Paper Review: Iterative Hessian Sketch: Fast and accurate solution approximation for constrained least-squares

Authors: Mert Pilanci, Martin J. Wainwright

July 26, 2016

Outline

Model Description

- Given a data vector $y \in \mathbb{R}^n$, a data matrix $A \in \mathbb{R}^{n \times d}$ and a convex constraint set $\mathcal{C}$, a constrained least-squares problem can be written as follows

$$x^{LS} = \arg\min_{x \in \mathcal{C}} f(x) \qquad \text{where} \qquad f(x) = \frac{1}{2n}\|Ax - y\|_2^2.$$

- *Classical Least-squares Sketch:* $(A, y) \to (SA, Sy)$, where $S \in \mathbb{R}^{m \times n}$ is the random projection matrix.

$$\widetilde{x} = \arg\min_{x \in \mathcal{C}} \frac{1}{2n}\|SAx - Sy\|_2^2,$$

- Consider the solution aproximation as the prediction norm

$$\|\widetilde{x} - x^{LS}\|_A = \frac{1}{\sqrt{n}}\|A(\widetilde{x} - x^{LS})\|_2.$$

Model Description

- Consider the model

$$y = Ax^* + w, \qquad \text{where } w \sim N(0, \sigma^2 I_n),$$

the data matrix $A \in \mathbb{R}^{n \times d}$ is fixed, x^* belongs to some compact subset $\mathcal{C}_0 \subseteq \mathcal{C}$.

- Quesiton: The minimal projection dimension m required to achieve an error guarantee $\|\widetilde{x} - x^{LS}\|_A \approx \|x^{LS} - x^*\|_A$
- Consider any random matrix $S \in \mathbb{R}^{m \times n}$ satisfying

$$\|\mathbb{E}\left[S^T (SS^T)^{-1} S\right]\|_{\text{op}} \leq \eta \, \frac{m}{n},$$

where η is a constant independent of n and m, such random matrix includes following examples:

Examples of Random Sketch Matrix

- Sub Gaussian projection matrix, e.g., i.i.d. Gaussian entries;

Examples of Random Sketch Matrix

- Sub Gaussian projection matrix, e.g., i.i.d. Gaussian entries;
- Sub-sampling projection (sub-sampling s rows of the identity matrix $I_{n\times n}$ without replacement);

Examples of Random Sketch Matrix

- Sub Gaussian projection matrix, e.g., i.i.d. Gaussian entries;
- Sub-sampling projection (sub-sampling s rows of the identity matrix $I_{n\times n}$ without replacement);
- Randomized orthogonal system (ROS) projection (each row randomly sampled in rescaled orthonormal matrix).

Sub-Optimality based on (SA, Sy)

Theorem 1 (Sub-optimality)

For any random sketching matrix $S \in \mathbb{R}^{m \times n}$ satisfying the above condition, any estimator $(SA, Sy) \to x^\dagger$ has MSE lower bounded as

$$\sup_{x^* \in \mathcal{C}_0} \mathbb{E}_{S,w}\left[\|x^\dagger - x^*\|_A^2\right] \geq \frac{\sigma^2}{128\,\eta} \frac{\log(\frac{1}{2}M_{1/2})}{\min\{m,n\}} \tag{1}$$

where $\mathbb{B}_A(1)$ denotes the unit ball defined by the norm $\|\cdot\|_A$, and $M_{1/2}$ is the $1/2$-packing number of $\mathcal{C}_0$ in the semi-norm $\|\cdot\|_A$.

Example 1 (Sub-optimality for ordinary least-squares)

Consider when $\mathcal{C} = \mathbb{R}^d$ and $rank(A) = d$, the least-squares solution x^{LS} has its prediction mean-squared error

$$\mathbb{E}\left[\|x^{LS} - x^*\|_A^2\right] \precsim \frac{\sigma^2 d}{n}. \tag{2}$$

On the other hand, with the choice $\mathcal{C}_0 = \mathbb{B}_2(1)$, a $1/2$-packing with $M = 2^d$ elements can be constructed. Theorem 1 implies that any estimator $x^\dagger$ based on (SA, Sy) has its prediction MSE lower bounded as

$$\mathbb{E}_{S,w}\left[\|\widehat{x} - x^*\|_A^2\right] \succsim \frac{\sigma^2\, d}{\min\{m, n\}}. \tag{3}$$

Thus, the sketch dimension m must grow proportionally to n. Sub-optimality!

