Stochastic-Gradient-based Interior-Point Algorithms

Frank E. Curtis, Lehigh University

presented at

INFORMS Optimization Society Conference

March 23, 2024

Collaborators and references

Submitted paper (second-round review):

- 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.

Working paper:

- F. E. Curtis, X. Jiang, and Q. Wang, "Single-Loop Deterministic and Stochastic Interior-Point Algorithms for Nonlinearly Constrained Optimization."

Outline

Outline

Stochastic gradient method

Consider $\min_{x \in \mathbb{R}^n} f(x)$, where $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$ 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** $k \in \{1, 2, \dots\} =: \mathbb{N}$ **do**
3: set $x_{k+1} \leftarrow x_k - \alpha_k g_k$, where $g_k \approx \nabla f(x_k)$
4: **end for**

Algorithm behavior is defined by $(\Omega, \mathcal{F}, \mathbb{P})$, where

- $\Omega = \Gamma \times \Gamma \times \Gamma \times \cdots$ (sequence of draws determining stochastic gradients);
- $\mathcal{F}$ 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.

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

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

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

Analyze through an associated sequence of sub-σ-algebras:

$$\mathcal{F}_0 = \{\emptyset, \Omega\}, \quad \mathcal{F}_1 = 2^\Gamma \times \Omega, \quad \mathcal{F}_2 = 2^\Gamma \times 2^\Gamma \times \Omega, \quad \dots$$

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 Lipschitz continuity of ∇f and construction of the algorithm, one finds

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

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

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

Theorem

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

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

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

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

Outline

- Stochastic Gradient Method
- Single-Loop Interior-Point (SLIP) Method
- Stochastic Setting
- Conclusion

Motivation

Interior-point methods are the workhorse for large-scale nonlinearly constrained optimization.

- Ipopt, Knitro, LOQO, etc.

As far as we are aware, there were no stochastic interior-point methods with convergence guarantees.

Huh? Why not?

- Stochastic optimization with nonlinear, nonconvex constraints is not well studied.
- For large-scale problems, people focus on projection methods, manifold methods, etc.
- Stochastic-gradient-based algorithms require gradients to be bounded and Lipschitz continuous
- ... but the typical (e.g., logarithmic) barrier function has neither property.

Bound-constrained setting

Given $f : \mathbb{R}^n \to \mathbb{R}$ and $(l, u) \in \mathbb{R}^n \times \mathbb{R}^n$ with $l < u$, consider

$$\boxed{\begin{aligned} \min_{x \in \mathbb{R}^n} \ & f(x) \\ \text{s.t. } & l \leq x \leq u \end{aligned}}$$

If x is a minimizer, then for some (y, z) one has

$$\nabla f(x) - y + z = 0, \quad 0 \leq (x - l) \perp y \geq 0, \quad 0 \leq (u - x) \perp z \geq 0.$$

(In what follows, we can handle infinite bounds, but consider finite bounds for simplicity....)

Textbook algorithm

For a given $\mu \in \mathbb{R}_{>0}$, consider the barrier-augmented function

$$\phi(x,\mu) = f(x) - \mu \sum_{i=1}^{n} \log(x_i - l_i) - \mu \sum_{i=1}^{n} \log(u_i - x_i).$$

Algorithm IPM : Interior-point method (textbook version)

1: choose an initial point $x_1 \in \mathbb{R}^n$ and barrier parameter $\mu_0 \in \mathbb{R}_{>0}$
2: **for** $k \in \{1,2,\dots\}$ **do**
3: **if** $\|\nabla_x \phi(x_k, \mu_{k-1})\|_2 \leq \theta \mu_{k-1}$ **then** set $\mu_k \ll \mu_{k-1}$ **else** set $\mu_k \leftarrow \mu_{k-1}$
4: compute descent direction d_k (e.g., $-\nabla\phi(x_k,\mu_k)$)
5: set $\alpha_{k,\max} \in (0,1]$ by fraction-to-the-boundary rule to ensure

$$x_k + \alpha_{k,\max} d_k \in [l + \epsilon x_k, u - \epsilon x_k]$$

6: set $\alpha_k \in (0, \alpha_{k,\max}]$ to ensure sufficient decrease $\phi(x_{k+1},\mu_k) \ll \phi(x_k,\mu_k)$
7: **end for**

