High-Dimensional Optimization in Adaptive Random Subspaces

Jonathan Lacotte, Mert Pilanci and Marco Pavone

Stanford University

Convex, Smooth Optimization Problem

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex and μ-strongly smooth function, i.e., $\nabla^2 f(w) \preceq \mu I_n$ for all $w \in \mathbb{R}^n$, and $A \in \mathbb{R}^{n \times d}$ a high-dimensional matrix. We are interested in solving the primal problem

$$x^* = \operatorname*{argmin}_{x \in \mathbb{R}^d} f(Ax) + \frac{\lambda}{2}\|x\|_2^2 \,. \tag{1}$$

Approximate Recovery in Low-dimensional Space

Given a random matrix $S \in \mathbb{R}^{d \times m}$ with $m \ll d$, we consider instead the *sketched primal* problem

$$\alpha^* \in \operatorname*{argmin}_{\alpha \in \mathbb{R}^m} f(AS\alpha) + \frac{\lambda}{2}\alpha^\top S^\top S \alpha \,, \tag{2}$$

where we effectively restrict the optimization domain to a lower m-dimensional subspace. In this work, we explore the following questions: How can we estimate the original solution x^* given the sketched solution α^*? Is a uniformly random subspace the optimal choice, e.g., $S \sim$ Gaussian i.i.d.? Or, can we come up with an adaptive sampling distribution that is related to the matrix A, which yields stronger guarantees?

Let $f^*(z) := \sup_{w \in \mathbb{R}^n} \{w^\top z - f(w)\}$ be the Fenchel conjugate of f. Standard Fenchel duality holds,

$$\min_x f(Ax) + \frac{\lambda}{2}\|x\|_2^2 = \max_z -f^*(z) - \frac{1}{2\lambda}\|A^\top z\|_2^2 \,.$$

Strong duality also holds for the sketched program

$$\min_\alpha f(AS\alpha) + \frac{\lambda}{2}\|S\alpha\|_2^2 = \max_z -f^*(z) - \frac{1}{2\lambda}\|P_S A^\top z\|_2^2 \,,$$

where $P_S = S(S^\top S)^\dagger S^\top$ is the orthogonal projector onto the range of S. Intuitively, provided that S is well-chosen, the regularizers of the dual programs are close to each other

$$\|A^\top z\|_2^2 \approx \|P_S A^\top z\|_2^2 \,.$$

Key quantity to control the error between the two programs:

$$Z_f(A,S) = \sup_{\Delta \in (\operatorname{dom} f^* - z^*)} \left(\frac{\Delta^\top A P_S^\perp A^\top \Delta}{\|\Delta\|_2^2} \right)^{\frac{1}{2}} \,, \tag{3}$$

where z^* is the optimal dual solution of the original optimization problem, and $P_S^\perp = I - P_S$.

Deterministic Guarantee

We consider the candidate solution $\widetilde{x} = -\lambda^{-1} A^\top \nabla f(AS\alpha^*)$. Then, under the condition $\lambda \geqslant 2\mu Z_f^2$, we have

$$\|\widetilde{x} - x^*\|_2 \leqslant \sqrt{\frac{\mu}{2\lambda}} Z_f \|x^*\|_2 \,, \tag{4}$$

High-Probability Guarantee

Let $k \geqslant 2$, and $R_k(A) = \left(\sigma_k^2 + \frac{1}{k}\sum_{j=k+1}^{\rho} \sigma_j^2\right)^{\frac{1}{2}}$, where $\rho = \operatorname{rank}(A)$ and $\sigma_1 \geqslant \sigma_2 \geqslant \ldots \geqslant \sigma_\rho$ its singular values. We set $m = 2k$ and choose a sketching matrix

$$S = A^\top \widetilde{S} \,, \tag{5}$$

with $\widetilde{S} \in \mathbb{R}^{n \times m}$ Gaussian i.i.d. Then, for some universal constant $c_0 \leqslant 36$, provided $\lambda \geqslant 2\mu c_0^2 R_k^2(A)$, it holds with probability at least $1 - 12e^{-k}$ that

$$\|\widetilde{x} - x^*\|_2 \leqslant c_0 \sqrt{\frac{\mu}{2\lambda}} R_k(A) \|x^*\|_2 \,. \tag{6}$$

The above result is a consequence of (4) and classical results [2] on randomized low-rank approximations.

Adaptive versus Oblivious Sketching

Let $\nu_k = \sigma_k^2$ be the eigenvalues of $AA^\top$. We compare our theoretical predictions for different types of spectral decays: low rank ρ; κ-exponential decay $\nu_j \sim e^{-\kappa j}$ with $\kappa > 0$, and, β-polynomial decay $\nu_k \sim j^{-2\beta}$ with $\beta > 1/2$.

