On the Almost-Sure Convergence of the Primal Iterates and Lagrange Multipliers in a Stochastic Sequential Quadratic Optimization Method

Frank E. Curtis, Lehigh University

joint work with

Xin Jiang (Lehigh), **Qi Wang** (Lehigh)

presented at

Modeling and Optimization: Theory and Applications (MOPTA) 2023

August 16, 2023

Collaborators and reference

- F. E. Curtis, X. Jiang, and Q. Wang, "Almost-sure convergence of iterates and multipliers in stochastic sequential quadratic optimization," https://arxiv.org/abs/2308.03687.

Convergence of random variables

Consider a stochastic process $\{V_k\}$ and random variable V defined with respect to $(\Omega, \mathcal{F}, \mathbb{P})$

Convergence in probability: $\{V_k\} \xrightarrow{p} V$ if and only if

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

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

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

Outline

Outline

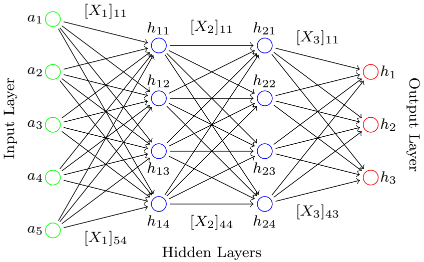
Stochastic optimization (unconstrained)

$$\min_{x \in \mathbb{R}^n} \; f(x)$$

where

- $f : \mathbb{R}^n \to \mathbb{R}$
- $f(x) = \mathbb{E}_\iota[F(x,\iota)]$ for all $x \in \mathbb{R}^n$
- ι has probability space $(\Omega_\iota, \mathcal{F}_\iota, \mathbb{P}_\iota)$
- $F : \mathbb{R}^n \times \Omega_\iota \to \mathbb{R}$
- $\mathbb{E}_\iota[\cdot]$ denotes expectation w.r.t. $\mathbb{P}_\iota$

e.g., $f(x) := \ell(\phi(x,a),b)$ in deep learning:



Stochastic approximation/gradient method

Robbins and Monro (1951) shows that for

- solving an equation with a unique root (and other assumptions)
- using an algorithm with unbiased derivative estimates
- and unsummable and square-summable step sizes (e.g., $\alpha = \mathcal{O}(1/k)$)

one can show

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

Cast into the context of minimization of (potentially nonconvex) $f : \mathbb{R}^n \to \mathbb{R}$, one can show that

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

Almost-sure convergence

Robbins and Siegmund (1971) proves the following lemma.

Lemma RS (simplified)

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $\{\mathcal{F}_k\}$ with $\mathcal{F}_k \subseteq \mathcal{F}_{k+1}$ for all $k \in \mathbb{N}$ be a sequence of sub-σ-algebras of $\mathcal{F}$. Let $\{R_k\}$, $\{P_k\}$, and $\{Q_k\}$ be sequences of nonnegative random variables such that for all $k \in \mathbb{N}$ the tuple (R_k, P_k, Q_k) is $\mathcal{F}_k$-measurable. If $\sum_{k=1}^{\infty} Q_k < \infty$ and, for all $k \in \mathbb{N}$, one has

$$\mathbb{E}[R_{k+1}|\mathcal{F}_k] \leq R_k - P_k + Q_k,$$

then, almost-surely, $\sum_{k=1}^{\infty} P_k < \infty$ and $\lim_{k\to\infty} R_k$ exists and is finite.

Therefore, it can be shown under certain assumptions that for

- stochastic approximation (solving an equation): $\{X_k\} \xrightarrow{a.s.} x_*$
- stochastic gradient (minimization): $\{\nabla f(X_k)\} \xrightarrow{a.s.} 0$ (Bertsekas and Tsitsiklis (2000))

Constrained stochastic optimization

$$\min_{x \in \mathbb{R}^n} \; f(x) \quad \text{s.t. } c(x) = 0$$

where

- $f(x) = \mathbb{E}_\iota[F(x,\iota)]$, as before
- c is continuously differentiable
- ∇f has Lipschitz constant L
- ∇c has Lipschitz constant Γ
- stationarity conditions:

$$\begin{aligned} \nabla f(x) + \nabla c(x) y &= 0 \\ c(x) &= 0 \end{aligned}$$

Algorithm : Stochastic SQP

1: choose $x_1 \in \mathbb{R}^n$, $\tau \in \mathbb{R}_{>0}$
2: **for** $k \in \{1, 2, \dots\}$ **do**
3: estimate gradient: $g_k \approx \nabla f(x_k)$
4: compute step: solve

