Gradient Projection Iterative Sketch for Large-Scale Constrained Least-Squares

Junqi Tang

University of Edinburgh

International Conference on Machine Learning, Sydney, 2017

Joint work with Mohammad Golbabaee and Mike Davies

Consider a constrained linear regression task in large data setting:

$$x^\star = \arg\min_{x \in \mathcal{K}} \left\{ f(x) := \frac{1}{2}\|y - Ax\|_2^2 \right\}. \tag{1}$$

- Training data matrix $A \in \mathbb{R}^{n \times d}$ with $n \gg d$, observation $y \in \mathbb{R}^n$
- $\mathcal{K}$: convex constrained set

First-order optimization:

- **Deterministic gradients** $\rightarrow$ large per-iteration cost scales with n
- **Stochastic gradients** $\rightarrow$ small per-iteration cost
- **Alternatives ?**

Sketching for reduced computation

- Sketching operators $S^t \in \mathbb{R}^{m \times n}$, $m \ll n$
- Classic sketching method provides an approximate solution:

$$\hat{x} = \arg\min_{x \in \mathcal{K}} \left\{ f_0(x) := \frac{1}{2m} \|Sy - SAx\|_2^2 \right\}, \tag{2}$$

- **Iterative Hessian Sketch** [Pilanci & Wainwright 2014]:

$$x^{t+1} = \arg\min_{x \in \mathcal{K}} \left\{ f_t(x) := \frac{1}{2m} \|S^t A(x - x^t)\|_2^2 - x^T A^T (y - Ax^t) \right\}, \tag{3}$$

provides progressively refined solution approximation.

Sketching for reduced computation

Proposition [Pilanci & Wainwright 2014]

Solution convergence. If the sketching operators S^t are sub-Gaussian matrices, the solution sequence of IHS sub-programs obeys:

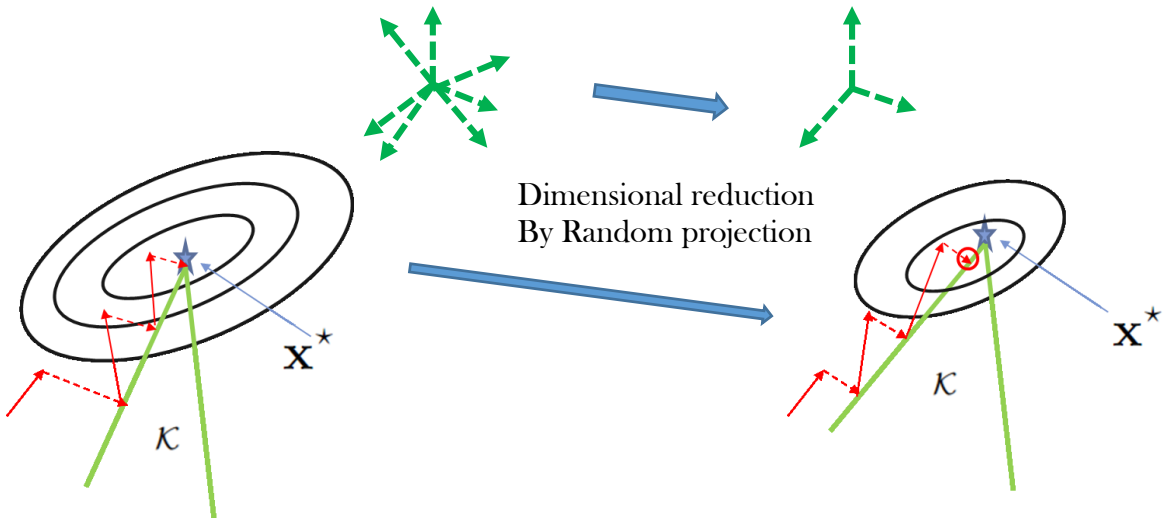
$$\|x^{t+1} - x^\star\|_A \leq \mathcal{O}(\frac{\mathcal{W}}{\sqrt{m}})\|x^t - x^\star\|_A, \tag{4}$$

with probability $\to 1$

- Linear convergence rate scales with the Gaussian Width $\mathcal{W} := \mathcal{W}(A\mathcal{C}_{x^\star} \cap \mathbb{S}^{n-1}) \leq \sqrt{d}$
- Requires the **exact** computation of sketched solutions

Sketched Gradients in a nutshell

- A new type of randomized 1st order optimization algorithm: sketching meta-algorithms + the deterministic gradient methods.
- Sketched cost functions are minimized **approximately**.



