Stochastic Gradient Methods for Minimization in Relative Scale

Anton Rodomanov

(Joint work with Y. Nesterov)

UCLouvain, Belgium

March 16, 2023

Université Grenoble Alpes, Grenoble

Outline

Outline

Motivating Example

Spectral Linear Regression (SLR) problem

$$\min_{x \in \mathbb{R}^d} \|A(x) - C\|,$$

where

$$A(x) := \sum_{i=1}^{d} x_i A_i,$$

and $A_1, \ldots, A_d, C \in \mathbb{R}^{n \times m}$ $(n \leq m)$, $\|\cdot\|$ is the matrix spectral norm.

Semidefinite Programming (SDP)

- SLR can be reduced to an SDP problem:

$$\min_{x \in \mathbb{R}^d, t \in \mathbb{R}} \quad t \qquad \text{s.t.} \quad \begin{pmatrix} tI & A(x) - C \\ (A(x) - C)^T & tI \end{pmatrix} \succeq 0.$$

- The SDP problem can be solved by Interior-Point methods.
- But this is expensive. Each iteration requires $O(n^3)$ time.
- Difficult to use sparsity of A_i, C.

Our Approach

Problem: $\phi^* := \min_{x \in \mathbb{R}^d} \big[\phi(x) := \|A(x) - C\|\big].$

- We propose randomized first-order methods that can solve this problem with relative accuracy $\delta \in (0, 1)$:

$$(1 - \delta)\, \mathbb{E}[\phi(\bar{x}_k)] \leq \phi^*.$$

- The main operation in our methods is the matrix-vector product:

$$A(x)v = \sum_{i=1}^{d} x_i (A_i v).$$

Can be evaluated in $O\big(\mathrm{nnz}(A)\big)$, where $\mathrm{nnz}(A) := \sum_{i=1}^{d} \mathrm{nnz}(A_i)$.

Outline

Problem Formulation

Problem

$$\min_{x \in Q} f(x),$$

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

Main assumptions:

- f has quadratic growth: there exists $x_0 \in Q$ and $\gamma_0 > 0$ such that

$$f(x) \geq \gamma_0 \|x - x_0\|_B^2, \qquad \forall x \in Q,$$

  where $\|h\|_B := \langle Bx, x \rangle^{1/2}$.
- We have a δ-relative stochastic subgradient oracle $g(x, \xi)$:

$$f(y) \geq (1 - \delta) f(x) + \langle \mathbb{E}_\xi[g(x, \xi)], y - x \rangle, \qquad \forall x, y \in Q.$$

- The size of $g(x, \xi)$ is uniformly relatively bounded:

$$\mathbb{E}_\xi[(\|g(x, \xi)\|_B^*)^2] \leq 2L f(x), \qquad \forall x \in Q.$$

Example: Squared Spectral Norm

Squared spectral norm ($n \leq m$)

$$F(X) := \|X\|^2 = \lambda_{\max}(XX^T), \qquad X \in \mathbb{R}^{n \times m}.$$

Quadratic growth: We have (w.r.t. Frobenius norm):

$$\gamma_0 = \frac{1}{n}, \qquad X_0 = 0.$$

Subgradient:

$$F'(X) = 2vv^T X, \qquad v := \mathrm{MaxEigVec}(XX^T),$$

where $v \in \mathbb{R}^n$ is a unit leading eigenvector of XX^T:

$$(XX^T)v = \lambda_{\max}(XX^T)v, \qquad \|v\| = 1.$$

Relative boundedness: This subgradient is bounded w.r.t. F:

$$\|F'(X)\|_F^2 \equiv 4F(X) \quad \Longrightarrow \quad L = 2.$$

Relative Boundedness

For any function $f : \mathbb{E} \to \mathbb{R}$, define

$$F(x) := \frac{1}{2} f^2(x).$$

Then, for any $x \in \mathbb{E}$, we have

$$\|\nabla f(x)\| \leq M \quad \iff \quad \|\nabla F(x)\|^2 \leq 2M^2 F(x).$$

Indeed, $\nabla F(x) = f(x) \nabla f(x)$. Hence,

$$\|\nabla F(x)\|^2 = f^2(x) \|\nabla f(x)\|^2 = 2\|\nabla f(x)\|^2 F(x).$$

Thus:

M-boundedness of $f \iff M^2$-relative boundedness of $\frac{1}{2} f^2$.

