Stochastic Algorithms with Adaptive Parameters for Solving Constrained Optimization Problems

Frank E. Curtis, Lehigh University

presented at

INFORMS Annual Meeting

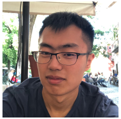
October 17, 2023

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, V. Kungurtsev, D. P. Robinson, and Q. Wang, "A Stochastic-Gradient-based Interior-Point Algorithm for Solving Smooth Bound-Constrained Optimization Problems," https://arxiv.org/abs/2304.14907.
- 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.

Outline

Outline

Stochastic algorithms

Consider an algorithm whose behavior (over an entire run) is dictated by a random draw from

$$\Gamma \times \Gamma \times \Gamma \times \cdots .$$

Our aim is to prove conclusions with respect to a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where

- $\Omega = \Gamma \times \Gamma \times \Gamma \times \cdots$;
- $\mathcal{F}$ is a σ-algebra on Ω, specifically, the set of events (i.e., measurable subsets of Ω); and
- $\mathbb{P} : \mathcal{F} \to [0, 1]$ is a probability measure.

Probability space $(\Omega, \mathcal{F}, \mathbb{P})$

One can understand $\Omega = \Gamma \times \Gamma \times \Gamma \times \cdots$ through the axiom of choice.

An algebra $\mathcal{A}$ on Ω is a collection of subsets of Ω that are

- closed under finite numbers of union operations ($X \in \mathcal{A}$ and $Y \in \mathcal{A}$ implies $X \cup Y \in \mathcal{A}$);
- closed under finite numbers of complement operations ($X \in \mathcal{A}$ implies $X^c \in \mathcal{A}$).

A σ-algebra $\mathcal{F}$ is an algebra that is also closed under countable union operations, i.e.,

$$X_i \in \mathcal{F} \text{ for all } i \in \mathbb{N} \text{ implies } \bigcup_{i \in \mathbb{N}} X_i \in \mathcal{F}.$$

The probability measure $\mathbb{P}$ has unit mass (i.e., $\mathbb{P}(\Omega) = 1$) and is countably additive in that

$$\mathbb{P}\left(\bigcup_{i \in \mathbb{N}} \mathcal{X}_i\right) = \sum_{i \in \mathbb{N}} \mathbb{P}(\mathcal{X}_i) \text{ for any sequence of disjoint events } \{\mathcal{X}_i\}.$$

Example

Consider for simplicity the setting of only two iterations with flip-of-a-coin randomness, so

$$\Omega = \Gamma \times \Gamma = \{0,1\} \times \{0,1\}.$$

The σ-algebra $\mathcal{F}$ of all possible events has the form

$$\mathcal{F} = 2^{\Omega} = \left\{ \begin{array}{l} \emptyset, \\ \{00\}, \{01\}, \{10\}, \{11\}, \\ \{00,01\}, \{00,10\}, \{00,11\}, \{01,10\}, \{01,11\}, \{10,11\}, \\ \{00,01,10\}, \{00,01,11\}, \{00,10,11\}, \{01,10,11\}, \\ \{00,01,10,11\} \equiv \Omega \end{array} \right\}.$$

A corresponding probability measure $\mathbb{P}$ would give us probabilities for all possible events.

Sub-σ-algebras

A sub-σ-algebra of a σ-algebra $\mathcal{F}$ is any subset of $\mathcal{F}$ that is also a σ-algebra.

Using our example, one can consider the information before the first iteration as

$$\mathcal{F}_0 = \{\emptyset, \Omega\} \subset \mathcal{F}.$$

Similarly, one can consider the information after the first iteration as

$$\mathcal{F}_1 = 2^{\{0,1\}} \times \{0,1\} = \begin{Bmatrix} \emptyset, \\ \{0\}, \\ \{1\}, \\ \{0,1\} \end{Bmatrix} \times \{0,1\} = \begin{Bmatrix} \emptyset, \\ \{00, 01\}, \\ \{10, 11\}, \\ \{00, 01, 10, 11\} \equiv \Omega \end{Bmatrix}.$$

And again, one can consider the information after the second iteration as

$$\mathcal{F}_2 = 2^{\{0,1\}} \times 2^{\{0,1\}} = \mathcal{F}.$$

Overall, one finds that $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \mathcal{F}_2 \equiv \mathcal{F}$.