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

5: choose step size: for small $\beta_k \in \mathbb{R}_{>0}$,

$$\alpha_k \leftarrow \frac{\beta_k \tau}{\tau L + \Gamma}$$

6: update iterate: set $x_{k+1} \leftarrow x_k + \alpha_k d_k$
7: **end for**

Motivation #1: Physics-informed learning (e.g., PINNs)

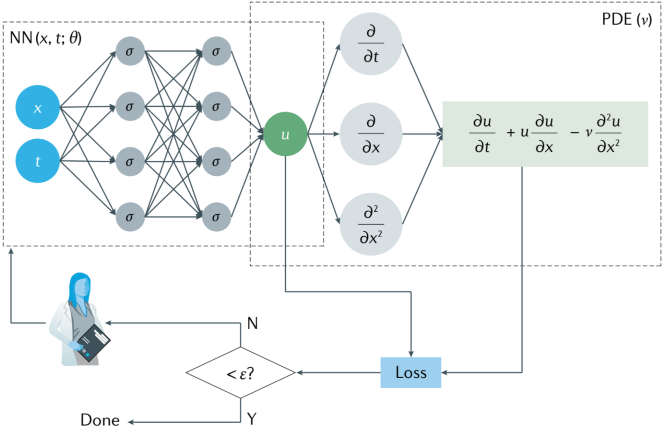


Photo: Karniadakis et al.

Motivation #2: Fair learning

Let

- Y be a feature vector
- A be a sensitive feature vector
- Z be the output/label

and consider

$$\min_{x\in\mathbb{R}^n} \mathbb{E}_{(Y,A,Z)}\left[\ell\left(\underbrace{\phi\left(x, \begin{bmatrix} Y \\ A \end{bmatrix}\right)}_{\hat{Z}}, Z\right)\right].$$

This loss might not be fair between subgroups in the population.

- Various criteria related to fairness (e.g., demographic parity, equalized odds, equalized opportunity) leading to various measures (e.g., accuracy equality, disparate impact, measures conditioned on outcome, measures conditioned on prediction)
- For example, in binary classification, disparate impact asks for the following *constraints* to hold:

$$\boxed{\mathbb{P}[\hat{Z} = z | A = 1] = \mathbb{P}[\hat{Z} = z | A = 0] \quad \text{for each} \quad z \in \{-1, 1\}}$$

Convergence to stationarity

Assumption

- *τ is sufficiently small*
- *$\{\beta_k\} = \mathcal{O}(1/k)$ with β_1 sufficiently small*

Theorem (Berahas, Curtis, Robinson, Zhou (2021))

$$\liminf_{k\to\infty} \mathbb{E}\left[\|\nabla f(X_k) + \nabla c(X_k)^T Y_k^{\text{true}}\|^2 + \|c(X_k)\|\right] = 0$$

This shows that over some sequence the expected stationarity measure vanishes, but

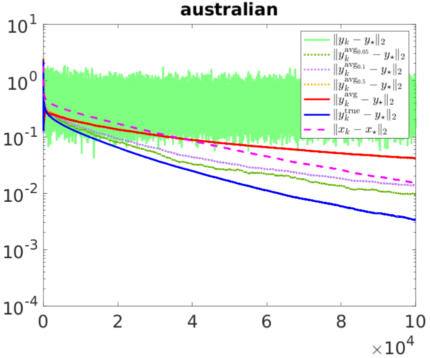
- it does not guarantee that $\{X_k\}$ converges in any sense and
- the values $\{Y_k^{\text{true}}\}$ are not realized by the algorithm, so
- it does not guarantee anything about $\{Y_k\}$

Multipliers are important for verifying stationarity, active-set identification, etc.

Preview

We are going to see conditions that guarantee behavior as seen below.