Composition with Affine Mapping

Consider

$$f(x) = F(Ax + b),$$

where $A\colon \mathbb{E} \to \mathbb{E}_1$, $b \in \mathbb{E}_1$, and F satisfies our assumptions:

- F has quadratic growth w.r.t. $\|\cdot\|_{B_1}$ with parameters γ_0 and y_0.
- We have δ-relative stochastic oracle $G(y, \xi)$ for F.
- Oracle $G(y, \xi)$ is uniformly relatively bounded with constant L.

Define the seminorm induced by $B = A^* B_1 A$:

$$\|x\|_B = \|Ax\|_{B_1}, \quad \forall x \in \mathbb{E}$$

and stochastic oracle

$$g(x, \xi) := A^* G(Ax + b, \xi).$$

Then, all properties are satisfied with the same constants γ_0, L, and

$$x_0 = \operatorname*{argmin}_{x \in Q} \|Ax + b - y_0\|_B.$$

Outline

Relative Stochastic Oracle for Spectral Norm

Computing $\text{MaxEigVec}(XX^T)$ exactly is very expensive.
Instead, we would like to approximate it by a random vector:

$$\text{MaxEigVec}(XX^T) \approx \text{MaxEigVec}_\delta(XX^T, \xi).$$

Need the following subroutine:

δ-relatively inexact stochastic eigenvector ($\delta \in (0,1)$)

Given a matrix $A \in \mathbb{S}^n_+$, compute $\hat{v} := \text{MaxEigVec}_\delta(A, \xi)$ such that

$$\mathbb{E}_\xi \langle A\hat{v}, \hat{v} \rangle \geq (1-\delta)\lambda_{\max}(A), \qquad \|\hat{v}\| = 1.$$

Then, we have a δ-relative inexact stochastic oracle:

$$G(X,\xi) := 2\hat{v}\hat{v}^T X, \qquad \hat{v} := \text{MaxEigVec}_\delta(XX^T, \xi).$$

It is still relatively bounded:

$$\|G(x,\xi)\|_F^2 = 4\langle XX^T\hat{v}, \hat{v}\rangle \leq 4\lambda_{\max}(XX^T) = 2F(X).$$

Power Method

Let $A \in \mathbb{S}^n_+$. For an integer degree $p \geq 1$, define

$$\hat{v}_p(A, \xi) := \frac{A^p \xi}{\|A^p \xi\|}, \qquad \xi \sim \mathrm{Unif}(\mathcal{S}^{n-1}).$$

Should be computed in a numerically stable way:

Power Method

$$\hat{v}_{k+1} := \frac{A \hat{v}_k}{\|A \hat{v}_k\|}, \quad k = 0, \ldots, p-1, \qquad \hat{v}_0 := \xi.$$

Complexity: p matrix-vector products.

Main result (Kuczyński and Woźniakowski, 1992)

$$\delta \leq \frac{\ln n}{p}.$$

Lanczos Method

$$\hat{v}_p \in \underset{x \in \mathcal{K}_p \cap \mathcal{S}^{n-1}}{\mathrm{Argmax}} \langle Ax, x \rangle, \qquad \mathcal{K}_p := \mathrm{span}(\xi, A\xi, A^2\xi, \ldots, A^p\xi).$$

Accuracy estimate (Kuczyński and Woźniakowski, 1992)

For $\xi \sim \mathrm{Unif}(\mathcal{S}^{n-1})$, we have

$$\delta \leq 3\Big(\frac{\ln n}{p}\Big)^2.$$

Implementing Lanczos Method

Lanczos tridiagonalization

Set $q_0 = 0$, $r_0 = \xi$. Iterate for $0 \leq k \leq p - 1$:

$$q_{k+1} = \frac{r_k}{\|r_k\|}, \qquad r_{k+1} = Aq_{k+1} - \langle Aq_{k+1}, q_{k+1} \rangle q_{k+1} - \|r_k\| q_k.$$