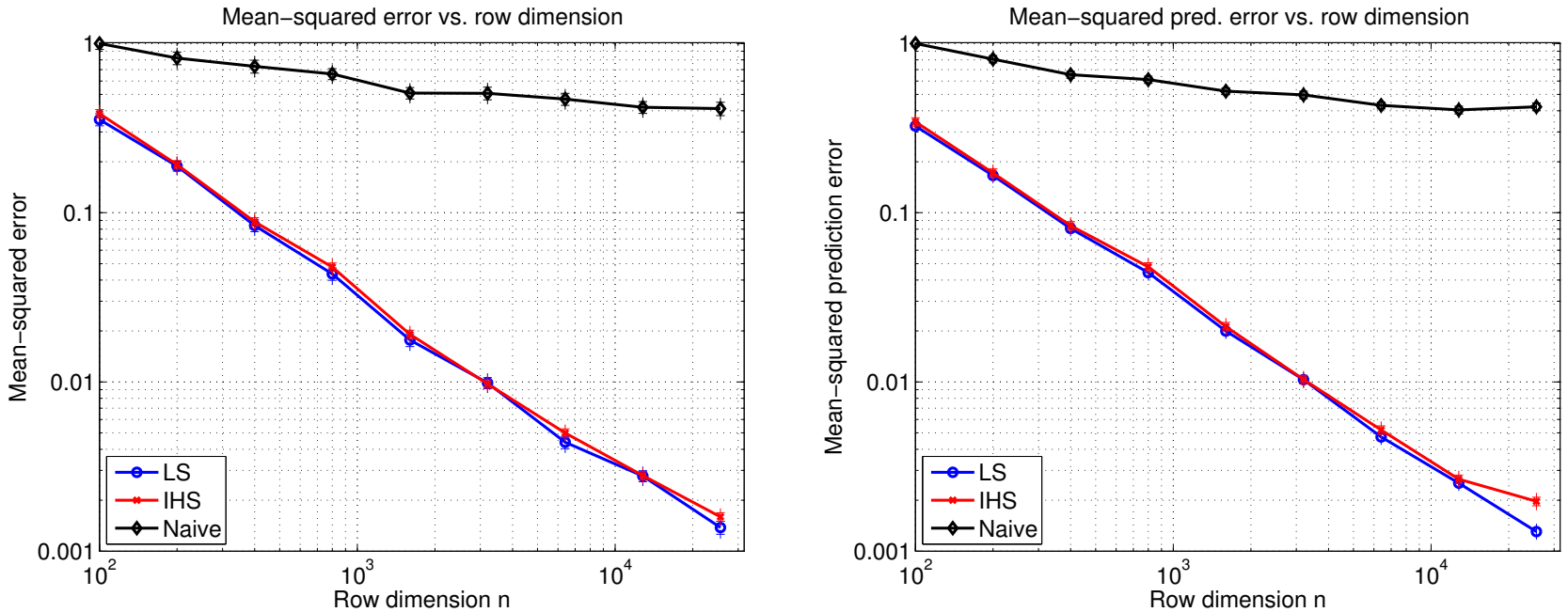

Figure 1: Plots of mean-squared error versus the row dimension $n \in \{100, 200, 400, \ldots, 25600\}$ for unconstrained least-squares in dimension $d = 10$. The blue curves correspond to the error $x^{LS} - x^*$ of the unsketched least-squares estimate. Red curves correspond to the IHS method applied for $N = 1 + \lceil \log(n) \rceil$ rounds using a sketch size $m = 7d$. Black curves correspond to the naive sketch applied using $M = Nm$ projections in total, corresponding to the same number used in all iterations of the IHS algorithm. (a) Error $\|\widetilde{x} - x^*\|_2^2$. (b) Prediction error $\|\widetilde{x} - x^*\|_A^2 = \frac{1}{n}\|A(\widetilde{x} - x^*)\|_2^2$.

Example 2 (Sub-optimality for sparse linear models)

Consider the ℓ_1-ball constrain $\mathcal{C} = \{x \in \mathbb{R}^n \mid \|x\|_1 \leq R\}$, the corresponding lasso estimator has MSE at most

$$\mathbb{E}\left[\|x^{LS} - x^*\|_A^2\right] \precsim \frac{\sigma^2 s \log\left(\frac{ed}{s}\right)}{n}. \tag{4}$$

On the other hand, the $\frac{1}{2}$-packing number M can be lower bounded as $\log M \succsim s \log\left(\frac{ed}{s}\right)$; By Theorem 1, any estimator $x^\dagger$ based on (SA, Sy) has mean-squared error lower bounded as

$$\mathbb{E}_{w,S}\left[\|x^\dagger - x^*\|_A^2\right] \succsim \frac{\sigma^2 s \log\left(\frac{ed}{s}\right)}{\min\{m, n\}}. \tag{5}$$

Again, the sketch dimension m must grow proportionally to n.
Sub-optimality!

