Randomized algorithms for optimization: Statistical and computational guarantees

Martin Wainwright

UC Berkeley
Statistics and EECS

Based on joint work with:

Mert Pilanci (UC Berkeley)
Yun Yang (UC Berkeley)

Sketching via random projections

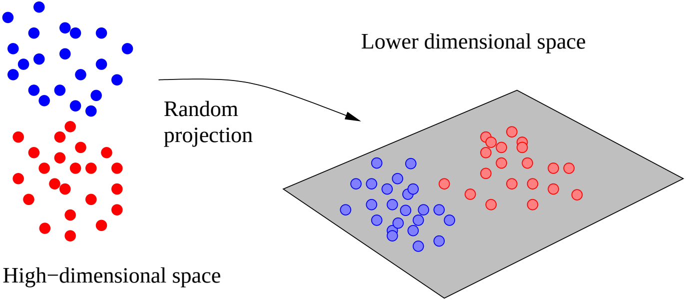
Massive data sets require fast algorithms but with rigorous guarantees.

Sketching via random projections

Massive data sets require fast algorithms but with rigorous guarantees.

A general purpose tool:

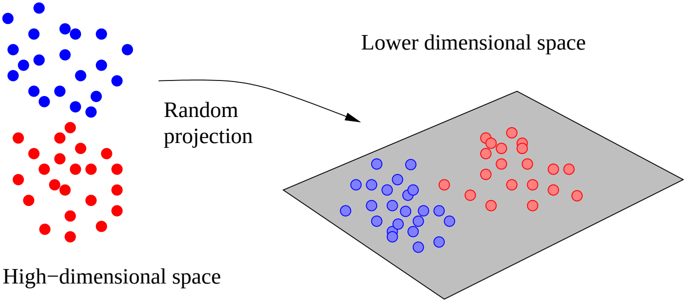
- Choose a random subspace of "low" dimension m.
- Project data into subspace, and solve reduced dimension problem.



Sketching via random projections

A general purpose tool:

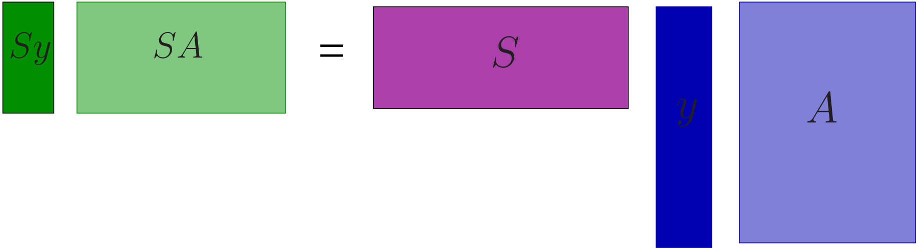
- Choose a random subspace of "low" dimension m.
- Project data into subspace, and solve reduced dimension problem.



Basic underlying idea now widely used in practice:

- Johnson & Lindenstrauss (1984): for Hilbert spaces
- various surveys and books: Vempala, 2004; Mahoney et al., 2011 Cormode et al., 2012

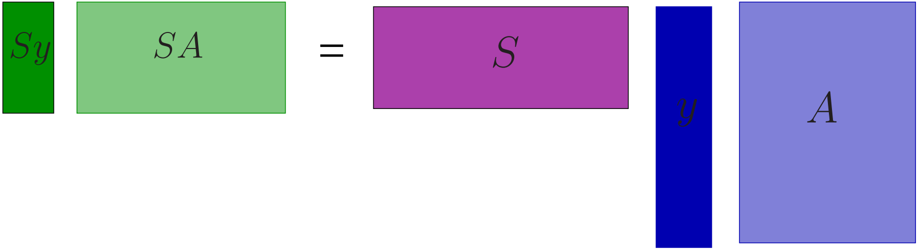
Classical sketching for constrained least-squares



Original problem: data $(y, A) \in \mathbb{R}^n \times \mathbb{R}^{n \times d}$, and convex constraint set $\mathcal{C} \subseteq \mathbb{R}^d$

$$x_{\text{LS}} = \arg\min_{x \in \mathcal{C}} \|Ax - y\|_2^2$$

Classical sketching for constrained least-squares



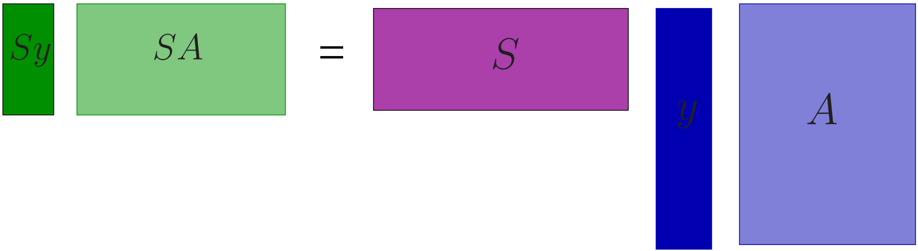
Original problem: data $(y, A) \in \mathbb{R}^n \times \mathbb{R}^{n \times d}$, and convex constraint set $\mathcal{C} \subseteq \mathbb{R}^d$

$$x_{\text{LS}} = \arg\min_{x \in \mathcal{C}} \|Ax - y\|_2^2$$

Sketched problem: data $(Sy, SA) \in \mathbb{R}^m \times \mathbb{R}^{m \times d}$:

$$\widehat{x} = \arg\min_{x \in \mathcal{C}} \|SAx - Sy\|_2^2$$

Classical sketching for constrained least-squares



Sketched problem: data $(Sy, SA) \in \mathbb{R}^m \times \mathbb{R}^{m \times d}$:

$$\widehat{x} = \arg\min_{x \in \mathcal{C}} \|SAx - Sy\|_2^2$$

Some history:

- random projections and Johnson-Lindenstrauss: 1980s onwards
- sketching for unconstrained least-squares: Sarlos, 2006
- leverage scores, cores sets: Drineas et al., 2010, 2011
- overview paper: Mahoney et al., 2011

Sketches based on randomized orthonormal systems

Step 1: Choose some fixed orthonormal matrix $H \in \mathbb{R}^{n \times n}$.
Example: Hadamard matrices