Result:

$$AQ_k = Q_k T_k + r_k e_k^T,$$

where $Q_k = [q_1, \dots, q_k]$ has orthonormal columns spanning $\mathcal{K}_k$, $e_k \in \mathbb{R}^n$ is the kth coordinate vector, where $T_k \in \mathbb{R}^{k \times k}$ is a tridiagonal matrix:

$$T_k = \mathrm{TriDiag}(\alpha_1, \dots, \alpha_k; \beta_1, \dots, \beta_k),$$

where $\alpha_k := \langle Aq_k, q_k \rangle$ and $\beta_k = \|r_k\|$.

Outline

Stochastic Gradient Method

Stochastic Gradient method

$$x_{k+1} = \mathrm{GradStep}_{Q,B}(x_k, a_k g_k), \qquad g_k := g(x_k, \xi_k), \qquad k \geq 0,$$

where $a_k \geq 0$ are certainly chosen step sizes.

Gradient step: For any $x \in \mathbb{E}$ and $g \in (\ker B)^{\perp}$, denote

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

(Also referred to as the "prox-mapping" by some authors.)

- When $B \succ 0$, this is the projected gradient step (w.r.t. B-norm):

  $$\mathrm{GradStep}_{Q,B}(x, g) = \mathrm{Proj}_{Q,B}(x - B^{-1} g),$$

  where $\mathrm{Proj}_{Q,B}(x) := \mathrm{argmin}_{y \in Q} \|y - x\|_B$.
- If $Q = \mathbb{E}$, point $T := \mathrm{GradStep}_{Q,B}(x, g)$ is a solution of linear system

  $$B(T - x) = -g.$$

Convergence Guarantee

Suppose a_i are deterministic step sizes and $a_i < \frac{1-\delta}{L}$.

Output point: For $c_i := a_i(1 - \delta - La_i)$, define and

$$\bar{x}_k := \frac{1}{C_k} \sum_{i=0}^{k-1} c_i x_i, \qquad C_k := \sum_{i=0}^{k-1} c_i.$$

Theorem. For any $k \geq 0$, we have

$$(1 - \Delta_k)\, \mathbb{E}[f(\bar{x}_k)] \leq f^*,$$

where

$$\Delta_k := \delta + \frac{1 - \delta + 2\gamma_0 L \sum_{i=0}^{k-1} a_i^2}{1 + 2\gamma_0 \sum_{i=0}^{k-1} a_i}.$$

Choice of Stepsizes I

$$\Delta_k := \delta + \frac{1 - \delta + 2\gamma_0 L \sum_{i=0}^{k-1} a_i^2}{1 + 2\gamma_0 \sum_{i=0}^{k-1} a_i} \qquad (\geq 0).$$

General recipe

To make $\Delta_k \to \delta$, it suffices to ensure that

$$\sum_{k=0}^{\infty} a_k = \infty, \qquad \sum_{k=0}^{\infty} a_k^2 < \infty.$$

Choice of Stepsizes II

$$\Delta_k := \delta + \frac{1 - \delta + 2\gamma_0 L \sum_{i=0}^{k-1} a_i^2}{1 + 2\gamma_0 \sum_{i=0}^{k-1} a_i} \qquad (\geq 0).$$

Optimal step sizes for a fixed horizon $N \geq 1$

$$a_k = a_N^* := \frac{1 - \delta}{\sqrt{2\gamma_0 N L (1 - \delta) + L^2} + L}, \qquad k \geq 0.$$

Under this choice, we have

$$\Delta_N \leq \delta + \sqrt{\frac{2L}{\gamma_0 N}}.$$

In particular,

$$N \geq N(\delta) := \frac{2L}{\gamma_0 \delta^2} \quad \Longrightarrow \quad \Delta_N \leq 2\delta.$$

Choice of Stepsizes III

Constant step size based on target accuracy $\delta \in (0, 1)$:

$$a_k = \frac{\delta}{2L} \quad \Longrightarrow \quad \Delta_N \leq 2\delta, \quad \forall N \geq N(\delta).$$

Outline

MaxCut Problem

Let $G = (V, E)$ be an undirected weighted graph with $V = \{1, \dots, n\}$ and weights $w(\{i, j\}) > 0$ for each edge $\{i, j\} \in E$.