Outline

Stochastic Gradient method

Let's return to: An algorithm whose behavior (over an entire run) is dictated by a random draw from

$$\Omega_1^\infty = \Gamma \times \Gamma \times \Gamma \times \cdots .$$

Consider $\min_{x \in \mathbb{R}^n} f(x)$, where $\inf_{x \in \mathbb{R}^n} f(x) > -\infty$ and $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$ is Lipschitz continuous with constant L.

Algorithm SG : Stochastic Gradient method

1: choose an initial point $x_1 \in \mathbb{R}^n$ and step sizes $\{\alpha_k\} > 0$
2: **for** $k \in \{1, 2, \ldots\}$ **do**
3: set $x_{k+1} \leftarrow x_k - \alpha_k g_k$, where $g_k \approx \nabla f(x_k)$
4: **end for**

One can view any $\{(x_k, g_k)\}$ as a realization of $\{(X_k, G_k)\}$, where for all $k \in \mathbb{N}$

$$x_k = X_k(\omega) \text{ and } g_k = G_k(\omega) \text{ given } \omega \in \Omega.$$

Filtration

What is the associated sequence of sub-σ-algebras?

- The information before the first iteration is simply given by

$$\mathcal{F}_0 = \{\emptyset, \Omega_1^\infty\}.$$

- After the stochastic gradient computation in the first iteration, let

$$\mathcal{F}_1 = 2^\Gamma \times \Omega_2^\infty.$$

- After the stochastic gradient computation in the second iteration, let

$$\mathcal{F}_2 = 2^\Gamma \times 2^\Gamma \times \Omega_3^\infty$$

- ... and so on.

Random variables measurable with respect to $\mathcal{F}_k$

Consider a random variable for which a realization is determined by the draw, e.g., X_k.

- $\mathcal{F}_j$ for all $j < k$ *does not* give enough information about X_k.
- $\mathcal{F}_j$ for all $j \geq k$ *does* give enough information about X_k.

We say X_k is measurable with respect to $\mathcal{F}_k$ if and only if all "inverses" of X_k are in $\mathcal{F}_k$.

- For our purposes going forward, it is sufficient to understand that this means

$$X_k = \mathbb{E}[X_k | \mathcal{F}_k] \text{ for all } k \in \mathbb{N}.$$

For the stochastic gradient method, one finds that

- X_k is $\mathcal{F}_k$-measurable for all $k \in \mathbb{N}$
- G_k is $\mathcal{F}_{k+1}$-measurable for all $k \in \mathbb{N}$.

Convergence of SG

Let $\mathbb{E}[\cdot]$ denote expectation with respect to $\mathbb{P}[\cdot]$.

Assumption

For all $k \in \mathbb{N}$, one has that

- $\mathbb{E}[G_k|\mathcal{F}_k] = \nabla f(X_k)$ *and*
- $\mathbb{E}[\|G_k\|_2^2|\mathcal{F}_k] \leq M + M_{\nabla f}\|\nabla f(X_k)\|_2^2$

By Lipschtiz continuity of ∇f and construction of the algorithm, one finds

$$\begin{aligned} f(X_{k+1}) - f(X_k) &\leq \nabla f(X_k)^T(X_{k+1} - X_k) + \tfrac{1}{2}L\|X_{k+1} - X_k\|_2^2 \\ &= -\alpha_k \nabla f(X_k)^T G_k + \tfrac{1}{2}\alpha_k^2 L\|G_k\|_2^2 \\ \implies \mathbb{E}[f(X_{k+1})|\mathcal{F}_k] - f(X_k) &\leq -\alpha_k\|\nabla f(X_k)\|_2^2 + \tfrac{1}{2}\alpha_k^2 L\mathbb{E}[\|G_k\|_2^2|\mathcal{F}_k] \\ &\leq -\alpha_k\|\nabla f(X_k)\|_2^2 + \tfrac{1}{2}\alpha_k^2 L(M + M_{\nabla f}\|\nabla f(X_k)\|_2^2), \end{aligned}$$

where the last inequalities follow by the assumption and since $f(X_k)$ and $\nabla f(X_k)$ are $\mathcal{F}_k$-measurable.