Solving a constrained logistic regression problem with the **australian** dataset from LIBSVM:



Outline

Short version

Main result: If

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

then

$$\{X_k\} \xrightarrow{a.s.} x_*.$$

Respectively, these are assumptions about

- the problem, similar to "local convexity" (generalized "P–Ł condition")
- the algorithm behavior(!)…necessary for the nonconvex setting to say anything about $\{X_k\}$

Merit function

Convergence of the algorithm is driven by the exact merit function

$$\phi_\tau(X) = \tau f(X) + \|c(X)\|$$

Reductions in a local model of ϕ_τ can be tied to a stationarity measure

$$\Delta q_\tau(X, \nabla f(X), H, D^{\text{true}}) \qquad \sim \qquad \|\nabla f(X) + \nabla c(X) Y\|^2 + \|c(X)\|$$

Lemma

Suppose $\mathbb{E}[G_k|\mathcal{F}_k] = \nabla f(X_k)$ *and* $\mathbb{E}[\|G_k - \nabla f(X_k)|\mathcal{F}_k\|^2] \leq \sigma^2$. *Lemma RS with*

$$P_k := \tfrac{\beta_k \tau}{\tau L + \Gamma} \Delta q_\tau(X_k, \nabla f(X_k), H_k, D_k^{\text{true}}), \quad Q_k := \tfrac{\beta_k^2 \tau^2 \sigma^2}{2\zeta(\tau L + \Gamma)}, \quad \text{and} \quad R_k := \phi_\tau(X_k) - \tau f_{\inf}$$

shows that, almost surely,

$$\lim_{k\to\infty} \{\phi_\tau(X_k)\} \text{ exists and is finite and}$$

$$\liminf_{k\to\infty} \Delta q_\tau(X_k, \nabla f(X_k), H_k, D_k^{\text{true}}) = 0$$

Almost-sure convergence of the primal iterates

If $\{X_k\}$ stays within a neighborhood of x_* almost surely, where x_* is a stationary point at which a generalization of the Polyak–Łojasiewicz condition holds, then almost-sure convergence follows:

Theorem

Suppose that there exists $x_* \in \mathcal{X}$ with $c(x_*) = 0$, $\mu \in \mathbb{R}_{>1}$, and $\epsilon \in \mathbb{R}_{>0}$ such that for all

$$x \in \mathcal{X}_{\epsilon,x_*} := \{x \in \mathcal{X} : \|x - x_*\|_2 \leq \epsilon\}$$

one finds that

$$\phi_\tau(x) - \phi_\tau(x_*) \begin{cases} = 0 & \text{if } x = x_* \\ \in (0, \mu(\tau\|Z(x)^T\nabla f(x)\|_2^2 + \|c(x)\|_2)] & \text{otherwise,} \end{cases}$$

where for all $x \in \mathcal{X}_{\epsilon,x_*}$ one defines $Z(x) \in \mathbb{R}^{n\times(n-m)}$ as some orthonormal matrix whose columns form a basis for the null space of $\nabla c(x)^T$. Then, if $\limsup_{k\to\infty}\{\|X_k - x_*\|_2\} \leq \epsilon$ almost surely, it follows that

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

Outline

Lagrange multipliers as a (noisy) mapping of the primal iterates

In a standard manner, it can be shown 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 of the derivative of c at X_k and a projection matrix:

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

If $\{X_k\} \xrightarrow{a.s.} x_*$, then one would expect

- $\{Y_k^{\text{true}}\} \xrightarrow{a.s.} y_*$ (i.e., 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

Assumption

Given $x_* \in \mathcal{X}$ *as a primal stationary point, there exist* $\epsilon \in \mathbb{R}_{>0}$, $\mathcal{H} : \mathbb{R}^n \to \mathbb{S}^n$, $L_{\mathcal{H}} \in \mathbb{R}_{>0}$, $\mathcal{M} : \mathbb{R}^n \to \mathbb{R}^{m \times n}$, *and* $L_{\mathcal{M}} \in \mathbb{R}_{>0}$ *such that:*

(i) $H_k = \mathcal{H}(X_k)$ *whenever* $X_k \in \mathcal{X}_{\epsilon, x_*}$;