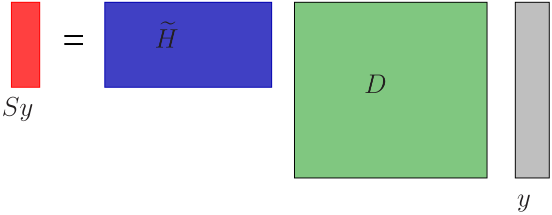
$$H_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \qquad H_{2^t} = \underbrace{H_2 \otimes H_2 \otimes \cdots \otimes H_2}_{\text{Kronecker product } t \text{ times}}$$

(E.g., Ailon & Liberty, 2010)

Sketches based on randomized orthonormal systems

Step 1: Choose some fixed orthonormal matrix $H \in \mathbb{R}^{n \times n}$.
Example: Hadamard matrices

$$H_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \qquad H_{2^t} = \underbrace{H_2 \otimes H_2 \otimes \cdots \otimes H_2}_{\text{Kronecker product } t \text{ times}}$$



Step 2:

(A) Multiply data vector y with a diagonal matrix of random signs $\{-1, +1\}$

(B) Choose m rows of H to form sub-sampled matrix $\widetilde{H} \in \mathbb{R}^{m \times n}$

(C) Requires $\mathcal{O}(n \log m)$ time to compute sketched vector $Sy = \widetilde{H}\, Dy$.

(E.g., Ailon & Liberty, 2010)

Different notions of approximation

Given a convex set $\mathcal{C} \subseteq \mathbb{R}^d$:

Original least-squares problem

$$x_{\text{LS}} = \arg\min_{x \in \mathcal{C}} \Big\{ \underbrace{\|Ax - y\|_2^2}_{f(x)} \Big\}$$

Sketched solution

$$\widehat{x} = \arg\min_{x \in \mathcal{C}} \Big\{ \|SAx - Sy\|_2^2 \Big\}$$

Question: When is sketched solution $\widehat{x}$ a "good" approximation to x_{LS}?

Different notions of approximation

Given a convex set $\mathcal{C} \subseteq \mathbb{R}^d$:

Original least-squares problem

$$x_{\text{LS}} = \arg\min_{x \in \mathcal{C}} \Big\{ \underbrace{\|Ax - y\|_2^2}_{f(x)} \Big\}$$

Sketched solution

$$\widehat{x} = \arg\min_{x \in \mathcal{C}} \Big\{ \|SAx - Sy\|_2^2 \Big\}$$

Question: When is sketched solution $\widehat{x}$ a "good" approximation to x_{LS}?

Cost approximation

Sketched solution $\widehat{x} \in \mathcal{C}$ is a δ-accurate cost approximation if

$$f(x_{\text{LS}}) \;\le\; f(\widehat{x}) \;\le\; (1+\delta)^2 \, f(x_{\text{LS}}).$$

Different notions of approximation

Given a convex set $\mathcal{C} \subseteq \mathbb{R}^d$:

Original least-squares problem

$$x_{\text{LS}} = \arg\min_{x \in \mathcal{C}} \Big\{ \underbrace{\|Ax - y\|_2^2}_{f(x)} \Big\}$$

Sketched solution

$$\widehat{x} = \arg\min_{x \in \mathcal{C}} \Big\{ \|SAx - Sy\|_2^2 \Big\}$$

Question: When is sketched solution $\widehat{x}$ a "good" approximation to x_{LS}?

Cost approximation

Sketched solution $\widehat{x} \in \mathcal{C}$ is a δ-accurate cost approximation if

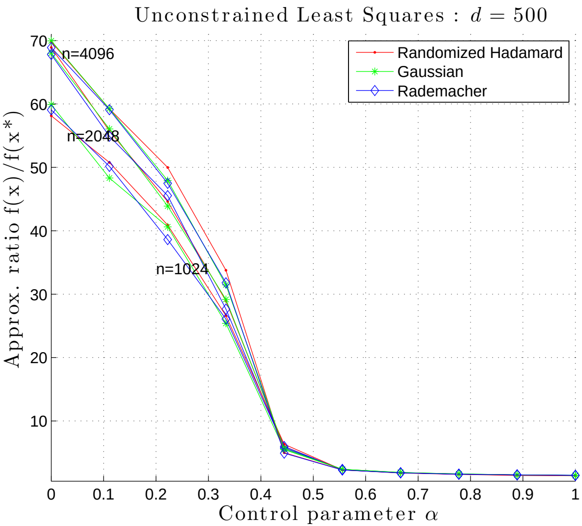
$$f(x_{\text{LS}}) \;\le\; f(\widehat{x}) \;\le\; (1+\delta)^2 \, f(x_{\text{LS}}).$$

Solution approximation

Sketched solution $\widehat{x} \in \mathcal{C}$ is a δ-accurate solution approximation if

$$\underbrace{\|\widehat{x} - x_{\text{LS}}\|_A}_{\frac{1}{\sqrt{n}}\|A(\widehat{x} - x_{\text{LS}})\|_2} \;\le \delta$$

Cost approx. for unconstrained LS



Sketch size $m = 4\alpha \, \text{rank}(A)$

What if solution approximation is our goal?

- often the least-squares solution x_{LS} itself is of primary interest
- unfortunately, δ-accurate cost approximation does not ensure high solution accuracy

What if solution approximation is our goal?

- often the least-squares solution x_{LS} itself is of primary interest
- unfortunately, δ-accurate cost approximation does not ensure high solution accuracy

Thought experiment: Consider random ensembles of linear regression problems:

$$y = Ax^* + w, \qquad \text{where } x^* \in \mathbb{R}^d, \text{ and } w \sim N(0, \sigma^2 I_n).$$

What if solution approximation is our goal?