SG theory

Taking total expectation, one arrives at

$$\mathbb{E}[f(X_{k+1}) - f(X_k)] \leq -\alpha_k(1 - \tfrac{1}{2}\alpha_k L M_{\nabla f})\mathbb{E}[\|\nabla f(X_k)\|_2^2] + \tfrac{1}{2}\alpha_k^2 LM$$

Theorem

$$\alpha_k = \frac{1}{LM_{\nabla f}} \qquad \Longrightarrow \mathbb{E}\left[\frac{1}{k}\sum_{j=1}^{k}\|\nabla f(X_j)\|_2^2\right] \leq M_k \xrightarrow{k\to\infty} \mathcal{O}\left(\frac{M}{M_{\nabla f}}\right)$$

$$\alpha_k = \Theta\left(\frac{1}{k}\right) \qquad \Longrightarrow \mathbb{E}\left[\frac{1}{\left(\sum_{j=1}^{k}\alpha_j\right)}\sum_{j=1}^{k}\alpha_j\|\nabla f(X_j)\|_2^2\right] \to 0$$

$$\Longrightarrow \liminf_{k\to\infty}\ \mathbb{E}[\|\nabla f(X_k)\|_2^2] = 0$$

(further steps) *and* $\nabla f(X_k) \to \infty$ *almost surely.*

Outline

Sequential quadratic optimization (SQP)

Consider

$$\boxed{\begin{aligned}\min_{x\in\mathbb{R}^n}\ & f(x)\\ \text{s.t. } & c(x)=0\end{aligned}}$$

with $J \equiv \nabla c$ and H positive definite over $\text{Null}(J)$, either viewpoint

$$\boxed{\begin{bmatrix}\nabla f(x)+J(x)^T y\\ c(x)\end{bmatrix}=0} \quad \text{or} \quad \boxed{\begin{aligned}\min_{d\in\mathbb{R}^n}\ & f(x)+\nabla f(x)^T d+\tfrac{1}{2}d^T H d\\ \text{s.t. } & c(x)+J(x)d=0\end{aligned}}$$

leads to the same "Newton-SQP system"

$$\boxed{\begin{bmatrix}H_k & J_k^T\\ J_k & 0\end{bmatrix}\begin{bmatrix}d_k\\ y_k\end{bmatrix}=-\begin{bmatrix}\nabla f(x_k)\\ c_k\end{bmatrix}}$$

Stochastic SQP

Algorithm guided by merit function with adaptive parameter τ defined by

$$\phi(x,\tau) = \tau f(x) + \|c(x)\|_1$$

Algorithm : Stochastic SQP

choose $x_1 \in \mathbb{R}^n$, $\tau_0 \in (0,\infty)$, $\{\beta_k\} \in (0,1]^{\mathbb{N}}$
for $k \in \{1,2,\dots\}$ **do**
    compute step: solve

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

    update merit parameter: set τ_k to ensure

$$\phi'(x_k,\tau_k,d_k) \le -\Delta q(x_k,\tau_k,g_k,d_k) \ll 0$$

    compute step size: set

$$\alpha_k = \Theta\left(\frac{\beta_k \tau_k}{\tau_k L_{\nabla f} + L_{\nabla c}}\right)$$

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

Deterministic vs. stochastic setting

Convergence analysis hinges on the behavior of the sequence $\{\mathcal{T}_k\}$.