Major challenges for the stochastic setting

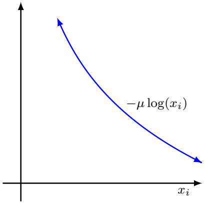
Stationarity test:

- Computing $\|\nabla_x \phi(x_k, \mu_{k-1})\|_2$ is intractable
- Could estimate it using a stochastic gradient, but then a probabilistic guarantee, at best

Fraction-to-the-boundary rule:

- Tying fraction to current iterate x_k leads to issues
- ... stochastic gradients could push iterate sequence to boundary too quickly

Unbounded gradients and lack of Lipschitz continuity:



Our approach

Our approach is based on two coupled ideas:

- prescribed decreasing barrier parameter sequence $\{\mu_k\} \searrow 0$ (single-loop algorithm!)
- prescribed $\{\theta_k\} \searrow 0$ and enforcing

$$x_{k+1} \in \mathcal{N}_{[l,u]}(\theta_k) := \{x \in \mathbb{R}^n : l + \theta_k \leq x \leq u - \theta_k\}$$

"Wait! I thought interior-points worked well because of their complexity properties?!"

- This algorithm is completely different and doesn't have those properties
- Is it worthwhile to do this? (Our experiments say yes!)

Proposed algorithm

Algorithm SLIP : Single-loop interior-point method

1: choose an initial point $x_1 \in \mathbb{R}^n$, $\{\mu_k\} \searrow 0$, $\{\theta_k\} \searrow 0$
2: **for** $k \in \{1, 2, \dots\}$ **do**
3: compute descent direction d_k (e.g., estimating $-\nabla \phi(x_k, \mu_k)$)
4: set
$$\alpha_k \leftarrow \frac{1}{L + 2\mu_k \theta_k^{-2}}$$
5: set $\gamma_k \in (0, 1]$ to ensure
$$x_{k+1} \leftarrow x_k + \gamma_k \alpha_k d_k \in \mathcal{N}_{[l,u]}(\theta_k)$$
6: **end for**

*Paper considers a more general framework; this is a simplified example

Critical lemmas, deterministic setting

Lemma

For all $k \in \mathbb{N}$, one finds for $L_k := L + 2\mu_k\theta_k^{-2}$ that

$$\phi(x_{k+1},\mu_k) \leq \phi(x_k,\mu_k) + \nabla_x\phi(x_k,\mu_k)^T(x_{k+1}-x_k) + \tfrac{1}{2}L_k\|x_{k+1}-x_k\|_2^2,$$

$$\text{so } \{\alpha_k\} = \{L_k^{-1}\} \implies \phi(x_{k+1},\mu_{k+1}) \leq \phi(x_k,\mu_k) - \tfrac{1}{2}\gamma_k\alpha_k\|\nabla_x\phi(x_k,\mu_k)\|_2^2.$$

Lemma

For all $k \in \mathbb{N}$, one finds that γ_k is bounded below by the minimum of 1 and

$$\alpha_k^{-1}\left(\frac{\frac{1}{2}\mu_k\Delta}{\mu_k + \frac{1}{2}\kappa_{\nabla f}\Delta} - \theta_k\right)(\kappa_{\nabla f} + \mu_k\theta_{k-1}^{-1})^{-1}.$$

Thus, with $t \in [-1,0)$, $\{\mu_k\} = \{\mu_1 k^t\}$, $\{\theta_{k-1}\} = \{\theta_0 k^t\}$, and $\{\alpha_k\} = \{L_k^{-1}\}$, one finds that

$$\sum_{k=1}^{\infty}\gamma_k\alpha_k = \infty \quad \text{and} \quad \{\mu_k\theta_{k-1}^{-1}\} \quad \text{is bounded.}$$