(ii) $\|\mathcal{H}(x) - \mathcal{H}(\bar{x})\|_2 \leq L_{\mathcal{H}} \|x - \bar{x}\|_2$ *for all* $(x, \bar{x}) \in \mathcal{X}_{\epsilon, x_*} \times \mathcal{X}_{\epsilon, x_*}$;

(iii) $M_k = \mathcal{M}(X_k)$ *whenever* $X_k \in \mathcal{X}_{\epsilon, x_*}$; *and*

(iv) $\|\mathcal{M}(x) - \mathcal{M}(\bar{x})\|_2 \leq L_{\mathcal{M}} \|x - \bar{x}\|_2$ *for all* $(x, \bar{x}) \in \mathcal{X}_{\epsilon, x_*} \times \mathcal{X}_{\epsilon, x_*}$.

Theorem

Suppose (x_*, y_*) *is a stationary point. Then, for any* $k \in \mathbb{N}$, *one finds* $\|X_k - x_*\|_2 \leq \epsilon$ *implies*

$$\|Y_k - y_*\|_2 \leq \kappa_y \|X_k - x_*\|_2 + r^{-1} \|\nabla f(X_k) - G_k\|_2$$
$$\text{and} \quad \|Y_k^{\text{true}} - y_*\|_2 \leq \kappa_y \|X_k - x_*\|_2,$$

where $\kappa_y := \kappa_H L_c r^{-2} + L r^{-1} + \kappa_{\nabla f} L_{\mathcal{M}}$.

$\{Y_k\}$ has error and $\{Y_k^{\text{true}}\}$ is not computed! Average Y_k's?

Unfortunately, this means that

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

Idea: Average elements of $\{Y_k\}$?

- If $X_k = x_*$ for all $k \in \mathbb{N}$, then one can leverage the classical central limit theorem
- However, since $\{X_k\}$ is a random process, multipliers are not IID estimators of y_*

Martingale central limit theorem

Assumption

Suppose that $\{M_k\}$ *and* $\{\Delta_k\} = \{\nabla f(X_k) - G_k\}$ *satisfy*

$$\tfrac{1}{k}\mathbb{E}[\|M_i\Delta_i\|_2^2] < \infty \text{ for all } (k,i) \in \mathbb{N} \times [k],$$

$$\left\{\frac{1}{k}\sum_{i=1}^{k}\mathbb{E}\left[\|M_i\Delta_i\|_2^2 \mathbf{1}_{\left\{\frac{\|M_i\Delta_i\|_2}{\sqrt{k}} > \delta\right\}}\middle|\mathcal{F}_i\right]\right\} \xrightarrow{p} 0 \text{ for all } \delta \in \mathbb{R}_{>0},$$

$$\left\{\frac{1}{k}\sum_{i=1}^{k}\mathbb{E}[M_i\Delta_i\Delta_i^T M_i^T|\mathcal{F}_i]\right\} \xrightarrow{p} \Sigma \text{ for some } \Sigma \in \mathbb{S}^n, \text{ and}$$

$$\sup_{k\in\mathbb{N}} \mathbb{E}\left[\left\|\sum_{i=1}^{k}\frac{1}{\sqrt{k}}M_i\Delta_i\right\|_2^{\Theta}\right] < \infty \text{ for some } \Theta \in \mathbb{R}_{>1}.$$

True and average Lagrange multiplier convergence

Theorem

If the iterate sequence converges almost surely to x_*, i.e., $\{X_k\} \xrightarrow{a.s.} x_*$, then

$$\{Y_k^{\text{true}}\} \xrightarrow{a.s.} y_* \quad \textit{and} \quad \{Y_k^{\text{avg}}\} \xrightarrow{a.s.} y_*.$$

Outline

Test problem

Consider constrained logistic regression of the form

$$\min_{x \in \mathbb{R}^n} \frac{1}{N} \sum_{i=1}^{N} \log(1 + e^{-\gamma_i d_i^T x}) \quad \text{s.t.} \quad Ax = b, \quad \|x\|_2^2 = 1,$$