Deterministic setting under nice *function* assumptions:

- $\tau_k = \tau_{\min}$ for all $k \geq k_{\min}$ for some $\tau_{\min} \in (0, \infty)$ and $k_{\min} \in \mathbb{N}$.
- Note, however, that $(\tau_{\min}, k_{\min})$ is NOT knowable *a priori and* depends on x_1.

Stochastic setting under nice *function* assumptions, but general *noise* assumptions:

- $E_{\text{big}} := \{\{\mathcal{T}_k\}$ decreases, but not enough$\}$
- $E_{\text{good}} := \{\{\mathcal{T}_k\}$ decreases sufficiently and does not vanish to zero$\}$
- $E_{\text{zero}} := \{\{\mathcal{T}_k\}$ vanishes to zero$\}$

Even the good case is not straightforward!

- Imagine a sequence of events in E_{good} over which $k_{\min} \to \infty$.

Assumptions that are reasonable?

Need to have an honest discussion in the community about what assumptions are reasonable.

Prove probability of events E_{big}, E_{good}, and E_{zero}?

- Seems quite impossible in the general nonconvex landscape.
- If this means that we abandon certain settings/algorithms, that's a shame.

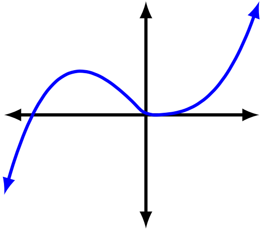
$E_{\text{big}} \cup E_{\text{good}}$ essentially requires bounded noise.

- Enough to focus on bounded noise over a finite number of iterations?
- Enough to focus on the event that the noise remains bounded (over infinite iterations)?

Reality check

Note that even in the deterministic setting, some assumptions can be unreasonable.

- The merit function for $\min_{x \in \mathbb{R}} \; x^3$ s.t. $x \geq 0$ is not bounded below.



- People understand that in practice certain safeguards can be incorporated.

For other stochastic algorithms, noise assumptions are not verifiable in practice.

- For example, probabilistic guarantee of certain accuracy.

Proposal

My feeling is that it should be considered sufficient to analyze the algorithm under reasonable events, e.g.,

$$E := E(\tau_{\min}, k_{\min}) := \{\mathcal{T}_k = \mathcal{T} \text{ for sufficiently small } \mathcal{T} \in [\tau_{\min}, \infty) \text{ for all } k \geq k_{\min}\}.$$

(Recall that $\{\tau_k\}$ can be bounded below in deterministic setting, although $k_{\min}$ not known.)

For the purposes of analysis, this involves focusing on the *trace* σ-algbra $\mathcal{G} := \mathcal{F} \cap \{E\}$.

- Redefine the sequence of sub-σ-algebras as $\{\mathcal{G}_k\}$, where

$$\mathcal{G}_k := \mathcal{F}_k \cap \{E\} \quad \text{for all} \quad k \in \mathbb{N}.$$

- **Key**: The *macroparameter* $\mathcal{T} \geq \tau_{\min}$ is $\mathcal{G}_{k_{\min}}$-measurable.

Outline

Stochastic Processes

Stochastic Gradient Method

Stochastic Methods with Adaptive Parameters

Conclusion

Summary

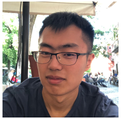
Discussed procedures for analyzing stochastic algorithms for smooth nonconvex optimization.

- Each realization of the algorithm corresponds to a draw from $\Omega = \Gamma \times \Gamma \times \Gamma \times \cdots$.
- Step-by-step analysis conducted with sequence of sub-σ-algebras $\{\mathcal{F}_k\}$.

Algorithms with random *macroparameters* cannot satisfy idealized assumptions.

- Need to consider what assumptions are reasonable in practice
- . . . or else we throw out the baby (good algorithms)
- . . . with the bath water (unreasonable demands for analysis)!

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, V. Kungurtsev, D. P. Robinson, and Q. Wang, "A Stochastic-Gradient-based Interior-Point Algorithm for Solving Smooth Bound-Constrained Optimization Problems," https://arxiv.org/abs/2304.14907.
- 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.