- often the least-squares solution x_{LS} itself is of primary interest
- unfortunately, δ-accurate cost approximation does not ensure high solution accuracy

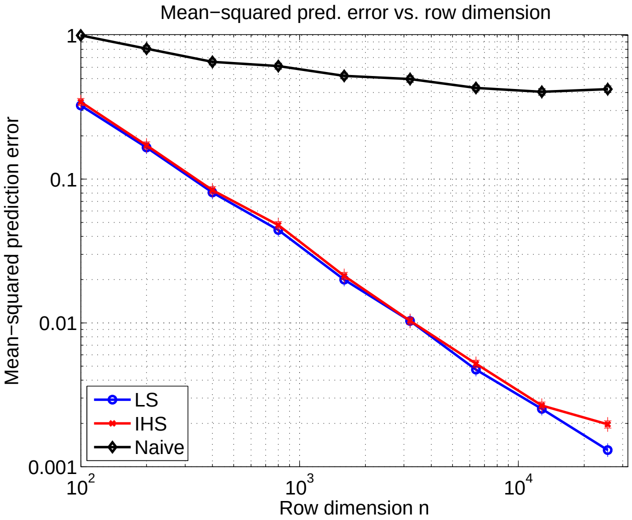
Thought experiment: Consider random ensembles of linear regression problems:

$$y = Ax^* + w, \qquad \text{where } x^* \in \mathbb{R}^d, \text{ and } w \sim N(0, \sigma^2 I_n).$$

Least-squares solution x_{LS} has mean-squared error

$$\mathbb{E}\|x_{\text{LS}} - x^*\|_A^2 = \underbrace{\frac{\sigma^2 \ \text{rank}(A)}{n}}_{\text{Nominal } \delta}$$

Unconstrained LS: Solution approximation



Sketch size $m = 4\,\mathrm{rank}(A)\log n$.

Fundamental cause of poor performance?

Recall planted ensembles of problems:

$$y = Ax^* + w, \qquad \text{where } x^* \in \mathcal{C}, \text{ and } w \sim N(0, \sigma^2 I_n).$$

Fundamental cause of poor performance?

Recall planted ensembles of problems:

$$y = Ax^* + w, \qquad \text{where } x^* \in \mathcal{C}, \text{ and } w \sim N(0, \sigma^2 I_n).$$

Any random sketching matrix $S \in \mathbb{R}^{m \times n}$ such that

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

Fundamental cause of poor performance?

Recall planted ensembles of problems:

$$y = Ax^* + w, \qquad \text{where } x^* \in \mathcal{C}, \text{ and } w \sim N(0, \sigma^2 I_n).$$

Any random sketching matrix $S \in \mathbb{R}^{m \times n}$ such that

$$|||\mathbb{E}_S\Big[S^T(SS^T)^{-1}S\Big]|||_{\text{op}} \precsim \frac{m}{n}$$

Theorem (Pilanci & W, 2014)

Any possible estimator $(Sy, SA) \mapsto \widetilde{x}$ *has error lower bounded as*

$$\sup_{x^* \in \mathcal{C}} \mathbb{E}_{S,w}\Big[\|\widetilde{x} - x_{LS}\|_A^2\Big] \succsim \sigma^2 \frac{\log P_{1/2}(\mathcal{C})}{\min\{n, m\}}$$

where $P_{1/2}(\mathcal{C})$ *is the* $1/2$*-packing number of* $\mathcal{C} \cap \mathbb{B}_2(1)$ *in the norm* $\|\cdot\|_A$.

Fundamental cause of poor performance?

Any random sketching matrix $S \in \mathbb{R}^{m \times n}$ such that

$$|||\mathbb{E}_S \Big[S^T (SS^T)^{-1} S \Big] |||_{\text{op}} \precsim \frac{m}{n}$$

Theorem (Pilanci & W, 2014)

Any possible estimator $(Sy, SA) \mapsto \widetilde{x}$ *has error lower bounded as*

$$\sup_{x^* \in \mathcal{C}} \mathbb{E}_{S,w} \Big[\|\widetilde{x} - x_{\text{LS}}\|_A^2 \Big] \succsim \sigma^2 \frac{\log P_{1/2}(\mathcal{C})}{\min\{n, m\}}$$

where $P_{1/2}(\mathcal{C})$ *is the* $1/2$*-packing number of* $\mathcal{C} \cap \mathbb{B}_2(1)$ *in the norm* $\|\cdot\|_A$.

Concretely: For unconstrained least-squares, we have

$$\sup_{x^* \in \mathcal{C}} \mathbb{E}_{S,w} \Big[\|\widetilde{x} - x_{\text{LS}}\|_A^2 \Big] \succsim \sigma^2 \frac{\text{rank}(A)}{\min\{n, m\}}.$$

Consequently, we need $m \geq n$ to match least-squares performance in estimating x^*.

A slightly different approach: Hessian sketch

Observe that

$$x_{\text{LS}} = \arg\min_{x \in \mathcal{C}} \|Ax - y\|_2^2 \;=\; \arg\min_{x \in \mathcal{C}} \Big\{ \frac{1}{2} x^T A^T A x - \langle A^T y,\, x \rangle \Big\}.$$

A slightly different approach: Hessian sketch

Observe that

$$x_{\text{LS}} = \arg\min_{x \in \mathcal{C}} \|Ax - y\|_2^2 \;=\; \arg\min_{x \in \mathcal{C}} \Big\{ \frac{1}{2} x^T A^T A x - \langle A^T y,\, x \rangle \Big\}.$$

Consider sketching only quadratic component:

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

A slightly different approach: Hessian sketch

Observe that

$$x_{\text{LS}} = \arg\min_{x \in \mathcal{C}} \|Ax - y\|_2^2 \;=\; \arg\min_{x \in \mathcal{C}} \Big\{ \frac{1}{2} x^T A^T A x - \langle A^T y,\, x \rangle \Big\}.$$

Consider sketching only quadratic component:

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

For a broad class of sketches, as long sketch dimension $m \gtrsim (1/\delta^2)\ \text{rank}(A)$, can prove that

$$\|\widetilde{x} - x_{\text{LS}}\|_A \precsim \delta\, \|x_{\text{LS}}\|_A$$

A slightly different approach: Hessian sketch

Observe that

$$x_{\text{LS}} = \arg\min_{x \in \mathcal{C}} \|Ax - y\|_2^2 \; = \; \arg\min_{x \in \mathcal{C}} \Big\{ \frac{1}{2} x^T A^T A x - \langle A^T y, \, x \rangle \Big\}.$$

Consider sketching only quadratic component:

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

For a broad class of sketches, as long sketch dimension $m \gtrsim (1/\delta^2) \; \text{rank}(A)$, can prove that

$$\|\widetilde{x} - x_{\text{LS}}\|_A \precsim \delta \, \|x_{\text{LS}}\|_A$$

Key point:

This one-step method is also provably sub-optimal, but the construction can be iterated to obtain an optimal method.