Gradient Projection Iterative Sketch — $\mathcal{G}(m, [\eta], [k])$

Warm start (optional):

Sketch $f_0(x) :\rightarrow \{SA, Sy\}$, the classic sketch objective

Gradient projection inner loop (i = 1, 2, 3, ... k_0)

$\rightarrow \quad x_{i+1}^0 = \mathcal{P}_{\mathcal{K}}[x_i^0 - \eta \nabla f_0(x_i^0)]$

Output $\quad x_0^1 = x_{k_0}^0$

Main iteration (t = 1, 2, 3, ... N):

Sketch $f_t(x) :\rightarrow \{S^t A,\ y,\ x_0^t\}$, the Iterative Hessian Sketch objective

Gradient projection inner loop (i = 1, 2, 3, ... k_t)

$\rightarrow x_{i+1}^t = \mathcal{P}_{\mathcal{K}}[x_i^t - \eta \nabla f_t(x_i^t)]$

Output $\quad x_0^{t+1} = x_{k_t}^t$

Accelerated GPIS — $\mathcal{A}(m, [\eta], [k])$

Warm start (optional):

Sketch $f_0(x) :\rightarrow \{SA, Sy\}$

Gradient projection inner loop (i = 1, 2, 3, ... k_0)

$\rightarrow$ $x_{i+1}^0 = \mathcal{P}_{\mathcal{K}}[z_i^0 - \eta \nabla f_0(z_i^0)]$

$\rightarrow$ **Extrapolate** z_{i+1}^0 with Nesterov's momentum

Output $x_0^1 = z_0^1 = x_{k_0}^0$

Main iteration (t = 1, 2, 3, ... N):

Sketch $f_t(x) :\rightarrow \{S^t A,\ y,\ x_0^t\}$

Gradient projection inner loop (i = 1, 2, 3, ... k_t)

$\rightarrow$ $x_{i+1}^t = \mathcal{P}_{\mathcal{K}}[z_i^t - \eta \nabla f_t(z_i^t)]$

$\rightarrow$ **Extrapolate** z_{i+1}^t with Nesterov's momentum

Output $x_0^{t+1} = z_0^{t+1} = x_{k_t}^t$

Theory

Linear convergence of GPIS with strong-convexity.
For $\mu I \preceq A^T A \preceq L I$, the output of GPIS algorithm obeys:

$$\|x_0^{N+1} - x^\star\|_A \leq \left\{ \prod_{t=1}^{N} \rho_t^\star \right\} \|x_0^1 - x^\star\|_A; \tag{5}$$

when S^t are Gaussian sketches and $d < m < n$, we show that:

$$\rho_t^\star \leq \mathcal{O}\left(\frac{\mathcal{W}}{\sqrt{m}} + (1 - \frac{\mu}{L}\frac{m-d}{m+d})^{k_t} \right), \quad \textit{with prob.} \ \to 1 \tag{6}$$

Without strong-convexity:

- $\mathcal{O}(\frac{1}{k})$ convergence for GPIS.
- $\mathcal{O}(\frac{1}{k^2})$ convergence for **Accelerated GPIS** .

