xi^2}{n}\left\|\frac{1}{g}\sum_{i=1}^{g}\left(\mathbf{U}^T\mathbf{S}_i\mathbf{S}_i^T\mathbf{U} + n\gamma\mathbf{\Sigma}^{-2}\right)^{\dagger}\mathbf{U}^T\mathbf{S}_i\mathbf{S}_i^T\right\|_2^2.$
  - Here $\mathbf{X} = \mathbf{U\Sigma V}^T$ is the SVD.

Statistical Perspective

- For sufficiently large s, the followings hold w.h.p.:

$$\frac{\mathrm{bias}(\mathbf{w}^{\mathrm{s}})}{\mathrm{bias}(\mathbf{w}^{\star})} \leq 1 + \epsilon \qquad \text{and} \qquad \frac{\mathrm{var}(\mathbf{w}^{\mathrm{s}})}{\mathrm{var}(\mathbf{w}^{\star})} \leq \frac{n}{s}(1 + \epsilon).$$

Without model averaging

Statistical Perspective

- For sufficiently large s, the followings hold w.h.p.:

$$\frac{\mathrm{bias}(\mathbf{w}^{\mathrm{s}})}{\mathrm{bias}(\mathbf{w}^{\star})} \leq 1+\epsilon \quad \text{and} \quad \frac{\mathrm{var}(\mathbf{w}^{\mathrm{s}})}{\mathrm{var}(\mathbf{w}^{\star})} \leq \frac{n}{s}(1+\epsilon).$$

Without model averaging

- Using the **same** sketching methods and **same** s, the followings hold w.h.p.:

$$\frac{\mathrm{bias}(\mathbf{w}^{\mathrm{s}})}{\mathrm{bias}(\mathbf{w}^{\star})} \leq 1+\epsilon \quad \text{and} \quad \frac{\mathrm{var}(\mathbf{w}^{\mathrm{s}})}{\mathrm{var}(\mathbf{w}^{\star})} \lesssim \frac{n}{s}\left(\frac{1}{\sqrt{g}}+\epsilon\right)^{2}$$

With model averaging

Statistical Perspective

- For sufficiently large s, the followings hold w.h.p.:

$$\frac{\mathrm{bias}(\mathbf{w}^{\mathrm{s}})}{\mathrm{bias}(\mathbf{w}^{\star})} \leq 1+\epsilon \quad \text{and} \quad \frac{\mathrm{var}(\mathbf{w}^{\mathrm{s}})}{\mathrm{var}(\mathbf{w}^{\star})} \leq \frac{n}{s}\boxed{(1+\epsilon)}.$$

Without model averaging

- Using the **same** sketching methods and **same** s, the followings hold w.h.p.:

$$\frac{\mathrm{bias}(\mathbf{w}^{\mathrm{s}})}{\mathrm{bias}(\mathbf{w}^{\star})} \leq 1+\epsilon \quad \text{and} \quad \frac{\mathrm{var}(\mathbf{w}^{\mathrm{s}})}{\mathrm{var}(\mathbf{w}^{\star})} \lesssim \frac{n}{s}\boxed{\left(\frac{1}{\sqrt{g}}+\epsilon\right)^{\mathbf{2}}}$$

With model averaging

Statistical Perspective

- For sufficiently large s, the followings hold w.h.p.:

$$\frac{\text{bias}(\mathbf{w}^{\text{s}})}{\text{bias}(\mathbf{w}^{\star})} \leq 1 + \epsilon \quad \text{and} \quad \frac{\text{var}(\mathbf{w}^{\text{s}})}{\text{var}(\mathbf{w}^{\star})} \leq \frac{n}{s} \boxed{(1 + \epsilon)}.$$

Without model averaging

- Using the **same** sketching methods and **same** s, the followings hold w.h.p.:

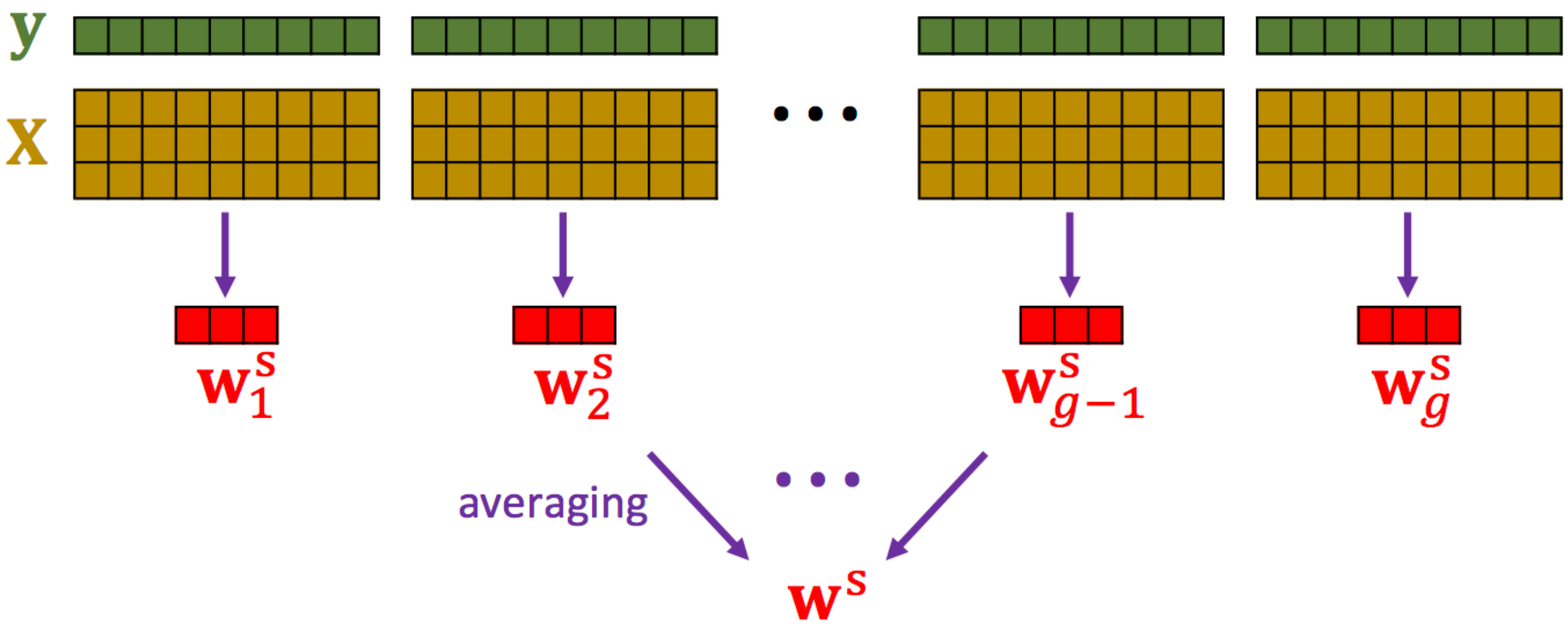
$$\frac{\text{bias}(\mathbf{w}^{\text{s}})}{\text{bias}(\mathbf{w}^{\star})} \leq 1 + \epsilon \quad \text{and} \quad \frac{\text{var}(\mathbf{w}^{\text{s}})}{\text{var}(\mathbf{w}^{\star})} \lesssim \frac{n}{s} \boxed{\left(\frac{1}{\sqrt{g}} + \epsilon\right)^{\mathbf{2}}}$$

With model averaging

If ϵ is small, then $\text{var}(\mathbf{w}^{\text{s}}) \propto \frac{1}{g}$.

Applications to Distributed Optimization

- $(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)$ are (randomly) split among g machines.
- Equivalent to uniform sampling with $s = \frac{n}{g}$.



Optimization Perspective

- **Application to distributed optimization**:
  - If $s = \frac{n}{g} \gg d$, $\mathbf{w}^{\mathrm{s}}$ is very close to $\mathbf{w}^{\star}$ (provably).
  - $\mathbf{w}^{\mathrm{s}}$ is good initialization of distributed optimization algorithms.

Statistical Perspective

- **Application to distributed machine learning**:
  - If $s = \frac{n}{g} \gg d$, then $R(\mathbf{w}^{\mathrm{s}})$ is comparable to $R(\mathbf{w}^{\star})$.
  - If low-precision solution suffices, then $\mathbf{w}^{\mathrm{s}}$ is a good substitute of $\mathbf{w}^{\star}$.
  - One-shot solution.

Thank You!

The paper is at arXiv:1702.04837