Hessian Sketch based on $(SA, A^T y) \in \mathbb{R}^{m \times d} \times \mathbb{R}^d$

- The Hessian sketch estimator can be obtained by

$$\widehat{x} = \arg\min_{x \in \mathcal{C}} \frac{1}{2} \|SAx\|_2^2 - \langle A^T y, x \rangle.$$

Proposition 1: Bounds on Hessian sketch

For any convex set $\mathcal{C}$ and any sketching matrix $S \in \mathbb{R}^{m \times n}$, the Hessian sketch solution $\widehat{x}$ satisfies the bound

$$\|\widehat{x} - x^{LS}\|_A \leq \frac{Z_2}{Z_1} \|x^{LS}\|_A.$$

Hessian Sketch

$$Z_1(S) = \inf_{v \in \mathcal{K}^{LS} \cap \mathcal{S}^{n-1}} \frac{1}{m} \|Sv\|_2^2 \quad \text{and}$$

$$Z_2(S) = \sup_{v \in \mathcal{K}^{LS} \cap \mathcal{S}^{n-1}} \left| \langle u, \, (\frac{S^T S}{m} - I_n) \, v \rangle \right|,$$

u is a fixed unit-norm vector, and the transformed tangent cone is defined as

$$\mathcal{K}^{LS} = \left\{ v \in \mathbb{R}^n \mid v = t \, A(x - x^{LS}) \quad \text{for some } t \geq 0 \text{ and } x \in \mathcal{C} \right\}.$$

Hessian Sketch

Consider the "good event" with a given tolerance parameter $\rho \in (0, \frac{1}{2}]$,

$$\mathcal{E}(\rho) = \left\{ Z_1 \geq 1 - \rho, \text{ and } Z_2 \leq \frac{\rho}{2} \right\}.$$

Conditioned on this event, Proposition 1 implies that

$$\|\widehat{x} - x^{LS}\|_A \leq \frac{\rho}{2\,(1-\rho)} \|x^{LS}\|_A \;\leq\; \rho \|x^{LS}\|_A.$$

- Remark 1: To obtain a high probability on event $\mathcal{E}(\rho)$, we need to choose the projection dimension m.
- Remark 2: Note that the error vector $\widehat{v} = A(\widehat{x} - x^{LS})$ of interest belongs to $\mathcal{K}^{LS}$. The "size" of the cone $\mathcal{K}^{LS}$?
- Remark 3: The Hessian sketch on its own does not provide an optimal approximation to the LS solution. Need iteration!

Answer to Remark 1

Lemma 1 (Sufficient conditions on sketch dimension)

(a) For sub-Gaussian sketch matrices, given a sketch size $m > \frac{c_0}{\rho^2}\mathcal{W}^2(\mathcal{K}^{LS})$, we have

$$\mathbb{P}\big[\mathcal{E}(\rho)] \geq 1 - c_1 e^{-c_2 m \delta^2}. \tag{8a}$$

(b) For randomized orthogonal system (ROS) sketches over the class of self-bounding cones, given a sketch size $m > \frac{c_0 \log^4(D)}{\rho^2}\mathcal{W}^2(\mathcal{K}^{LS})$, we have

$$\mathbb{P}\big[\mathcal{E}(\rho)] \geq 1 - c_1 e^{-c_2 \frac{m\rho^2}{\log^4(D)}}. \tag{8b}$$

Answer to Remark 2

Gaussian Width is defined as

$$\mathcal{W}(\mathcal{K}^{LS}) = \mathbb{E}_g\big[\sup_{v \in \mathcal{K}^{LS} \cap \mathbb{B}_2(1)} |\langle g, v \rangle|\big],$$

where $g \sim N(0, I_n)$ is a standard Gaussian vector.

- The Gaussian width measures how well the vecotrs in $\mathcal{K}^{LS}$ can align with a randomly chosen direction; i.e., it is a measure of the "size" of the set.
- For $\mathcal{K}^{LS}$, it can be proved that $\mathcal{W}(\mathcal{K}^{LS}) \leq \sqrt{rank(A)}$.

