On the almost-sure convergence of a stochastic sequential quadratic optimization method

Xin Jiang

Institute for Data, Intelligent Systems, and Computation
Industrial and Systems Engineering Department
Lehigh University

joint work with Frank E. Curtis and Qi Wang

2023 INFORMS Annual Meeting, Phoenix, Arizona

October 15, 2023

Convergence of random variables

- consider stochastic process $\{V_k\}$ and random variable V in $(\Omega, \mathcal{F}, \mathbb{P})$
- **convergence in probability:** $\{V_k\} \xrightarrow{p} V$ if and only if

$$\lim_{k\to\infty} \mathbb{P}[\|V_k - V\| > \epsilon] = 0 \quad \text{for all} \quad \epsilon > 0$$

- **almost-sure convergence:** $\{V_k\} \xrightarrow{\text{a.s.}} V$ if and only if

$$\mathbb{P}\Big[\lim_{k\to\infty} V_k = V\Big] = 1$$

- almost-sure convergence implies convergence in probability

$$\{V_k\} \xrightarrow{\text{a.s.}} V \qquad \Longrightarrow \qquad \{V_k\} \xrightarrow{p} V$$

Stochastic optimization (unconstrained)

$$\text{minimize} \quad f(x) \overset{\triangle}{=} \mathbb{E}_\omega[F(x,\omega)]$$

- $f\colon \mathbb{R}^n \to \mathbb{R}$ is smooth and potentially nonconvex
- random variable ω has probability space $(\Omega, \mathcal{F}, \mathbb{P})$

Stochastic approximation/gradient method

- using unbiased derivative estimates to solve a (nonlinear) equation

$$\lim_{k\to\infty} \mathbb{E}[(X_k - x_\star)^2] = 0 \qquad \Longrightarrow \qquad \{X_k\} \xrightarrow{\text{p}} x_\star$$

- cast into the context of stochastic (unconstrained) minimization:

$$\lim_{k\to\infty} \mathbb{E}[\|\nabla f(X_k)\|^2] = 0$$

Almost-sure convergence

- for stochastic approximation (solving an equation): $\{X_k\} \xrightarrow{\text{a.s.}} x_\star$
- for stochastic gradient (minimization): $\{\nabla f(X_k)\} \xrightarrow{\text{a.s.}} 0$

Robbins and Monro (1951), Robbins and Siegmund (1971), Bertsekas and Tsitsiklis (2000)

Constrained stochastic optimization

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & c(x) = 0 \end{array}$$

- $f(x) = \mathbb{E}[F(x, \omega)]$, and c is continuously differentiable
- ∇f and ∇c are Lipschitz continuous
- stationarity condition: $\nabla f(x) + \nabla c(x) y = 0$, and $c(x) = 0$

Stochastic sequential quadratic optimization (SQP):

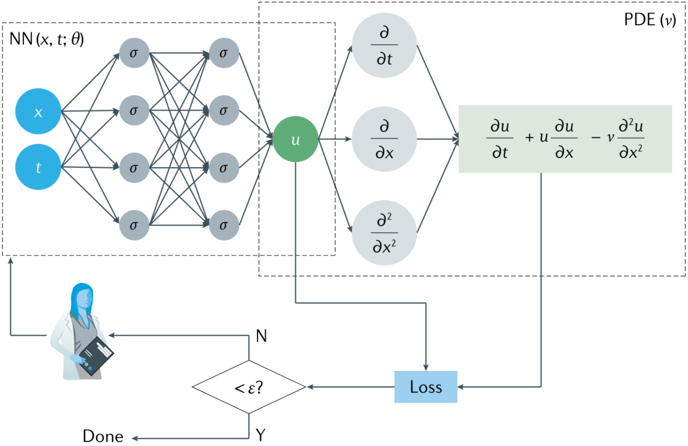
- solve a QP based on a local quadratic model of f and affine model of c
- equivalent to solve a linear system with gradient estimate $g_k \approx \nabla f(x_k)$:

$$\begin{bmatrix} H_k & J_k^T \\ J_k & 0 \end{bmatrix} \begin{bmatrix} d_k \\ y_k \end{bmatrix} = - \begin{bmatrix} g_k \\ c_k \end{bmatrix}$$

- update primal iterate with prescribed stepsizes $\{\alpha_k\}$:

$$x_{k+1} \leftarrow x_k + \alpha_k d_k$$

Application: physics-informed machine learning



Karniadakis et al. (2021)

Convergence to stationarity

with suitable choice of stepsizes $\{\alpha_k\}$,

$$\liminf_{k\to\infty} \mathbb{E}\big[\|\nabla f(X_k) + \nabla c(X_k)^T Y_k^{\text{true}}\| + \|c(X_k)\|\big] = 0$$

over some subsequence the expected stationarity measure vanishes, but

- it does not guarantee that $\{X_k\}$ converges in any sense
- the values $\{Y_k^{\text{true}}\}$ are not realized by the algorithm
- no information of the computed $\{Y_k\}$ is provided