An optimal method: Iterative Hessian sketch

Given an iteration number $T \geq 1$:

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

An optimal method: Iterative Hessian sketch

Given an iteration number $T \geq 1$:

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

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

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

An optimal method: Iterative Hessian sketch

Given an iteration number $T \geq 1$:

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

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

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

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

An optimal method: Iterative Hessian sketch

Given an iteration number $T \geq 1$:

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

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

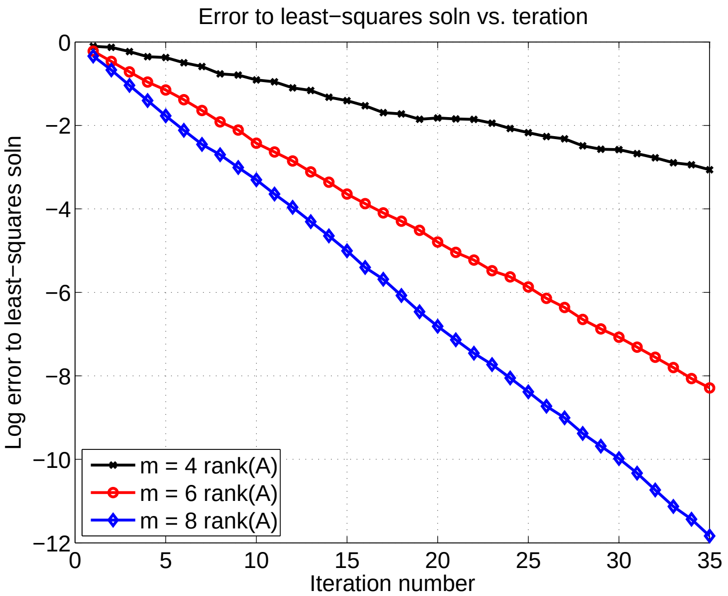
$$x^{t+1} = \arg\min_{x \in \mathcal{C}} \Big\{ \frac{1}{2} \| S^{t+1} A(x - x^t) \|_2^2 - \langle A^T (y - Ax^t), \, x \rangle \Big\}.$$

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

Intuition

- Step 1 returns the plain Hessian sketch $\widetilde{x} = x^1$.
- Step t is sketching a problem for which $x^t - x_{\text{LS}}$ is the optimal solution.
- The error is thus successively “localized”.

Geometric convergence for unconstrained LS



Theory for unconstrained least-squares

Theorem (Pilanci & W., 2014)

Given a sketch dimension $m \gtrsim \mathrm{rank}(A)$, *the error* *decays geometrically*

$$\|x^{t+1} - x_{LS}\|_A \leq \left(\frac{1}{2}\right)^t \|x_{LS}\|_A \qquad \textit{for all } t = 0, 1, \ldots, T-1$$

with probability at least $1 - c_1 T e^{-c_2 m}$.

Theory for unconstrained least-squares

Theorem (Pilanci & W., 2014)

Given a sketch dimension $m \gtrsim \operatorname{rank}(A)$, *the error decays geometrically*

$$\|x^{t+1} - x_{LS}\|_A \leq \Big(\frac{1}{2}\Big)^t \|x_{LS}\|_A \qquad \text{for all } t = 0, 1, \ldots, T-1$$

with probability at least $1 - c_1 T e^{-c_2 m}$.

- applies to any sub-Gaussian sketch; same result for fast JL sketches with additional logarithmic factors
- total number of random projections scales as $T\, m$
- for any $\epsilon > 0$, taking $T = \log\big(\frac{2\|x_{\mathrm{LS}}\|_A}{\epsilon}\big)$ iterations yields ϵ-accurate solution.

Experiments for planted ensembles

Linear regression problems with $A \in \mathbb{R}^{n \times d}$ and $n > d$:

$$y = Ax^* + w, \qquad \text{where } x^* \in \mathcal{C}, \text{ and } w \sim N(0, \sigma^2 I_n).$$

Experiments for planted ensembles

Linear regression problems with $A \in \mathbb{R}^{n \times d}$ and $n > d$:

$$y = Ax^* + w, \qquad \text{where } x^* \in \mathcal{C}, \text{ and } w \sim N(0, \sigma^2 I_n).$$

Least-squares solution has error

$$\mathbb{E}\|x_{\text{LS}} - x^*\|_A \precsim \sqrt{\frac{\sigma^2 d}{n}}$$

Experiments for planted ensembles

Linear regression problems with $A \in \mathbb{R}^{n \times d}$ and $n > d$:

$$y = Ax^* + w, \qquad \text{where } x^* \in \mathcal{C}, \text{ and } w \sim N(0, \sigma^2 I_n).$$