Answer to Remark 3

- Recall in Example 1, the least-squares solution x^{LS} has its prediction mean-squared error

$$\mathbb{E}\left[\|x^{LS} - x^*\|_A^2\right] \precsim \frac{\sigma^2 d}{n}.$$

- The Hessian sketch in Proposition 1 implies

$$\|\widehat{x} - x^{LS}\|_A \leq \frac{\rho}{2\,(1-\rho)}\|x^{LS}\|_A \;\leq\; \rho\|x^{LS}\|_A.$$

- To match the approximation error, we need to choose $\rho \asymp \sqrt{\frac{\sigma^2 d}{n}}$, which requires $m \asymp n$ by Lemma 1.

Iterative Hessian Sketch: Intuition

$$\|\widehat{x} - x^{LS}\|_A \leq \frac{\rho}{2\,(1-\rho)}\|x^{LS}\|_A \;\leq\; \rho\|x^{LS}\|_A.$$

- Hessian sketch returns an estimate with error within a ρ-factor of $\|x^{LS}\|_A$.
- Given the current iterate x^t, suppose that we can construct a new least-squares problem for which the optimal solution is $x^{LS} - x^t$.

$$\arg\min_{x \in \mathcal{C}} \frac{1}{2}\|A(x - x^t)\|_2^2 - \langle A^T(y - Ax^t),\, x - x^t\rangle,$$

- Applying the Hessian sketch to this new problem, the new iterate x^{t+1} whose distance to x^{LS} can be reduced to $\rho\|x^{LS} - x^t\|_A$.

Iterative Hessian Sketch: Algorithm

Formally, the iterative Hessian sketch algorithm takes the following form:

Given an iteration number $N \geq 1$:

(1) Initialize at $x^0 = 0$.

(2) For iterations $t = 0, 1, 2, \ldots, N-1$, generate an independent sketch matrix $S^{t+1} \in \mathbb{R}^{m \times n}$, and perform the update

$$x^{t+1} = \arg\min_{x \in \mathcal{C}} \frac{1}{2m} \|S^{t+1} A(x - x^t)\|_2^2 - \langle A^T(y - Ax^t),\, x \rangle.$$

(3) Return the estimate $\widehat{x} = x^N$.

Theorem 2 (Guarantees for iterative Hessian sketch)

The final solution $\widehat{x} = x^N$ satisfies the bound

$$\|\widehat{x} - x^{LS}\|_A \leq \left\{ \prod_{t=1}^{N} \frac{Z_2(S^t)}{Z_1(S^t)} \right\} \|x^{LS}\|_A.$$

Consequently, conditioned on the event $\cap_{t=1}^{N} \mathcal{E}^t(\rho)$ for some $\rho \in (0, 1/2)$, we have

$$\|\widehat{x} - x^{LS}\|_A \leq \rho^N \|x^{LS}\|_A.$$

- Lemma 1 can be combind with Theorem 2 to ensure that the compound event $\cap_{t=1}^{N} \mathcal{E}^t(\rho)$ holds with high probability.

Theorem 2 allows us to specify, for a given target accuracy $\varepsilon \in (0, 1)$, the number of iterations required.

Corollary 1

Fix some $\rho \in (0, 1/2)$, and choose a sketch dimension $m > \frac{c_0 \log^4(D)}{\rho^2} \mathcal{W}^2(\mathcal{K}^{LS})$. If we apply the IHS algorithm for $N(\rho, \varepsilon) = 1 + \frac{\log(1/\varepsilon)}{\log(1/\rho)}$ steps, then the output $\widehat{x} = x^N$ satisfies the bound

$$\frac{\|\widehat{x} - x^{LS}\|_A}{\|x^{LS}\|_A} \leq \varepsilon \tag{10}$$

with probability at least $1 - c_1 N(\rho, \varepsilon) e^{-c_2 \frac{m\rho^2}{\log^4(D)}}$.

Unconstrained Least Squares

Consider ordinary least squares with $\mathcal{W}^2(\mathcal{K}^{LS}) \asymp d$

Corollary 2

For some given $\rho \in (0, 1/2)$, suppose that we run the IHS algorithm for $N = 1 + \lceil \frac{\log \sqrt{n} \, \frac{\|x^{LS}\|_A}{\sigma\sqrt{d}}}{\log(1/\rho)} \rceil$ iterations using $m = \frac{c_0}{\rho^2} d$ projections per round. Then the output $\widehat{x}$ satisfies the bounds

$$\|\widehat{x} - x^{LS}\|_A \leq \sqrt{\frac{\sigma^2 d}{n}}, \qquad \text{and} \qquad \|x^N - x^*\|_A \leq \sqrt{\frac{\sigma^2 d}{n}} + \|x^{LS} - x^*\|_A \tag{11}$$

with probability greater than $1 - c_1 \, N \, e^{-c_2 \frac{m\rho^2}{\log^4(d)}}$.