Sketched Gradients in Practice

- **Count Sketch** [Clarkson & Woodruff, 2013]
  as sketching operator S^t
  $\rightarrow$ provides the best computational speed in practice
- **Line search** [Nesterov, 2007]
  for adaptive step sizes η
  $\rightarrow$ efficiently choose an aggressive step size at every iteration
- **Adaptive restart** [O'Donoghue & Candes, 2015]
  for Acc-GPIS
  $\rightarrow$ optimally utilize and control Nesterov's momentum

Experiments

Compared algorithms:

- SAGA with three different batch sizes $b = 10, 50, 100$.
- Acc-PGD (FISTA) : accelerated projected gradient with backtracking line-search and adaptive gradient restart

Measures:

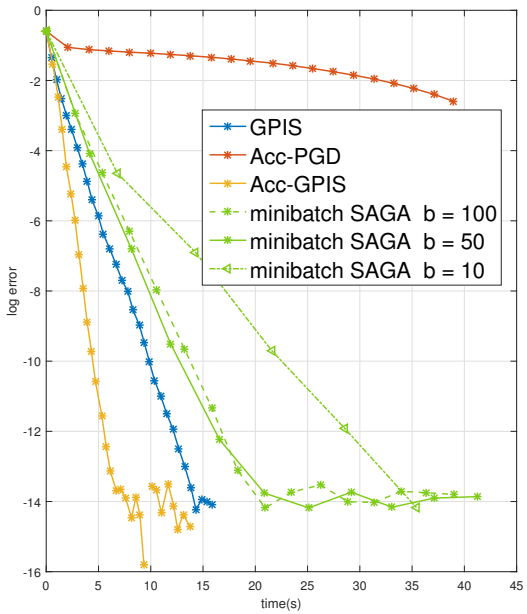
- **Epoch**: measure the cost of gradient calculation, the cost of calculating sketches
- **Actual running time**: count in the vectorized computational capability of modern processors, cost of loading minibatches, and cost of projections, etc..

Experiment: unconstrained Least-squares

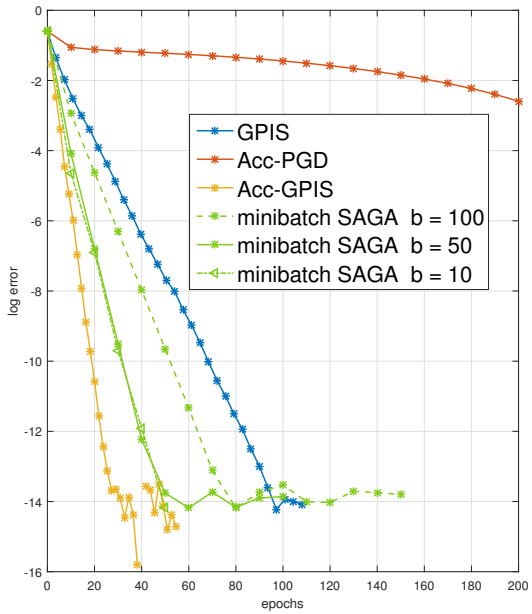
Millionsong data set $(n, d) = (500000, 90)$,

$$x^\star = \arg\min_{x \in \mathbb{R}^d} \left\{ f(x) := \frac{1}{2}\|y - Ax\|_2^2 \right\}. \tag{7}$$

Sketch size $m = 1000$ for GPIS/Acc-GPIS,



time(sec.)