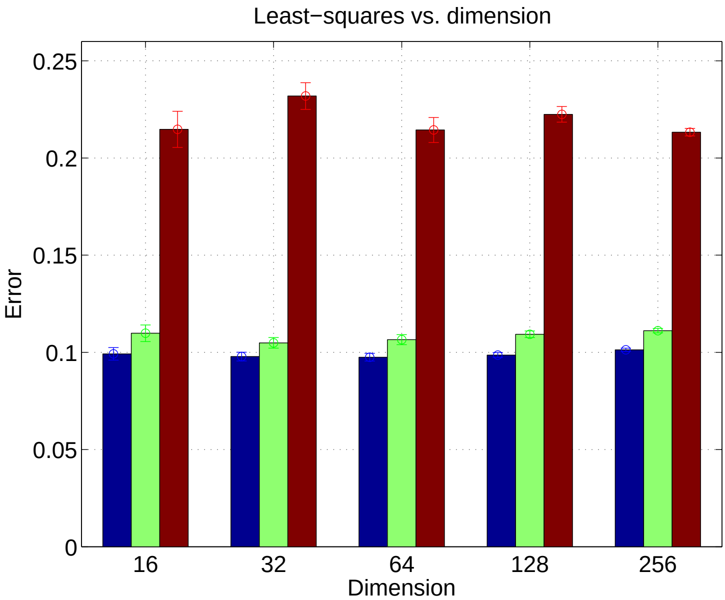
Least-squares solution has error

$$\mathbb{E}\|x_{\text{LS}} - x^*\|_A \precsim \sqrt{\frac{\sigma^2 d}{n}}$$

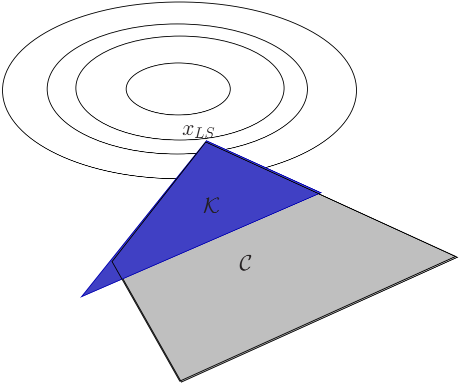
Scaling behavior:

- Fix $\sigma^2 = 1$ and sample size $n = 100d$, and vary $d \in \{16, 32, 64, 128, 256\}$.
- Run IHS with sketch size $m = 4d$ for $T = 4$ iterations.
- Compare to classical sketch with sketch size $16d$.

Sketched accuracy: IHS versus classical sketch



Extensions to constrained problems

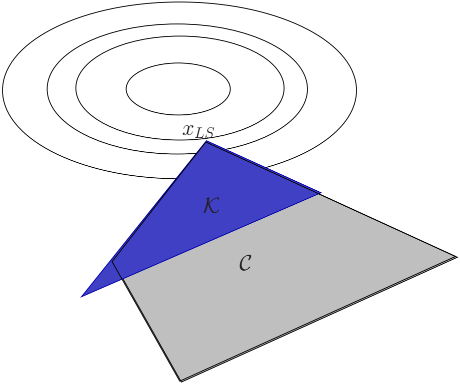


Constrained problem

$$x_{\mathrm{LS}} = \arg\min_{x\in\mathcal{C}} \|Ax - y\|_2^2$$

where $\mathcal{C} \subseteq \mathbb{R}^d$ is a convex set.

Extensions to constrained problems



Constrained problem

$$x_{\text{LS}} = \arg\min_{x \in \mathcal{C}} \|Ax - y\|_2^2$$

where $\mathcal{C} \subseteq \mathbb{R}^d$ is a convex set.

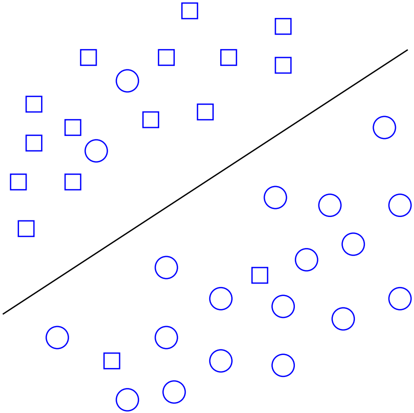
Tangent cone $\mathcal{K}$ at x_{LS}

Set of feasible directions at the optimum x_{LS}

$$\mathcal{K} = \left\{ \Delta \in \mathbb{R}^d \mid \Delta = t\,(x - x_{\text{LS}}) \quad \text{for some } x \in \mathcal{C}. \right\}.$$

Illustration: Binary classification with SVM

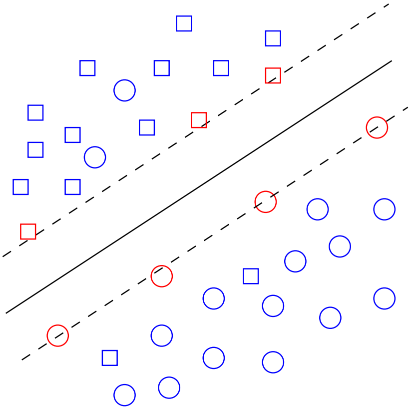
Observe labeled samples $(b_i, L_i) \in \mathbb{R}^D \times \{-1, +1\}$.



Goal: Find linear classifier $b \mapsto \mathrm{sign}(\langle w,\, b\rangle)$ with low classification error.

Illustration: Binary classification with SVM

Observe labeled samples $(b_i, L_i) \in \mathbb{R}^D \times \{-1, +1\}$.



- Support vector machine: produces classifier that depends only on samples lying on the margin
- Number of support vectors k typically $\ll$ total number of samples n

Sketching the dual of the SVM

Primal form of SVM:

$$\widehat{w} = \arg\min_{w \in \mathbb{R}^n} \left\{ \frac{1}{2\gamma} \sum_{i=1}^{d} \max\left\{0, 1 - L_i \langle w,\, b_i \rangle\right\} + \frac{1}{2} \|w\|_2^2 \right\}.$$

Sketching the dual of the SVM

Primal form of SVM:

$$\widehat{w} = \arg\min_{w \in \mathbb{R}^n} \left\{ \frac{1}{2\gamma} \sum_{i=1}^{d} \max\left\{0, 1 - L_i \langle w,\, b_i \rangle \right\} + \frac{1}{2} \|w\|_2^2 \right\}.$$

Dual form of SVM

$$x_{\text{LS}} := \arg\min_{x \in \mathcal{P}^n} \|\operatorname{diag}(L) B x\|_2^2,$$

$$\text{where } \mathcal{P}^n := \left\{ x \in \mathbb{R}^n \mid x \geq 0 \text{ and } \sum_{i=1}^{n} x_i = \gamma \right\}.$$

Sketching the dual of the SVM

Primal form of SVM:

$$\widehat{w} = \arg\min_{w \in \mathbb{R}^n} \Big\{ \frac{1}{2\gamma} \sum_{i=1}^{d} \max\big\{0, 1 - L_i \langle w,\, b_i \rangle\big\} + \frac{1}{2}\|w\|_2^2 \Big\}.$$