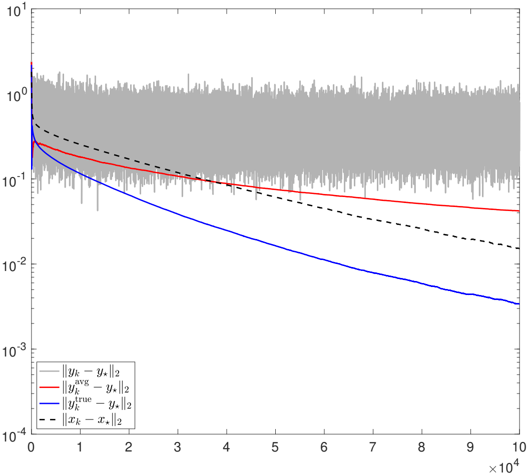
Lagrange multipliers are important for

- stationarity verification
- active-set identification
- *etc.*

Berahas et al. (2021)

Preview

we are going to see conditions that guarantee behavior as seen below



apply stochastic SQP to solve a constrained logistic regression problem

Outline

Primal iterates

Lagrange multipliers

Numerical demonstration

Main result I: short version

Almost-sure convergence of the primal iterates

$$\{X_k\} \xrightarrow{\text{a.s.}} x_\star$$

Assumptions

- a stationarity measure grows sufficiently away from $x_\star$
- $\{X_k\}$ remains within a small neighborhood of $x_\star$

respectively, these are assumptions about

- the problem, similar to "local convexity" or "generalized PL condition"
- algo. behavior: undesirable yet necessary in nonconvex, stochastic setting

Almost-sure convergence of the primal iterates

Convergence measure: *exact* penalty/merit function

$$\phi(x) = \tau f(x) + \|c(x)\|$$

Assumptions

- $\phi(x) \geq \phi(x_\star)$ for all $x \in \mathcal{B}(x_\star, \epsilon)$, with equality only if $x = x_\star$
- a generalized Polyak–Łojasiewicz condition holds for all $x \in \mathcal{B}(x_\star, \epsilon) \setminus \{x_\star\}$:

$$\phi(x) \leq \phi(x_\star) + \mu\big(\tau \|Z(x)^T \nabla f(x)\|^2 + \|c(x)\|\big)$$

  where $Z(x) \in \mathbb{R}^{n \times (n-m)}$ forms an orthogonal basis for $\mathrm{Null}(\nabla c(x)^T)$
- $\{X_k\} \subset \mathcal{B}(x_\star, \epsilon)$ almost surely: $\limsup_{k \to \infty} \{\|X_k - x_\star\|\} \leq \epsilon$

Main result I: almost-sure convergence of the primal iterates

$$\{\phi(X_k)\} \xrightarrow{\text{a.s.}} \phi(x_\star), \quad \{X_k\} \xrightarrow{\text{a.s.}} x_\star, \quad \left\{ \begin{bmatrix} \nabla f(X_k) + \nabla c(X_k) Y_k^{\mathsf{true}} \\ c(X_k) \end{bmatrix} \right\} \xrightarrow{\text{a.s.}} 0$$

Outline

Multipliers as a (noisy) mapping of the primal iterates

standard analysis of SQP shows that

$$Y_k = M_k(H_k(\nabla c(X_k)^\dagger)^T c(X_k) - G_k) \in \mathbb{R}^m,$$

where M_k is a product of a pseudoinverse and a projection matrix:

$$M_k = \nabla c(X_k)^\dagger (I - H_k Z_k (Z_k^T H_k Z_k)^{-1} Z_k^T) \in \mathbb{R}^{m \times n},$$

and Z_k is a basis for $\mathrm{Null}(\nabla c(X_k)^T)$

if $\{X_k\} \xrightarrow{\text{a.s.}} x_\star$, then one would expect

- $\{Y_k^{\text{true}}\} \xrightarrow{\text{a.s.}} y_\star$ (as above with $\nabla f(X_k)$ in place of G_k)
- $\{Y_k\}$ noisy with error proportional to error in stochastic gradient estimators

Initial result

Assumptions: $(x_\star, y_\star)$ is a stationary point, and in $\mathcal{B}(x_\star, \epsilon)$,

- $H_k = \mathcal{H}(X_k)$ is defined by a (locally) Lipschitz continuous function $\mathcal{H}$
- $M_k = \mathcal{M}(X_k)$ is defined by a (locally) Lipschitz continuous function $\mathcal{M}$

One-iteration analysis: if $X_k \in \mathcal{B}(x_\star, \epsilon)$, then

$$\begin{aligned} \|Y_k - y_\star\| &\leq \kappa_y \|X_k - x_\star\| + r^{-1} \|\nabla f(X) - G_k\| \\ \|Y_k^{\text{true}} - y_\star\| &\leq \kappa_y \|X_k - x_\star\|, \end{aligned}$$

where $(\kappa_y, r) \in \mathbb{R}_{++} \times \mathbb{R}_{++}$ are constants

unfortunately, this means

- $\{Y_k\}$ *always* has error
- $\{Y_k^{\text{true}}\}$ converges if $\{X_k\}$ does, but are not realized (require $\nabla f(X_k)$)

The averaged Lagrange multipliers

Idea: does averaging help reduce stochastic gradient errors?