Given $\varepsilon > 0$ and $\eta \in (0,1)$, denote by m_A (resp. m_O, m_S) a sufficient dimension for which adaptive (resp. oblivious, leverage score) sketching yields

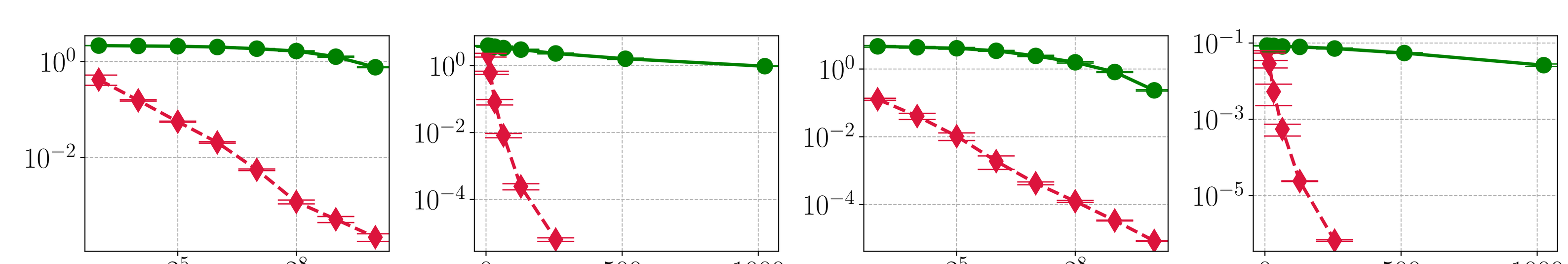
$$\|\widetilde{x} - x^*\|_2 / \|x^*\|_2 \leqslant \varepsilon$$

with probability at least $1 - \eta$. Bounds for oblivious sketching leverage results in [3] and bounds for Nystrom methods leverage results in [1].

	ρ-rank matrix ($\rho \ll n \wedge d$)	κ-exponential decay ($\kappa > 0$)	β-polynomial decay ($\beta > 1/2$)
Adaptive Gaussian (m_A)	$\rho + 1 + \log\left(\frac{12}{\eta}\right)$	$\kappa^{-1}\log\left(\frac{1}{\lambda\varepsilon}\right) + \log\left(\frac{12}{\eta}\right)$	$\lambda^{-12\beta}\varepsilon^{-1\beta} + \log\left(\frac{12}{\eta}\right)$
Oblivious Gaussian (m_O)	$(\rho+1)\varepsilon^{-2}\log\left(\frac{2\rho}{\eta}\right)$	$\kappa^{-1}\varepsilon^{-2}\log\left(\frac{1}{\lambda}\right)\log\left(\frac{2d}{\eta}\right)$	$\lambda^{-\frac{1}{2\beta}}\varepsilon^{-2}\log\left(\frac{2d}{\eta}\right)$
Leverage score (m_S)	$(\rho+1)\log\left(\frac{4\rho}{\eta}\right)$	$\kappa^{-1}\log\left(\frac{1}{\lambda\varepsilon}\right)\log\left(\frac{1}{\eta}\right)$	$\left(\lambda^{-\frac{1}{2\beta}}\varepsilon^{-\frac{1}{\beta}}\right)^{2\wedge\frac{\beta}{\beta-1}}\log\left(\frac{1}{\eta}\right)$
Lower bound on $\frac{m_O}{m_A}$	$\varepsilon^{-2}\log\rho$	$\varepsilon^{-2+h}\log 2d, \quad \forall h > 0$	$\varepsilon^{1\beta-2}\log(2d/\eta)$
Lower bound on $\frac{m_S}{m_A}$	$\log\rho$	$\min\left(\log\left(\frac{1}{\eta}\right), \kappa^{-1}\log\left(\frac{1}{\lambda\varepsilon}\right)\right)$	$\left(\lambda^{-\frac{1}{2\beta}}\varepsilon^{-\frac{1}{\beta}}\right)^{-1+2\wedge\frac{\beta}{\beta-1}}$

We compare numerically adaptive versus oblivious sketching. We use $n = 1000$ and $d = 2000$, A^{exp} and A^{poly}, satisfying respectively $\nu_j \sim ne^{-0.1j}$ (exponential) and $\nu_j \sim nj^{-2}$ (polynomial). We consider two loss functions:

- 'Logistic': $f(Ax) = n^{-1}\sum_{i=1}^n \ell_{y_i}(a_i^\top x)$ where $\ell_{y_i}(z) = y_i \log(1+e^{-z}) + (1-y_i)\log(1+e^z)$, $y \in \{0,1\}^n$.
- 'ReLU': $f(Ax) = (2n)^{-1}\sum_{i=1}^n (a_i^\top x)_+^2 - 2(a_i^\top x)y_i$.



(a) ReLU – polynomial (b) ReLU – exponential (c) Logistic – polynomial (d) Logistic – exponential

Fig. 1: In red, adaptive Gaussian sketching. In green, oblivious Gaussian sketching.