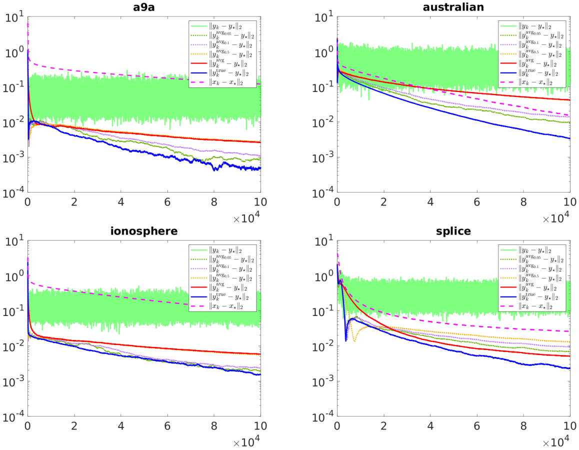
where

- $D = [d_1 \ \cdots \ d_N] \in \mathbb{R}^{n \times N}$ is a feature matrix
- $\gamma \in \mathbb{R}^N$ is a label vector
- $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$

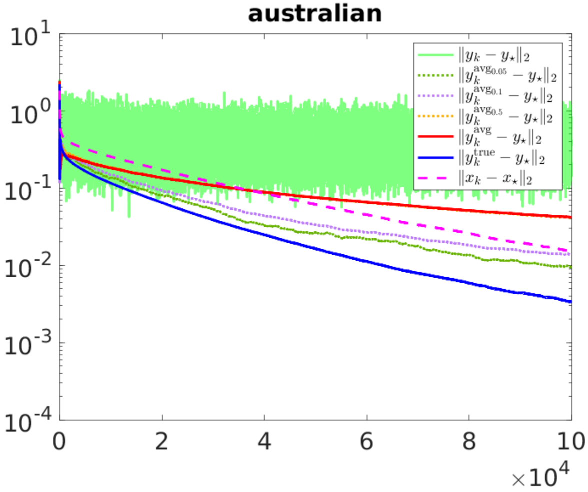
Consider prior sequences as well as Lagrange multiplier averages

$$Y_k^{\text{avg}_\epsilon} := \text{average of } Y_j\text{'s corresponding to } X_j\text{'s with } \|X_k - X_j\| \le \epsilon$$

LIBSVM datasets



australian dataset



Outline

Summary

$$\min_{x \in \mathbb{R}^n} f(x) \quad \text{s.t. } c(x) = 0$$

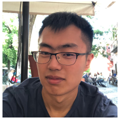
where

- $f(x) = \mathbb{E}_\iota[F(x, \iota)]$
- c is continuously differentiable

For Stochastic SQP, conditions that guarantee

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

Collaborators and references

- A. S. Berahas, F. E. Curtis, D. P. Robinson, and B. Zhou, "Sequential Quadratic Optimization for Nonlinear Equality Constrained Stochastic Optimization," *SIAM Journal on Optimization*, 31(2):1352–1379, 2021.
- A. S. Berahas, F. E. Curtis, M. J. O'Neill, and D. P. Robinson, "A Stochastic Sequential Quadratic Optimization Algorithm for Nonlinear Equality Constrained Optimization with Rank-Deficient Jacobians," https://arxiv.org/abs/2106.13015.
- F. E. Curtis, D. P. Robinson, and B. Zhou, "Inexact Sequential Quadratic Optimization for Minimizing a Stochastic Objective Subject to Deterministic Nonlinear Equality Constraints," https://arxiv.org/abs/2107.03512.
- F. E. Curtis, M. J. O'Neill, and D. P. Robinson, "Worst-Case Complexity of an SQP Method for Nonlinear Equality Constrained Stochastic Optimization," *Mathematical Programming* (online).
- F. E. Curtis, S. Liu, and D. P. Robinson, "Fair Machine Learning through Constrained Stochastic Optimization and an ϵ-Constraint Method," *Optimization Letters* (online).
- F. E. Curtis, D. P. Robinson, and B. Zhou, "Sequential Quadratic Optimization for Stochastic Optimization with Deterministic Nonlinear Inequality and Equality Constraints," https://arxiv.org/abs/2302.14790.
- F. E. Curtis, X. Jiang, and Q. Wang, "Almost-sure convergence of iterates and multipliers in stochastic sequential quadratic optimization," https://arxiv.org/abs/2308.03687.