Convergence guarantee, deterministic setting

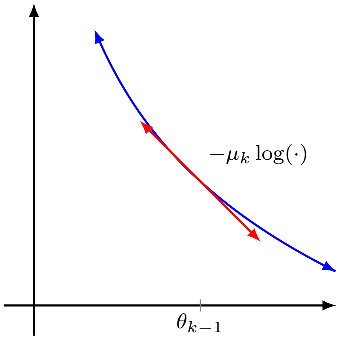
Theorem

One finds that

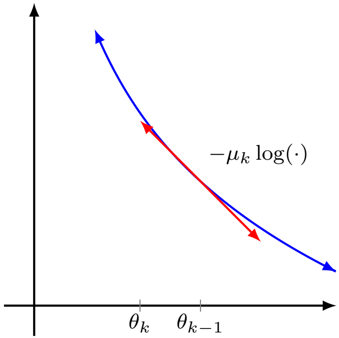
$$\liminf_{k\to\infty} \|\nabla_x \phi(x_k, \mu_k)\|_2^2 = 0.$$

Consequently, for any infinite-cardinality set $\mathcal{K} \subseteq \mathbb{N}$ *such that* $\{\nabla_x \phi(x_k, \mu_k)\}_{k\in\mathcal{K}} \to 0$ *and* $\{x_k\}_{k\in\mathcal{K}} \to \bar{x}$*, the limit point* $\bar{x}$ *is a KKT point (i.e., there exists* $\bar{y}$ *and* $\bar{z}$ *such that* $(\bar{x}, \bar{y}, \bar{z})$ *satisfies KKT conditions).*

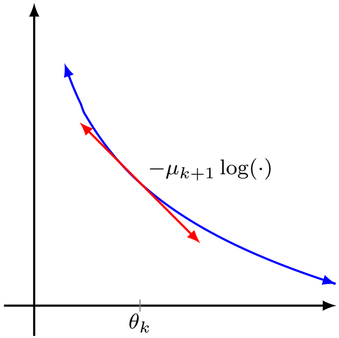
Why does it work?



Why does it work?



Why does it work?



Outline

Stochastic Gradient Method

Single-Loop Interior-Point (SLIP) Method

Stochastic Setting

Conclusion

Stochastic setting

In the stochastic setting, the algorithm parameters need to be chosen carefully!

- Notably, γ_k needs to be chosen based on knowledge of noise bound.
- Step-size sequence $\{\alpha_k\}$ can no longer decrease at same rate as $\{\mu_k\}$
- . . . needs to decrease more slowly (although rates can be arbitrarily close).

Convergence guarantee, stochastic setting

Theorem

Suppose $t \in (-1, -\frac{1}{2})$ *and* $t_\alpha \in (-\infty, 0)$ *have*

$$t + t_\alpha \in [-1, 0) \quad \textit{and} \quad t + 2t_\alpha \in (-\infty, -1)$$

and for some $\sigma \in \mathbb{R}_{>0}$ *one has for all* $k \in \mathbb{N}$ *that*

$$\mathbb{E}[G_k | \mathcal{F}_k] = \nabla f(X_k) \quad \textit{and} \quad \|G_k - \nabla f(X_k)\|_2 \leq \sigma.$$

Then, with $\{\mu_k\} = \{\mu_1 k^t\}$, $\{\theta_{k-1}\} = \{\theta_0 k^t\}$, *and* $\{\alpha_k\} = \{L_k^{-1} k^{t_\alpha}\}$, *one finds that*

$$\liminf_{k \to \infty} \|\nabla_x \phi(X_k, \mu_k)\|_2^2 = 0 \quad \textit{almost surely.}$$

Consequently, considering any realization $\{x_k\}$ *of* $\{X_k\}$, *for any infinite-cardinality set* $\mathcal{K} \subseteq \mathbb{N}$ *such that* $\{\nabla_x \phi(x_k, \mu_k)\}_{k \in \mathcal{K}} \to 0$ *and* $\{x_k\}_{k \in \mathcal{K}} \to \bar{x}$, *the limit point* $\bar{x}$ *is a KKT point.*