Iterative and Power Methods

Iterative method: Set $\widetilde{x}^{(0)} = 0$. At each iteration, compute $a^{(t)} = A\widetilde{x}^{(t-1)}$, and, $b^{(t)} = (S^\top S)^{-\frac{1}{2}} S^\top \widetilde{x}^{(t-1)}$, and solve

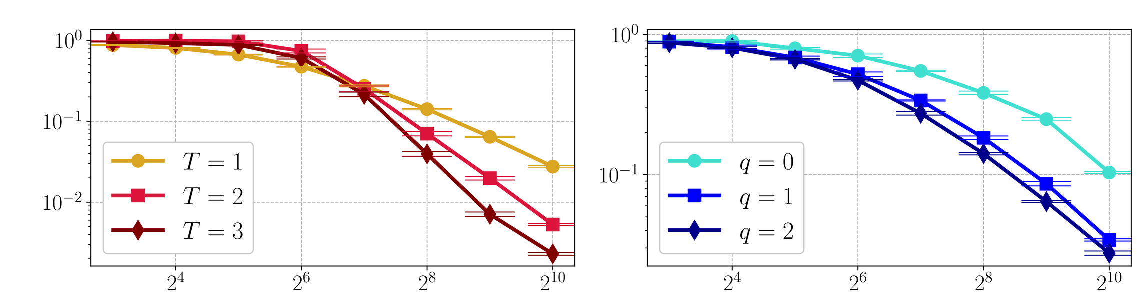
$$\alpha_\dagger^{(t)} = \operatorname*{argmin}_{\alpha_\dagger \in \mathbb{R}^m} f(A_{S,\dagger}\alpha_\dagger + a^{(t)}) + \frac{\lambda}{2}\|\alpha_\dagger + b^{(t)}\|_2^2 \,, \tag{7}$$

where $A_{S,\dagger} = AS(S^\top S)^{-\frac{1}{2}}$. Update the solution by $\widetilde{x}^{(t)} = -\frac{1}{\lambda}A^\top \nabla f(A_{S,\dagger}\alpha_\dagger^{(t)} + a^{(t)})$. Then, after T iterations, provided that $\lambda \geqslant 2\mu Z_f^2$, it holds that

$$\|\widetilde{x}^{(T)} - x^*\|_2 \leqslant \left(\frac{\mu Z_f^2}{2\lambda}\right)^{\frac{T}{2}} \|x^*\|_2 \,. \tag{8}$$

Power method [2]: Use the sketching matrix

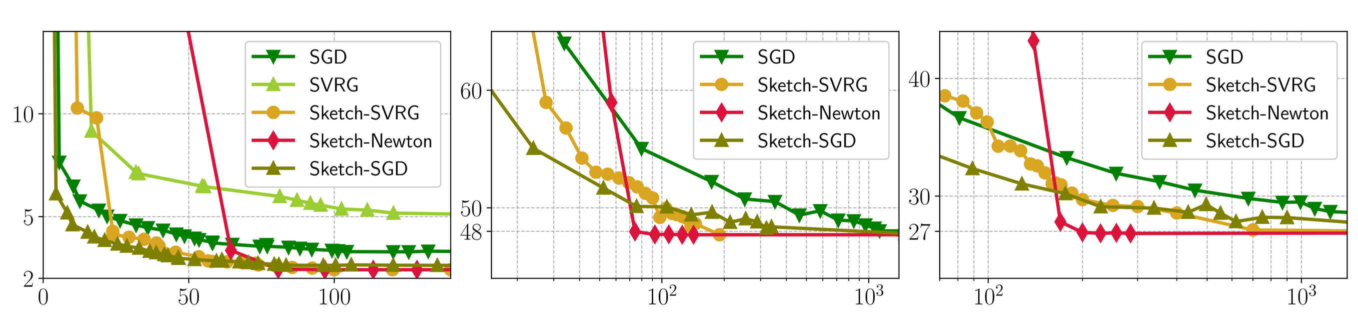
$$S = (A^\top A)^q A^\top \widetilde{S} \,, \quad \text{for some } q \geqslant 1 \,.$$



(a) Iterative method (b) Power method

Fig. 2: Benefits of iterative and power methods, evaluated on MNIST dataset.

Simulations on MNIST and CIFAR10 Datasets



(a) MNIST, $m = 512$, $\lambda = 10^{-5}$. (b) CIFAR, $m = 256$, $\lambda = 10^{-5}$. (c) CIFAR (random layer), $m = 1024$, $\lambda = 5.10^{-6}$.

Acknowledgements

This work was partially supported by the Office of Naval Research, ONR YIP Program, under contract N00014-17-1-2433, and, by the National Science Foundation under grant IIS-1838179.

References

[1] Alex Gittens and Michael W Mahoney. "Revisiting the Nyström method for improved large-scale machine learning". In: *The Journal of Machine Learning Research* 17.1 (2016), pp. 3977–4041.

[2] Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. "Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions". In: *SIAM review* 53.2 (2011), pp. 217–288.

[3] Lijun Zhang et al. "Recovering the optimal solution by dual random projection". In: *Conference on Learning Theory*. 2013, pp. 135–157.