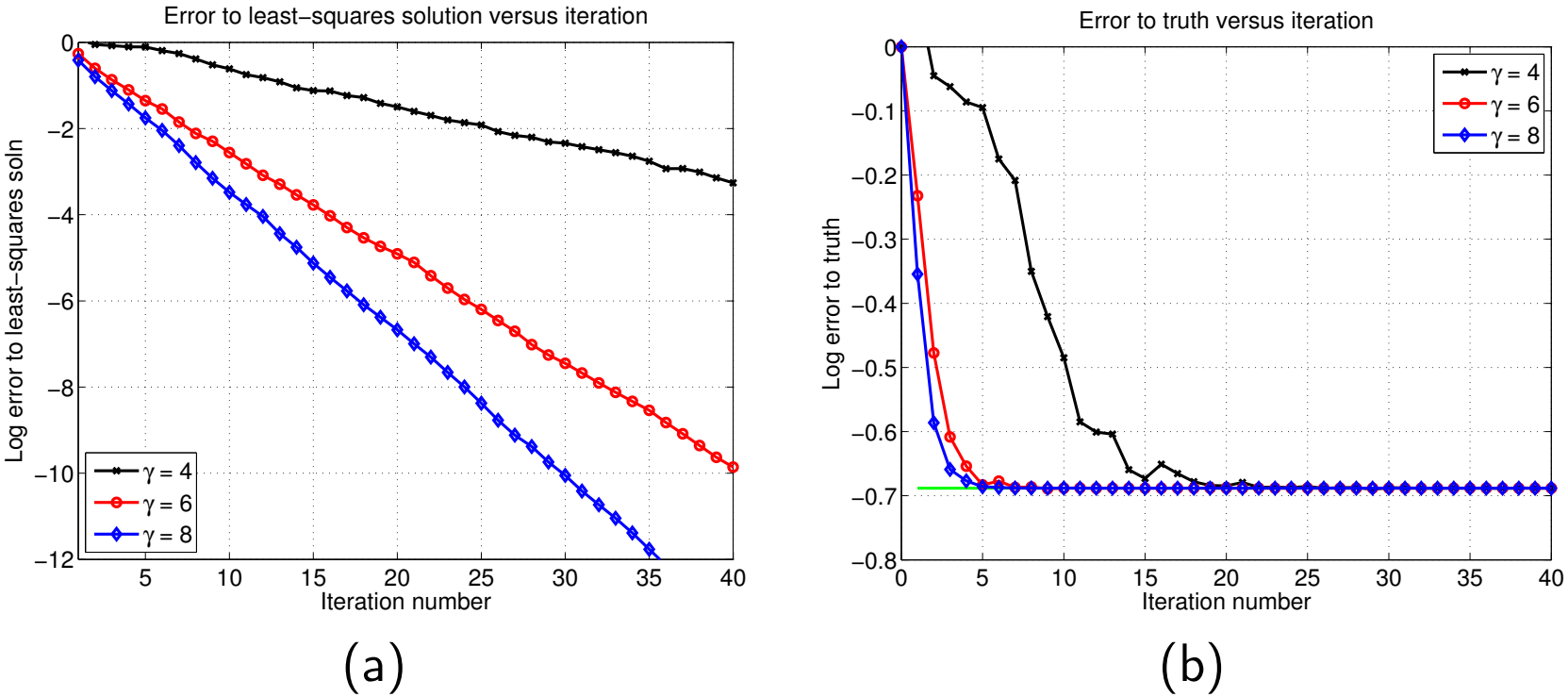

Figure 2: Simulations of the IHS algorithm for an unconstrained least-squares problem with noise variance $\sigma^2 = 1$, and of dimensions $(d, n) = (200, 6000)$. Simulations based on sketch sizes $m = \gamma d$, for a parameter $\gamma > 0$ to be set. (a) Plots of the log error $\|x^t - x^{LS}\|_A$ versus the iteration number t. (b) Plots of the log error $\|x^t - x^*\|_A$ versus the iteration number t.

