Randomized Minimization of Eigenvalue Functions

Yurii Nesterov Anton Rodomanov

Catholic University of Louvain (UCLouvain), Belgium

UCLouvain

Motivating Example

Spectral linear regression problem:

$$\phi^* := \min_{x \in \mathbb{R}^d} [\phi(x) := \|Ax - C\|_\infty], \qquad Ax := \sum_{i=1}^{d} x_i A_i,$$

where $A_1, \dots, A_d, C \in \mathbb{R}^{n \times m}$ and $\|\cdot\|_\infty$ is the spectral norm (Schatten ℓ_∞).

- Can be reduced to SDP problem and solved by Interior-Point methods.
- However, this approach works only when all matrices are small.
- Can we provably solve this problem for large matrices using cheaper gradient methods?

Stochastic Optimization in Relative Scale

Consider the following problem:

$$f^* := \min_{x \in Q} f(x),$$

where $f : \mathbb{E} \to \mathbb{R}$ is a convex function and $Q \subseteq \mathbb{E}$ is a simple convex set.

- f is consistent with some Euclidean seminorm $\|x\|_B := \langle Bx, x \rangle^{1/2}$:
$$f(x) \geq \gamma_0 \|x - x_0\|_B^2, \quad \forall x \in \mathbb{E}.$$
- We have access to unbiased stochastic subgradient oracle $g(x, \xi)$:
$$\mathbb{E}_\xi[g(x, \xi)] \in \partial f(x), \quad \forall x \in \mathbb{E}.$$
- The size of $g(x, \xi)$ w.r.t. $f(x)$ is bounded:
$$\mathbb{E}_\xi[(\|g(x, \xi)\|_B^*)^2] \leq 2Lf(x), \quad \forall x \in \mathbb{E}.$$

Stochastic Gradient Method

Input: Point x_0, oracle g, norm B, step size h, number of iterations N.

Initialize $\bar{x}_0 := x_0$.

for $k = 0, 1, \dots, N-1$ **do**

Sample ξ_k, compute $g_k := g(x_k, \xi_k)$

$\bar{x}_{k+1} := (1 - \tau_k)\bar{x}_k + \tau_k x_k$ for $\tau_k := 1/(k+1)$

$x_{k+1} := \mathrm{GradStep}_{Q,B}(x_k, hg_k)$

return $\bar{x}_N$ $\qquad \triangleright = \frac{1}{N}\sum_{k=0}^{N-1} x_k$ by construction

This algorithm uses the following gradient step operation:

$$\mathrm{GradStep}_{Q,B}(x, g) := \operatorname*{argmin}_{y \in Q} \Big\{ \langle g, y \rangle + \frac{1}{2}\|y - x\|_B^2 \Big\},$$

where $x \in \mathbb{E}$ and $g \in (\ker B)^\perp$.

- When $B \succ 0$, this is a standard projected gradient step (w.r.t. B-norm):
$$\mathrm{GradStep}_{Q,B}(x, g) = \mathrm{Proj}_{Q,B}(x - B^{-1}g),$$
where $\mathrm{Proj}_{Q,B}(x) := \operatorname{argmin}_{y \in Q} \|y - x\|_B$.
- When $Q = \mathbb{E}$, point $T := \mathrm{GradStep}_{Q,B}(x, g)$ is a solution of linear system
$$B(T - x) = -g.$$

Convergence Guarantees

Point $\bar{x}_N$ is an approximate solution to our problem in relative scale:

$$(1 - \delta_N)\,\mathbb{E}[f(\bar{x}_N)] \leq f^*, \qquad \delta_N := \frac{1 + 2\gamma_0 L h^2 N}{1 + 2\gamma_0 h N}.$$

- A (nearly) optimal choice of step size is
$$h^* = \frac{1}{\sqrt{2\gamma_0 N L}} \implies \delta_N^* = \sqrt{\frac{2L}{\gamma_0 N}}.$$
- Alternatively, we can tune the step size to the target accuracy $\delta \in (0, 1)$:
$$h^* = \frac{\delta}{2L} \implies \delta_N^* \leq \delta \quad \forall N \geq N(\delta) := \frac{2L}{\gamma_0 \delta^2}.$$
- In many applications, one does not need high accuracy: $\delta \in [0.01, 0.05]$.

Application: Spectral Linear Regression

- Without loss of generality, assume that $n \leq m$.
- Let us square the objective function:
$$(\phi^*)^2 = \min_{x \in \mathbb{R}^d} [f(x) := \phi^2(x) = F(Ax - C)], \qquad F(X) := \|X\|_\infty^2.$$
- This problem needs to be solved with accuracy $\delta_2 := \delta(2 - \delta)$ to obtain a δ-approximate solution to the original problem.
- Choose B as the Gram matrix:
$$B := A^* A = (\langle A_i, A_j \rangle),$$
and x_0 by solving the linear regression problem:
$$x_0 := \operatorname*{argmin}_{x \in \mathbb{R}^d} \|Ax - C\|_F^2 = \mathrm{GradStep}_B(0, -A^* C).$$
- Then, f is consistent with the seminorm with constant
$$\gamma_0 = \frac{1}{n}.$$
- Oracle $g(x, \xi)$ can be naturally chosen as follows:
$$g(x, \xi) := A^* G(Ax - C, \xi),$$
provided that we know a suitable oracle $G(X, \xi)$ for F (unbiased and L-bounded w.r.t. F in the Frobenius norm).