Cut: For each vertex $i = 1, \dots, n$, assign $x_i = \pm 1$.

Value of cut

$$c(x) = \frac{1}{2} \sum_{\{i,j\} \in E} w(\{i, j\})(1 - x_i x_j).$$

MaxCut problem

$$c^* := \max_{x \in B^n} c(x),$$

where

$$B^n := \{x \in \mathbb{R}^n : x_i^2 = 1, \ i = 1, \dots, n\}.$$

Note: NP-complete! But can be efficiently approximated.

MaxCut via Laplacian Matrix

Note that

$$c(x) = \frac{1}{2} \sum_{\{i,j\} \in E} w(\{i,j\})(1 - x_i x_j) = \frac{1}{4} \langle Ax, x \rangle,$$

where $A \in \mathbb{S}^n_+$ is the Laplacian matrix of G:

$$A_{i,j} := \begin{cases} \sum_{k:\ \{i,k\} \in E} w(\{i,k\}), & \text{if } i = j, \\ -w(\{i,j\}), & \text{if } \{i,j\} \in E, \\ 0, & \text{otherwise.} \end{cases}$$

MaxCut problem

$$4c^* = \max_{x \in B^n} \langle Ax, x \rangle.$$

SDP Relaxation

MaxCut problem:

$$s^* := \max_{x \in B^n} \langle Ax, x \rangle.$$

SDP relaxation:

$$f^* := \underbrace{\min_{z \in \mathbb{R}^n} \Big\{ \sum_{i=1}^n z_i : A \preceq D(z) \Big\}}_{\text{Dual SDP relaxation}} = \underbrace{\max_{Y \in \mathbb{S}^n} \big\{ \langle A, Y \rangle : Y \succeq 0,\ d(Y) = e \big\}}_{\text{Primal SDP relaxation}},$$

where $e := (1, \ldots, 1)^T \in \mathbb{R}^n$.

Accuracy of relaxation (Goemans and Williamson, 1995)

$$0.878 \cdot f^* \le s^* \le f^*.$$

Finding the Cut

Random hyperplane algorithm (Goemans and Williamson, 1995)

1. Solve Primal SDP relaxation, obtain optimal Y^*.
2. Compute decomposition $Y^* = R^T R$, where $R \in \mathbb{R}^{m \times n}$.
3. Sample $u \sim \mathrm{Unif}(\mathcal{S}^{m-1})$.
4. Compute $x^* = \mathrm{sign}(R^T u)$.

Quality of the cut

$$\mathbb{E}_u[c(x^*)] \geq 0.878 \cdot c^*.$$

Transforming Dual Problem

We can assume that $D(A) \coloneqq \mathrm{Diag}(A_{1,1}, \ldots, A_{n,n}) \succ 0$. Then,

$$f^* = \min_{z \in \mathbb{R}^n} \Big\{ \sum_{i=1}^n z_i : A \preceq D(z) \Big\}$$
$$= \min_{z \in \mathbb{R}^n_{++}} \Big\{ \sum_{i=1}^n z_i : \lambda_{\max}\big([D(z)]^{-1/2} A [D(z)]^{-1/2}\big) \le 1 \Big\}.$$

Make change of variables $x_i = z_i^{-1/2}$. Then:

$$f^* = \min_{x \in \mathbb{R}^n_{++}} \Big\{ \underbrace{\sum_{i=1}^n \frac{1}{x_i^2}}_{=: \phi(x)} : \underbrace{\lambda_{\max}\big(D(x) A D(x)\big) \le 1}_{=: f(x)} \Big\}$$
$$= \min_{x \in \mathbb{R}^n_{++}} [\phi(x) f(x)] = \min_{x \in \mathbb{R}^n_{++}} \{ f(x) : \phi(x) \le 1 \}.$$

Solving Transformed Dual

Problem

$$f^* = \min_{x \in Q} f(x), \qquad f(x) := \lambda_{\max}\bigl(S(x)\bigr),$$

where

$$S(x) := D(x) A D(x), \qquad Q := \Bigl\{ x \in \mathbb{R}^n_{++} : \sum_{i=1}^n \frac{1}{x_i^2} \le 1 \Bigr\}.$$

Note: $f(x) = \|P(x)\|^2$, where $P(x) := D(x) A^{1/2}$.