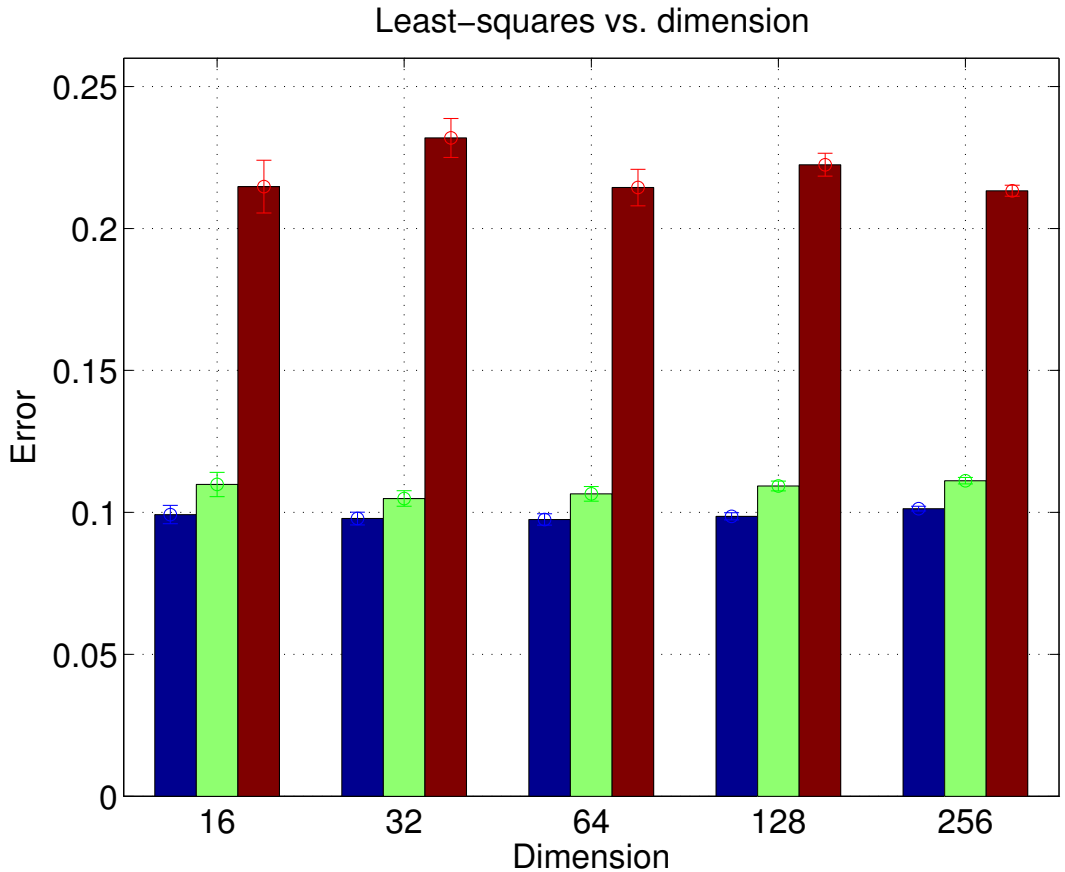

Figure 3: $d \in \{16, 32, 64, 128, 256\}$, sample size $n = 100d$. The initial least-squares solution has error $\|x^{LS} - x^*\|_A \approx 0.10$, as shown by the blue bars. We then ran the IHS algorithm for $N = 4$ iterations with a sketch size $m = 6d$. As shown by the green bars, these sketched solutions show an error $\|\widehat{x} - x^*\|_A \approx 0.11$ independently of dimension, consistent with the predictions of Corollary 2. Finally, the red bars show the error in the classical sketch, based on a sketch size $M = Nm = 24d$.

Sparse Least Squares with l_1-constrain

consider the convex program

$$x^{LS} = \arg \min_{\|x\|_1 \leq R} \{ \frac{1}{2} \|y - Ax\|_2^2 \},$$

where $R > 0$ is a user-defined radius, with $\mathcal{W}^2(\mathcal{K}^{LS}) \asymp s \log d$.

Corollary 3

For the sparse linear regression problems, suppose that we run the IHS algorithm for $N = 1 + \lceil \frac{\log \sqrt{n} \, \frac{\|x^{LS}\|_A}{\sigma}}{\log(1/\rho)} \rceil$ iterations using $m = \frac{c_0}{\rho^2} s \log \left(\frac{ed}{s} \right)$ projections per round. Then with probability greater than $1 - c_1 \, N \, e^{-c_2 \frac{m\rho^2}{\log^4(d)}}$, the output $\widehat{x}$ satisfies the bounds

$$\|\widehat{x} - x^{LS}\|_A \leq \sqrt{\frac{\sigma^2 s \log \left(\frac{ed}{s} \right)}{n}} \text{and} \|x^N - x^*\|_A \leq \sqrt{\frac{\sigma^2 s \log \left(\frac{ed}{s} \right)}{n}} + \|x^{LS} - x^*\|_A.$$