Dual form of SVM

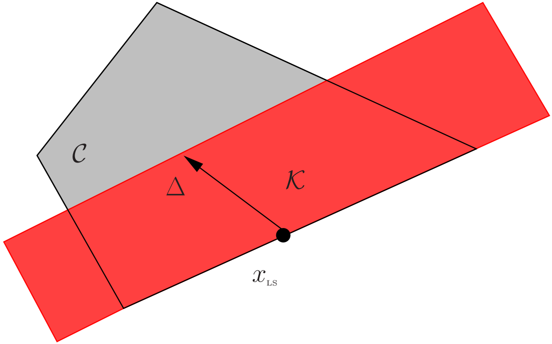
$$x_{\text{LS}} := \arg\min_{x \in \mathcal{P}^n} \|\operatorname{diag}(L) B x\|_2^2,$$

$$\text{where } \mathcal{P}^n := \Big\{ x \in \mathbb{R}^n \mid x \geq 0 \text{ and } \sum_{i=1}^{n} x_i = \gamma \Big\}.$$

Sketched dual SVM

$$\widehat{x} := \arg\min_{x \in \mathcal{P}^n} \|S \operatorname{diag}(L) B x\|_2^2$$

Unfavorable dependence on optimum x^*

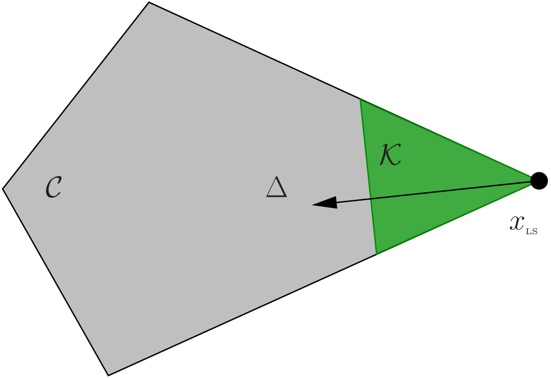


Tangent cone $\mathcal{K}$ at x_{LS}

Set of feasible directions at the optimum x_{LS}

$$\mathcal{K} = \left\{ \Delta \in \mathbb{R}^d \mid \Delta = t\,(x - x_{\text{LS}}) \quad \text{for some } x \in \mathcal{C}. \right\}.$$

Favorable dependence on optimum x^*

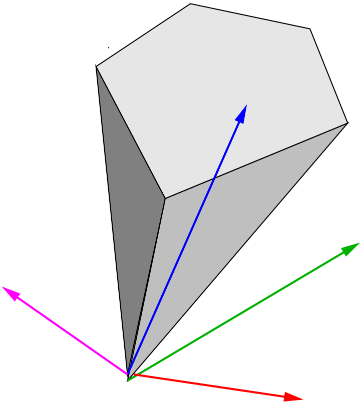


Tangent cone $\mathcal{K}$ at x_{LS}

Set of feasible directions at the optimum x_{LS}

$$\mathcal{K} = \big\{ \Delta \in \mathbb{R}^d \mid \Delta = t\,(x - x_{\text{LS}}) \quad \text{for some } x \in \mathcal{C}. \big\}.$$

Gaussian width of transformed tangent cone

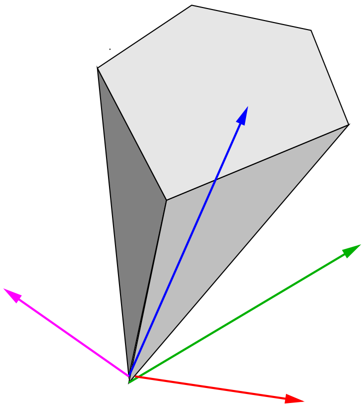


Gaussian width of set
$A\mathcal{K} \cap \mathcal{S}^{n-1} = \{A\Delta \mid \Delta \in \mathcal{K}, \|A\Delta\|_2 = 1\}$

$$\mathcal{W}(A\mathcal{K}) := \mathbb{E}\Big[\sup_{z \in A\mathcal{K} \cap \mathcal{S}^{n-1}} \langle g,\, z\rangle\Big]$$

where $g \sim N(0, I_{n\times n})$.

Gaussian width of transformed tangent cone



Gaussian width of set
$A\mathcal{K} \cap \mathcal{S}^{n-1} = \{A\Delta \mid \Delta \in \mathcal{K}, \|A\Delta\|_2 = 1\}$

$$\mathcal{W}(A\mathcal{K}) := \mathbb{E}\Big[\sup_{z \in A\mathcal{K} \cap \mathcal{S}^{n-1}} \langle g, z \rangle\Big]$$

where $g \sim N(0, I_{n \times n})$.

Gaussian widths used in many areas:

- Banach space theory: Pisier, 1986, Gordon 1988
- Empirical process theory: Ledoux & Talagrand, 1991, Bartlett et al., 2002
- Compressed sensing: Mendelson et al., 2008; Chandrasekaran et al., 2012

A general guarantee

Tangent cone at x_{LS}:

$$\mathcal{K} = \{\Delta \in \mathbb{R}^d \mid \Delta = t(x - x_{\mathrm{LS}}) \in \mathcal{C} \quad \text{for some } t \geq 0 \}.$$

Width of transformed cone $A\mathcal{K} \cap \mathcal{S}^{n-1}$:

$$\mathcal{W}(A\mathcal{K}) = \mathbb{E}\Big[\sup_{z \in A\mathcal{K} \cap \mathcal{S}^{n-1}} \langle g,\, z\rangle\Big] \qquad \text{where } g \sim N(0, I_{n\times n}).$$

Theorem (Pilanci & W., 2014)

Given a sketch dimension $m \gtrsim \mathcal{W}^2(A\mathcal{K})$, the error decays geometrically

$$\|x^{t+1} - x_{LS}\|_A \leq \Big(\frac{1}{2}\Big)^t \|x_{LS}\|_A \qquad \textit{for all } t = 0, 1, \ldots, T-1$$

with probability at least $1 - c_1 T e^{-c_2 m}$.