Power Iteration Oracle

- Function F has the following subgradient:
$$F'(X) := 2v(X)[v(X)]^T X \in \partial F(X),$$
where $v(X) \in \mathcal{S}^{n-1}$ is a leading unit eigenvector of XX^T.
- To approximate $v(X)$, we can use standard Power Method of degree q:
$$v(X) \approx v_u^q(X) := \frac{(XX^T)^q u}{\|(XX^T)^q u\|}, \qquad u \sim \mathrm{Unif}(\mathcal{S}^{n-1}).$$
- This gives us oracle $G(x, u)$ with $L = 2$. However, this oracle is biased, so we have no theoretical guarantees.

Our New Oracle for Squared Spectral Norm

- Introduce convex probabilistic approximation of F of degree $p \geq 1$:
$$F_p(X) := \mathbb{E}_u[\langle (XX^T)^p u, u \rangle^{1/p}], \quad u \sim \mathrm{Unif}(\mathcal{S}^{n-1}).$$
- We can quantify how close F_p is to F depending on p:
$$\beta_p F(X) \leq F_p(X) \leq F(X), \qquad \beta_p := \frac{p}{p+2}\Big(\frac{1}{n}\Big)^{1/p}.$$
- For any odd $p = 2q + 1$, we have unbiased stochastic oracle for F_p:
$$G_p(X, u) := 2\hat{v}_u^q(X)[\hat{v}_u^q(X)]^T X, \qquad \hat{v}_u^q(X) := \frac{(XX^T)^q u}{\langle (XX^T)^{2q+1} u, u \rangle^{q/(2q+1)}}.$$
This oracle is bounded w.r.t. F with constant $L_p := 2/\beta_p$.
- Instead of f, we can now minimize $f_p(x) := F_p(Ax - C)$ using oracle G_p and choosing oracle degree $p = 2q + 1$ sufficiently large:
$$\beta_p \geq 1 - \delta_2/2 \impliedby q = \lfloor (\ln n + 2)/\delta_2 \rfloor.$$
The final worst-case iteration complexity bound is
$$N_p(\delta) := \frac{8\beta_p^2 L_p}{\gamma_0 \delta_2^2} = \frac{16\beta_p n}{\delta_2^2} \leq \frac{16n}{\delta^2}.$$

Numerical Experiments

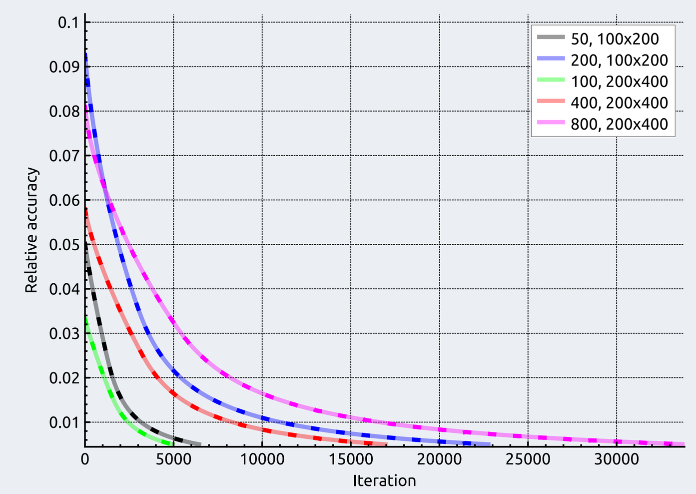
$\delta = 0.01$

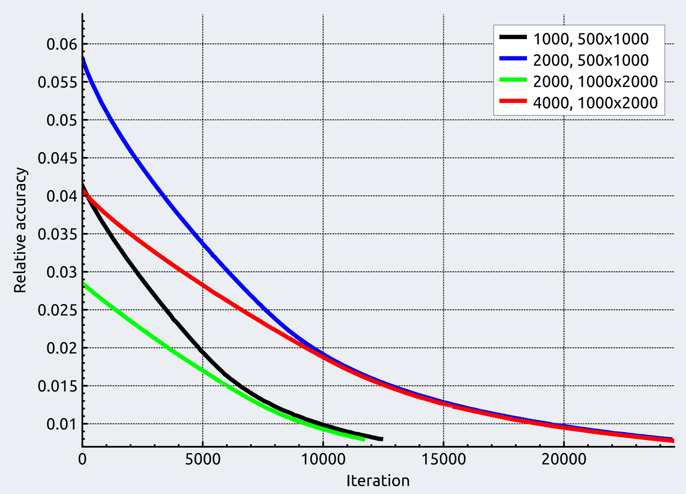
Dense data

d	n	m	p	$N_p(\delta)$
50	100	200	663	4000269
200	100	200		
100	200	400		
400	200	400	773	8000577
800	200	400		

Sparse data (5 nnz / column)

d	n	m	p	$N_p(\delta)$
1000	500	1000	825	20001419
2000	500	1000		
2000	1000	2000	895	40002992
4000	1000	2000		





- In all cases, performance is much better than predicted by theory.

Comparison of two oracles:

- Almost identical $\implies$ theoretical guarantees for Power Iteration Oracle?