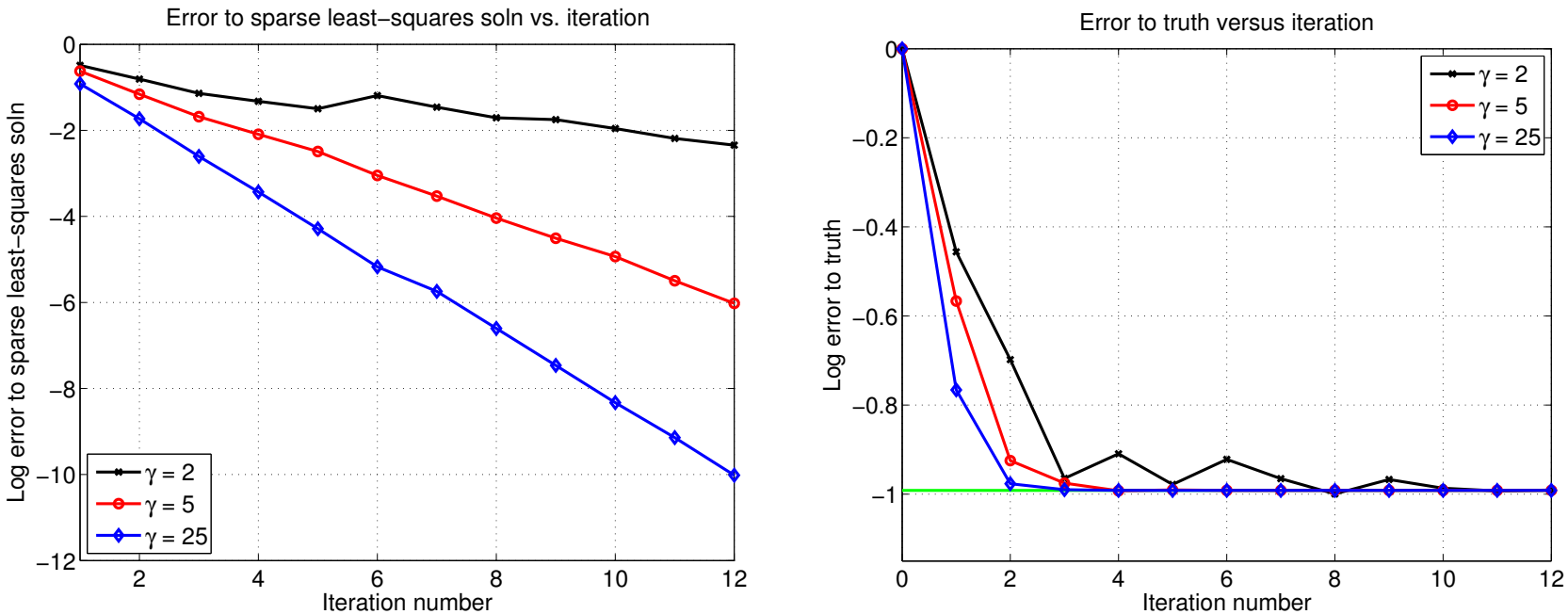

Figure 4: $(d, n, s) = (256, 8872, 32)$. Simulations based on sketch sizes $m = \gamma s \log d$, for a parameter $\gamma > 0$ to be set. (a) Plots of the log error $\|x^t - x^{LS}\|_2$ versus the iteration number t.(b) Plots of the log error $\|x^t - x^*\|_2$ versus the iteration number t. As expected, all three curves flatten out at the level of the least-squares error $\|x^{LS} - x^*\|_2 = 0.10 \approx \sqrt{\frac{s \log(ed/s)}{n}}$.