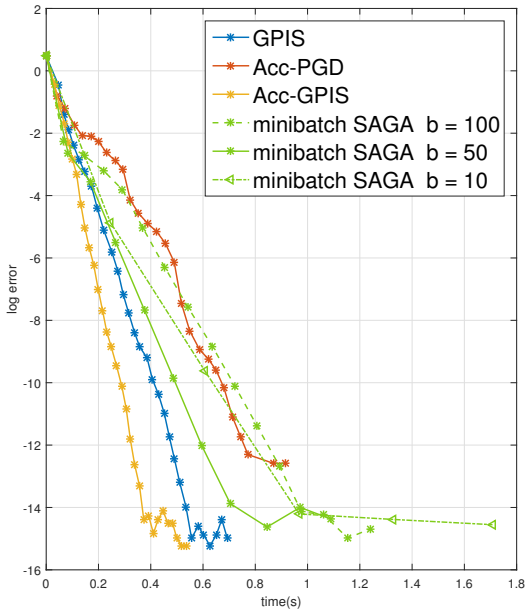
epoch counts

Experiment: Least-squares with l_1 constraint

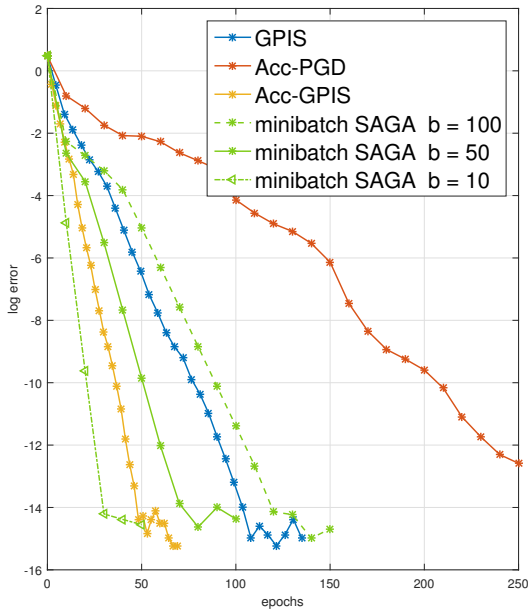
Augmented Magic04 data set $(n, d) = (19000, 10 + 40)$,
*40 additional irrelevant features.

$$x^\star = \arg \min_{x:\ \|x\|_1 \le c} \left\{ f(x) := \frac{1}{2}\|y - Ax\|_2^2 \right\}. \tag{8}$$

Sketch size $m = 475$ for GPIS/Acc-GPIS,



time(sec.)



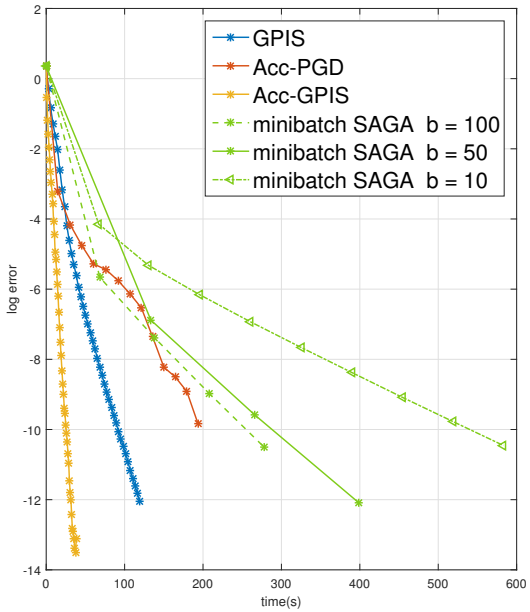
epoch counts

Experiment: Least-squares with Nuclear norm constraint

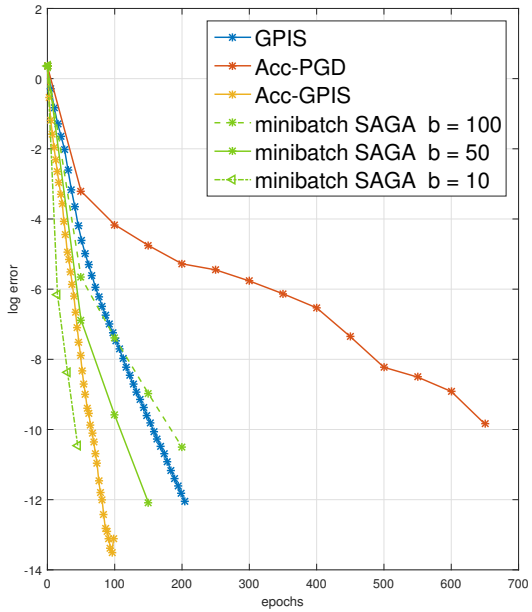
Synthetic multivariate matrix regression example $(n, d) = (50000, 100)$, matrix $X \in \mathbb{R}^{100 \times 100}$, with nuclear norm constraint to encourage low-rank solution.

$$X^\star = \arg \min_{X:\ \|X\|_\star \leq r} |||Y - AX|||_F^2. \tag{9}$$

Sketch size $m = 400$ for GPIS/Acc-GPIS,



time(sec.)