A general guarantee

Tangent cone at x_{LS}:

$$\mathcal{K} = \{\Delta \in \mathbb{R}^d \mid \Delta = t(x - x_{\text{LS}}) \in \mathcal{C} \quad \text{for some } t \geq 0\}.$$

Width of transformed cone $A\mathcal{K} \cap \mathcal{S}^{n-1}$:

$$\mathcal{W}(A\mathcal{K}) = \mathbb{E}\Big[\sup_{z \in A\mathcal{K} \cap \mathcal{S}^{n-1}} \langle g, \, z \rangle\Big] \qquad \text{where } g \sim N(0, I_{n \times n}).$$

Theorem (Pilanci & W., 2014)

Given a sketch dimension $m \gtrsim \mathcal{W}^2(A\mathcal{K})$, the error decays geometrically

$$\|x^{t+1} - x_{LS}\|_A \leq \Big(\frac{1}{2}\Big)^t \|x_{LS}\|_A \qquad \textit{for all } t = 0, 1, \ldots, T-1$$

with probability at least $1 - c_1 T e^{-c_2 m}$.

Similar results for fast JL sketches with additional logarithmic factors.

Ilustration: Width calculation for dual SVM

- Relevant constraint set is simplex in $\mathbb{R}^n$:

$$\mathcal{P}^n := \Big\{x \in \mathbb{R}^n \mid x \geq 0 \text{ and } \sum_{i=1}^{n} x_i = \gamma\Big\}.$$

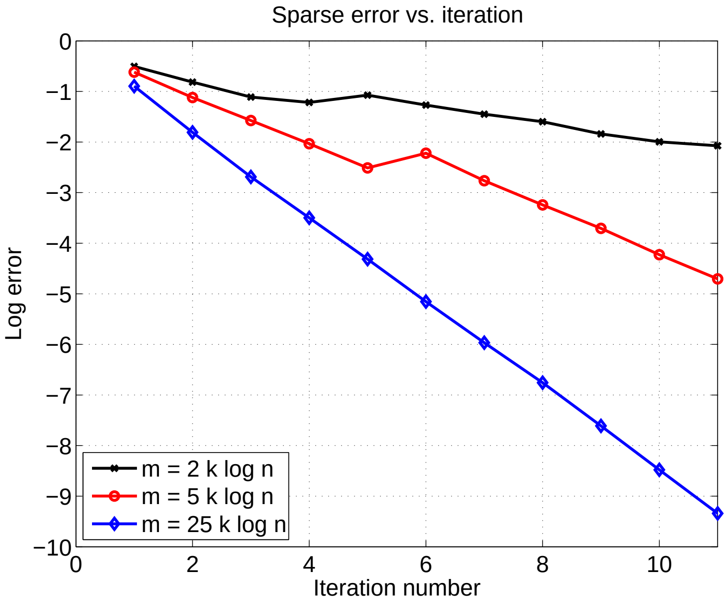
- in practice, SVM dual solution $\widehat{x}_{\text{dual}}$ is often sparse, with relatively few non-zeros
- under mild conditions on A, it can be shown that

$$\mathbb{E}\Big[\sup_{\substack{x \in \mathcal{P}^n \\ \|x\|_0 \leq k, \;\; \|Ax\|_2 \leq 1}} \langle g, \, Ax\rangle\Big] \;\precsim\; \sqrt{k \log n}.$$

Conclusion

For a SVM solution with k support vectors, a sketch dimension $m \succsim k \log n$ is sufficient to ensure geometric convergence.

Geometric convergence for SVM



A more general story: Newton Sketch

Convex program over set $\mathcal{C} \subseteq \mathbb{R}^d$:

$$x_{\text{opt}} = \arg\min_{x \in \mathcal{C}} f(x), \quad \text{where } f : \mathbb{R}^d \to \mathbb{R} \text{ is twice-differentiable.}$$

A more general story: Newton Sketch

Convex program over set $\mathcal{C} \subseteq \mathbb{R}^d$:

$$x_{\text{opt}} = \arg\min_{x \in \mathcal{C}} f(x), \quad \text{where } f : \mathbb{R}^d \to \mathbb{R} \text{ is twice-differentiable.}$$

Ordinary Newton steps:

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

where $\nabla^2 f(x^t)^{1/2}$ is a matrix square of the Hessian at x^t.

A more general story: Newton Sketch

Convex program over set $\mathcal{C} \subseteq \mathbb{R}^d$:

$$x_{\text{opt}} = \arg\min_{x \in \mathcal{C}} f(x), \quad \text{where } f : \mathbb{R}^d \to \mathbb{R} \text{ is twice-differentiable.}$$

Ordinary Newton steps:

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

where $\nabla^2 f(x^t)^{1/2}$ is a matrix square of the Hessian at x^t.

Sketched Newton steps:

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

A more general story: Newton Sketch

Convex program over set $\mathcal{C} \subseteq \mathbb{R}^d$:

$$x_{\text{opt}} = \arg\min_{x \in \mathcal{C}} f(x), \quad \text{where } f : \mathbb{R}^d \to \mathbb{R} \text{ is twice-differentiable.}$$

Ordinary Newton steps:

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

where $\nabla^2 f(x^t)^{1/2}$ is a matrix square of the Hessian at x^t.