- if $X_k = x_\star$ for all $k \in \mathbb{N}$, one can leverage classical central limit theorem
- yet, in practice, multipliers are not IID estimators of $y_\star$

Martingale central limit theorem: $\frac{1}{k}\mathbb{E}\left[\|\sum_{i=1}^{k} u_i\|\right] \xrightarrow{\text{a.s.}} 0$ if

$$\tfrac{1}{k}\mathbb{E}\left[\|u_i\|^2\right] < \infty, \qquad \left\{\tfrac{1}{k}\sum_{i=1}^{k}\mathbb{E}\left[\|u_i\|^2 \mathbf{1}_{\left\{\frac{\|u_i\|}{\sqrt{k}} > \delta\right\}}\right]\right\} \xrightarrow{\text{p}} 0,$$

$$\left\{\tfrac{1}{k}\sum_{i=1}^{k}\mathbb{E}[u_i u_i^T | \mathcal{F}_i]\right\} \xrightarrow{\text{p}} \Sigma, \qquad \sup_{k\in\mathbb{N}} \mathbb{E}\left[\left\|\sum_{i=1}^{k}\frac{1}{\sqrt{k}}u_i\right\|^2\right] < \infty,$$

where $u_k := M_k(\nabla f(X_k) - G_k)$

Main result II: almost-sure convergence of multipliers

$$\{X_k\} \xrightarrow{\text{a.s.}} x_\star \qquad \Longrightarrow \qquad \{Y_k^{\textsf{true}}\} \xrightarrow{\text{a.s.}} y_\star, \text{ and } \{Y_k^{\textsf{avg}}\} \xrightarrow{\text{a.s.}} y_\star$$

Outline

Primal iterates

Lagrange multipliers

Numerical demonstration

Test problem

consider constrained logistic regression of the form

$$\begin{array}{ll} \text{minimize} & \frac{1}{N}\sum_{i=1}^{N} \log\left(1 + e^{-\gamma_i d_i^T x}\right) \\ \text{subject to} & Ax = b, \quad \|x\|_2^2 = 1, \end{array}$$

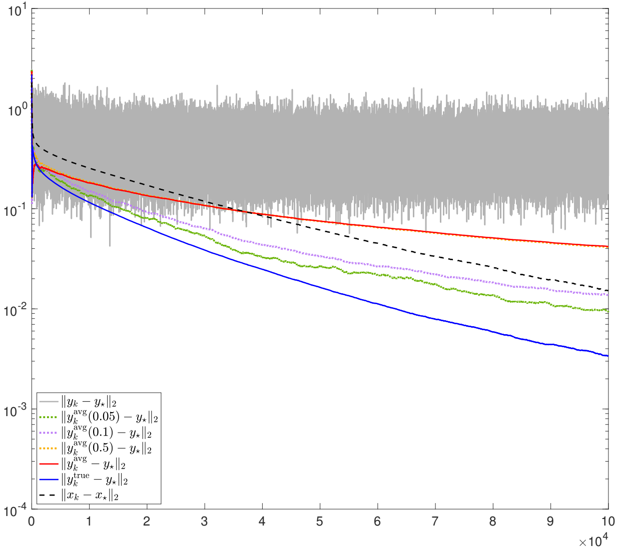
where $x \in \mathbb{R}^n$ is the optimization variable, and

- $D = \begin{bmatrix} d_1 & \cdots & d_N \end{bmatrix} \in \mathbb{R}^{n \times N}$ is a feature matrix
- $\gamma \in \mathbb{R}^N$ is a label vector
- $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$,

we plot prior sequences as well as Lagrange multiplier averages

$$Y_k^{\mathsf{avg}}(\epsilon) := \text{average of } Y_j\text{'s corresponding to } X_j\text{'s with } \|X_j - X_k\| \leq \epsilon$$

Numerical results



Summary

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & c(x) = 0, \end{array}$$

where f and c are continuously differentiable, and potentially nonconvex

for a stochastic SQP method, we present conditions that guarantee

- almost-sure convergence of $\{X_k\}$ to $x_\star$
- $\{\|Y_k - y_\star\|\}$ bounded by $\{\|G_k - \nabla f(X_k)\|\}$
- almost-sure convergence of $\{Y_k^{\text{true}}\}$ to $y_\star$
- almost-sure convergence of $\{Y_k^{\text{avg}}\}$ to $y_\star$