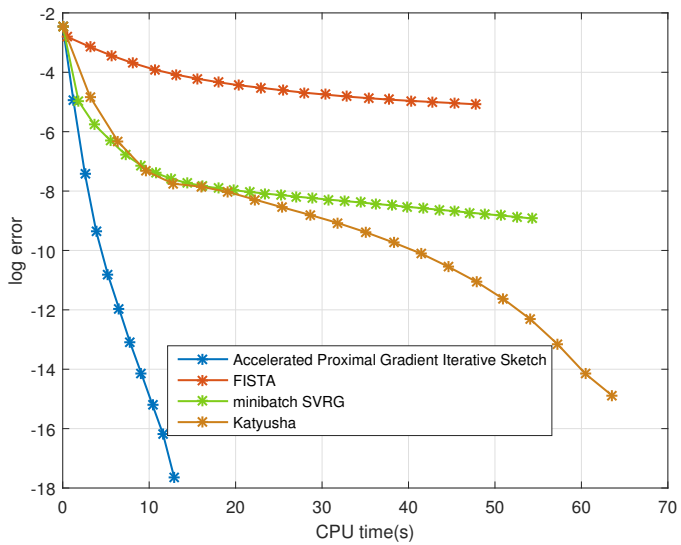
epoch counts

Extended results for Composite Least-squares

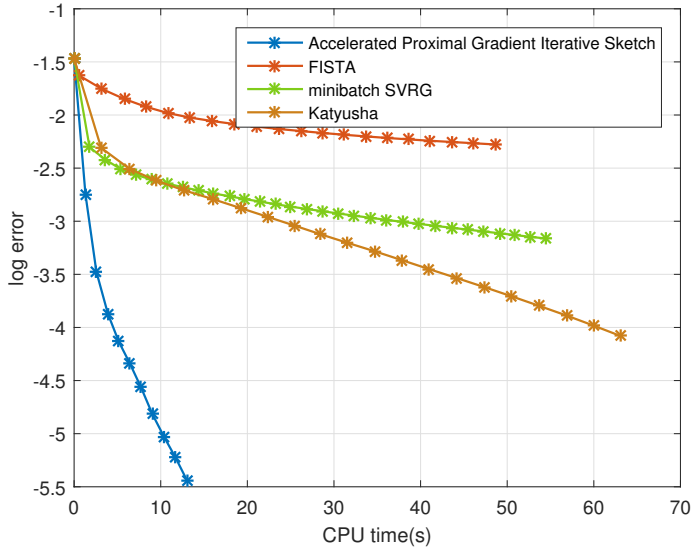
For composite minimization task reads:

$$x^\star = \arg\min_x \left\{ \frac{1}{2}\|y - Ax\|_2^2 + \lambda J(x) \right\}, J(x) = \|x\|_1 \qquad (10)$$

$(n, d) = (500000, 90)$
Sketch size $m = 800$ for Acc-GPIS (proximal setting),



(a) $\lambda = 4 \times 10^{-6}$ Millionsong



(b) $\lambda = 10^{-6}$ Millionsong

Figure: Experimental results on Composite Least-Squares with l_1 regularization

Summary

Take-home messages:

- Novel first order randomized algorithms *Sketched Gradients* are proposed.
- First convergence analysis is provided
- First randomized algorithm with practical implementations using a tailored *efficient line search* scheme and *adaptive restart* for Nesterov's momentum, and no need to foreknow L and μ
- **On going works:** Theory for fast sketching techniques and the proximal setting, projection-free variants.. etc..