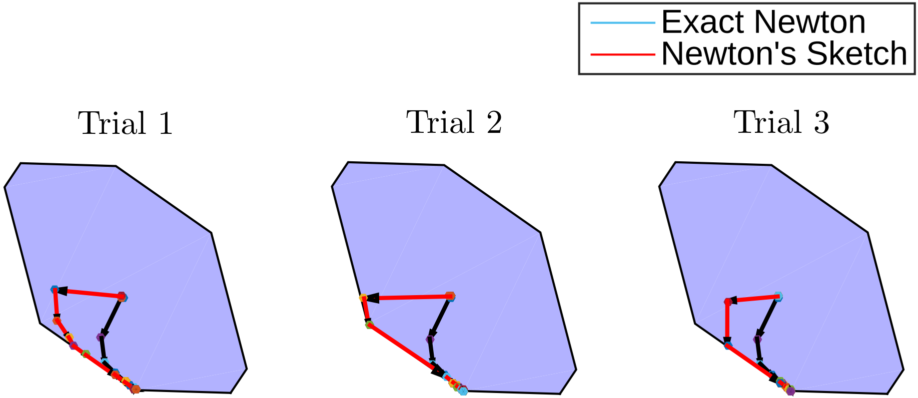
Sketched Newton steps:

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

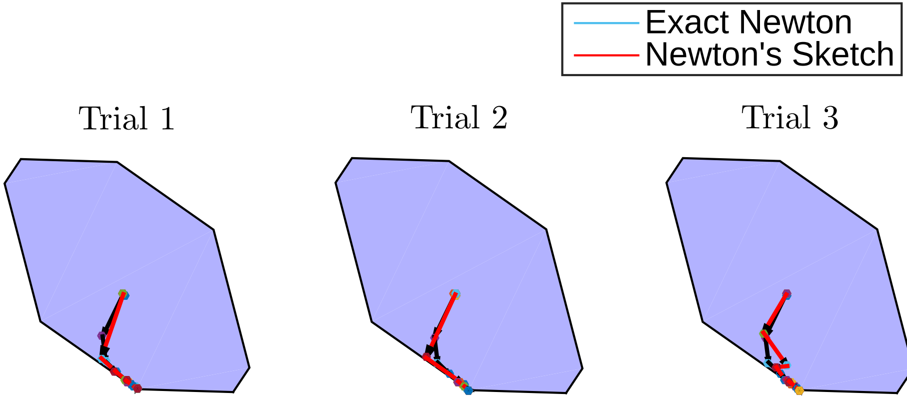
Question:

What is the minimal sketch dimension required to ensure that $\{\tilde{x}^t\}_{t=0}^T$ stays uniformly close to $\{x^t\}_{t=0}^T$?

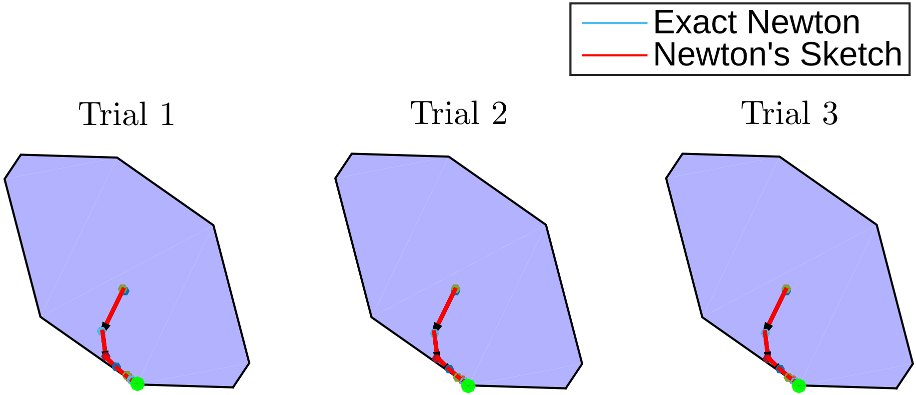
Sketching the central path: $m = d$



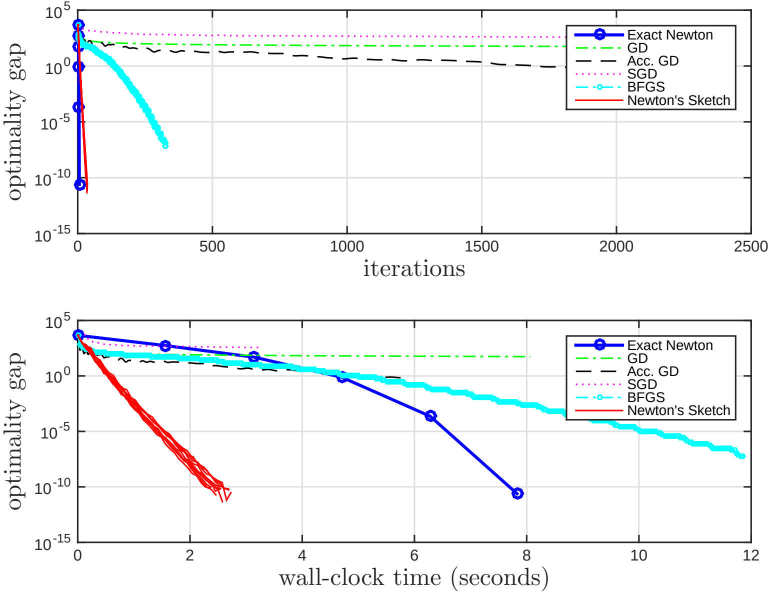
Sketching the central path: $m = 4d$



Sketching the central path: $m = 16d$



Running time comparisons



Summary

- important distinction: cost versus solution approximation
- classical least-squares sketch is provably sub-optimal for solution approximation
- iterative Hessian sketch: fast geometric convergence with guarantees in both cost/solution approximation
- sharp dependence of sketch dimension on geometry of solution and constraint set
- a more general perspective: sketched forms of Newton's method

Summary

- important distinction: cost versus solution approximation
- classical least-squares sketch is provably sub-optimal for solution approximation
- iterative Hessian sketch: fast geometric convergence with guarantees in both cost/solution approximation
- sharp dependence of sketch dimension on geometry of solution and constraint set
- a more general perspective: sketched forms of Newton's method

---

Papers/pre-prints:

- Pilanci & W. (2014a): Randomized sketches of convex programs with sharp guarantees, To appear in *IEEE Trans. Info. Theory*
- Pilanci & W. (2014b): Iterative Hessian Sketch: Fast and accurate solution approximation for constrained least-squares, Arxiv pre-print.
- Yang, Pilanci & W. (2015): Randomized sketches for kernels: fast and optimal non-parametric regression, Arxiv pre-print.
- Pilanci & W. (2015): Newton Sketch: A linear-time optimization algorithm with linear-quadratic convergence. Arxiv pre-print.