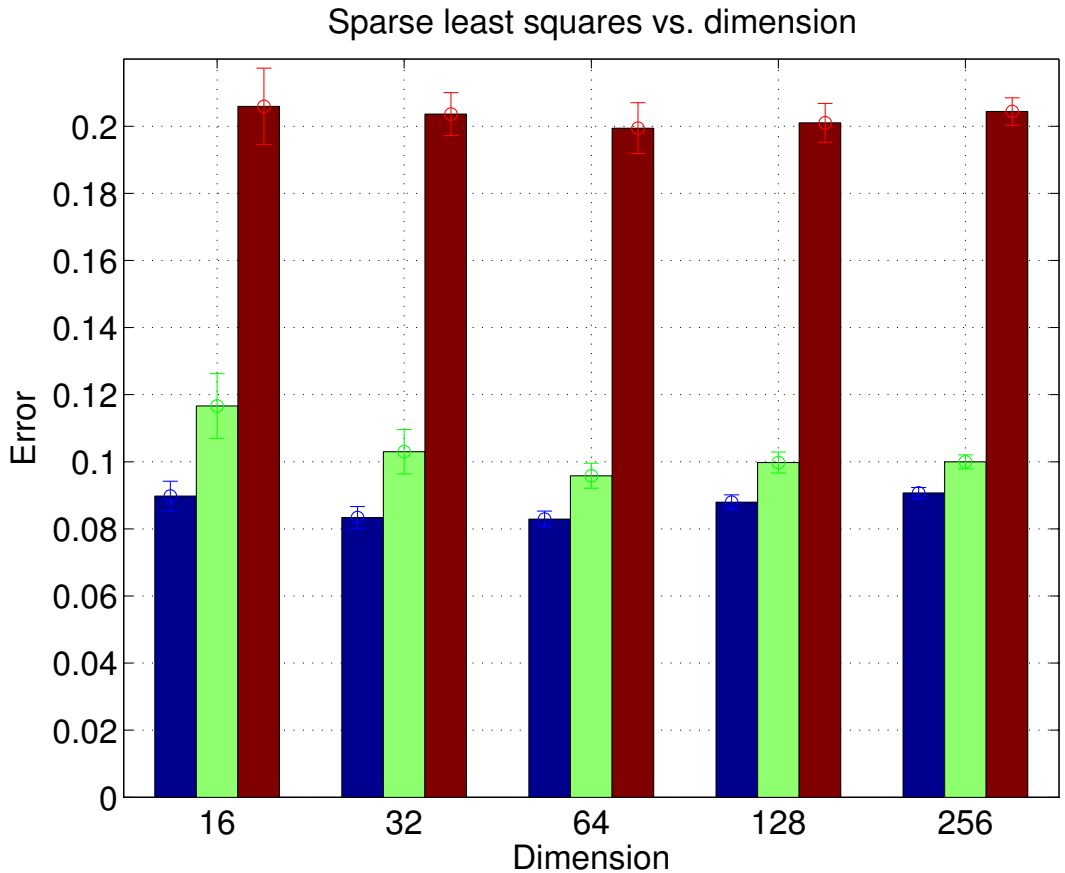

Figure 5: $d \in \{16, 32, 64, 128, 256\}$ and sparsity $s = \lceil 2\sqrt{d} \rceil$, on all occasions with a fixed sample size $n = 100s \log\left(\frac{ed}{s}\right)$. The initial Lasso solution has error $\|x^{LS} - x^*\|_2 \approx 0.10$, as shown by the blue bars. $N = 4$ iterations with a sketch size $m = 4s \log\left(\frac{ed}{s}\right)$.Red bars show the error in the naive sketch estimate, using a sketch of size $M = Nm = 16s \log\left(\frac{ed}{s}\right)$, equal to the total number of random projections used by the IHS algorithm.

Some Comments

- The classical sketching methods are sub-optimal, for the purposes of solution approximation.
- For IHS, the sketch dimension per iteration need grow only proportionally to the statistical dimension of the optimal solution, as measured by the Gaussian width of the tangent cone at the optimum.

Some Comments

- Sketched Newton's Method

$$\tilde{x}^{t+1} = \arg\min_{x \in \mathcal{C}} \frac{1}{2} \langle x - \tilde{x}^t, \nabla^2 f(\tilde{x}^t)\,(x - \tilde{x}^t) \rangle + \langle \nabla f(\tilde{x}^t),\, x - \tilde{x}^t \rangle.$$

$$\Updownarrow$$

$$\tilde{x}^{t+1} = \arg\min_{x \in \mathcal{C}} \frac{1}{2} \|\nabla^2 f(\tilde{x}^t)^{1/2}(x - \tilde{x}^t)\|_2^2 + \langle \nabla f(\tilde{x}^t),\, x - \tilde{x}^t \rangle,$$

$$\Downarrow$$

$$x^{t+1} = \arg\min_{x \in \mathcal{C}} \frac{1}{2} \|S^t \nabla^2 f(\tilde{x}^t)^{1/2}(x - x^t)\|_2^2 + \langle \nabla f(x^t),\, x - x^t \rangle,$$