Numerical experiments

Compare SLIP with a projected stochastic gradient method (PSGM) for which

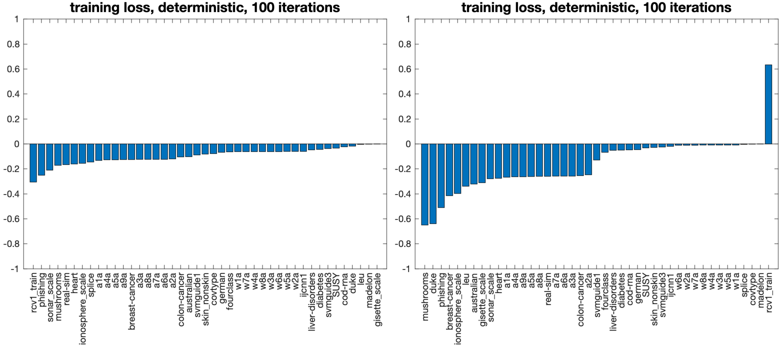
$$x_{k+1} \leftarrow \text{Proj}_{[l,u]}(x_k + \alpha_k d_k).$$

Experiments involve:

- binary classification problems with LIBSVM datasets
- two classifiers:
  - logistic regression (convex) and
  - neural network with one hidden layer and cross-entropy loss (nonconvex)
- performance measure

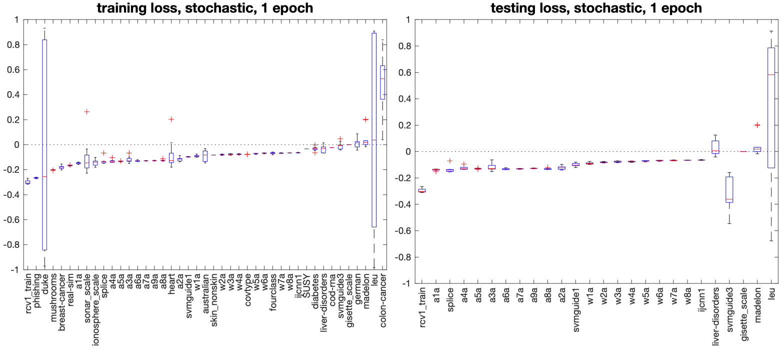
$$\frac{f(x_{\text{end}}^{\text{SLIP}}) - f(x_{\text{end}}^{\text{PSGM}})}{\max\{f(x_{\text{end}}^{\text{SLIP}}), f(x_{\text{end}}^{\text{PSGM}}), 1\}} \in (-1, 1)$$

Deterministic setting



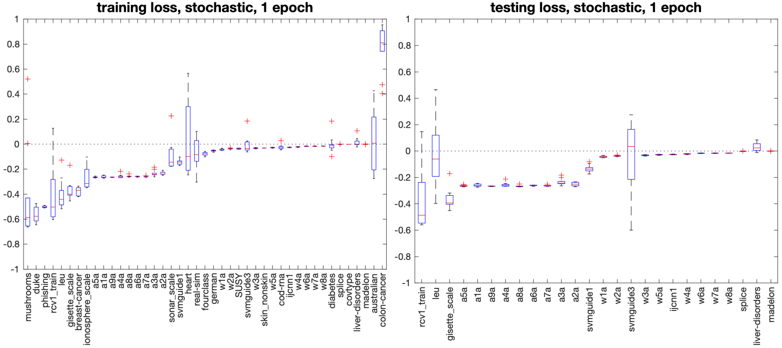
Relative performance of SLIP and PSGM, deterministic setting, training logistic regression (left) and neural network models with one hidden layer with cross-entropy loss (right).

Stochastic setting, logistic regression



Relative performance of SLIP and PSGM, stochastic setting (10 runs each), training logistic regression models; among 43 training datasets, 26 have testing datasets.

Stochastic setting, neural network with cross-entropy loss



Relative performance of SLIP and PSGM, stochastic setting (10 runs each), training neural network models (with one hidden layer) with cross-entropy loss; among 43 training datasets, 26 have testing datasets.

Outline

Summary

Presented a single-loop interior-point method for solving bound-constrained problems, with

- prescribed barrier and "neighborhood" parameter sequences,
- no need for stationarity tests, fraction-to-the-boundary rules, or line searches,
- convergence guarantees in deterministic and stochastic settings, and
- promising numerical performance!

What about the generally constrained setting???

- We've done it! (Happy to discuss outside of the talk.)
- Paper is forthcoming soon.

Collaborators and references




Submitted paper (second-round review):

- 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.

Working paper:

- F. E. Curtis, X. Jiang, and Q. Wang, "Single-Loop Interior-Point Methods for Deterministic and Stochastic Nonlinearly Constrained Optimization."