Oracle: $g(x, \xi) := 2d(AD(x)\hat{v}\hat{v}^T)$, $\hat{v} := \mathrm{MaxEigVec}_\delta\bigl(S(x), \xi\bigr)$.

Choice of norm: $B = D(A)$.

Then, f and $g(x, \xi)$ satisfy our assumptions with

$$\gamma_0 = \frac{1}{n}, \qquad x_0 = \operatorname*{argmin}_{x \in Q} \|x\|_B = \mathrm{Proj}_{Q,B}(0), \qquad L = 2.$$

Final Guarantee I

We can get a point $\bar{x}_k \in Q$ such that

$$(1 - \delta)\,\mathbb{E}[f(\bar{x}_k)] \leq f^*,$$

where

$$f(x) := \lambda_{\max}\big(S(x)\big)$$

in the following number of iterations:

$$N(\delta) = O\Big(\frac{L}{\gamma_0 \delta^2}\Big) = O\Big(\frac{n}{\delta^2}\Big).$$

Note: We cannot compute $f(\bar{x}_k)$ exactly (too expensive).

Final Guarantee II

Nevertheless, we can efficiently compute

$$\hat{f}_k := (1-\delta)^{-1}\langle S(\bar{x}_k)\hat{v}, \hat{v}\rangle, \qquad \hat{v} := \mathrm{MaxEigVec}_\delta\big(S(\bar{x}_k), \xi\big)$$

such that

$$\mathbb{E}[f(\bar{x}_k)] \le \mathbb{E}[\hat{f}_k] \le (1-\delta)^{-2} f^*.$$

Then:

$$f^* \le \mathbb{E}[\hat{f}_k] \le (1-\delta)^{-2} f^*.$$

Combining this with

$$0.878 \cdot f^* \le s^* \le f^*,$$

we get for the MaxCut problem:

$$\alpha\, \mathbb{E}[\hat{f}_k] \le s^* \le \mathbb{E}[\hat{f}_k],$$

where $\alpha := 0.878(1-\delta)^2$.

Final Guarantee III

Result

We can produce $\hat{f}_k$ such that

$$\alpha\,\mathbb{E}[\hat{f}_k] \leq s^* \leq \mathbb{E}[\hat{f}_k].$$

where $\alpha := 0.878(1-\delta)^2$.

Total arithmetical complexity:

$$N(\delta) \times \underbrace{O\left(\frac{\ln n}{\sqrt{\delta}}\right)}_{\substack{\text{Number of} \\ \text{mat-vec products}}} \times \underbrace{O(|E|)}_{\substack{\text{Cost of} \\ \text{mat-vec product}}} = O\left(\frac{n|E|\ln n}{\delta^{5/2}}\right).$$

Note: We do not need a very small δ:

$$\delta = 0.05 \quad \Longrightarrow \quad \alpha \approx 0.79,$$
$$\delta = 0.01 \quad \Longrightarrow \quad \alpha \approx 0.86.$$

Open Question

Open question: How to generate the cut corresponding to $\hat{f}_k$?

Main problem: We need an approximate optimal solution Y_k for the primal SDP relaxation and its factorization

$$Y_k = R_k^T R_k.$$

Thank you!

References

M. X. Goemans and D. P. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM*, 42(6):1115–1145, Nov. 1995. ISSN: 0004-5411. DOI: 10.1145/227683.227684.

J. Kuczyński and H. Woźniakowski. Estimating the Largest Eigenvalue by the Power and Lanczos Algorithms with a Random Start. *SIAM Journal on Matrix Analysis and Applications*, 13(4):1094–1122, Oct. 1992. DOI: 10.1137/0613066.