Stochastic Algorithms for Constrained Continuous Optimization: Stochastic-Gradient-based Interior-Point Algorithms

Frank E. Curtis, Lehigh University

presented at

ESI Workshop: One World Optimization Seminar

June 4, 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

Motivation

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

- Ipopt, Knitro, LOQO, etc.

Before our work, there were no stochastic interior-point methods with convergence guarantees.†

Why not?

- Stochastic algorithms for constrained optimization are not widely studied
- ...except for projection methods, manifold-based methods, and conditional gradient methods.
- Stochastic-gradient-based algorithms require gradients to be bounded and Lipschitz continuous
- ...but barrier functions (e.g., logarithmic barrier) have neither property.

In our work and this talk, we focus on the bound-constrained case.

- We will end with the additional challenges for the generally constrained case.

† One attempt was made, but there was a flaw in the algorithm and proof.

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

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

(We can handle infinite bounds, but consider finite bounds for simplicity. . . .)

Textbook algorithm

For all $\mu \in \mathbb{R}_{>0}$, 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 $x_1 \in (l,u)$ and barrier parameter $\mu_0 \in \mathbb{R}_{>0}$
2: **for** all $k \in \{1,2,\dots\}$ **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

$$x_k + \alpha_{k,\max} d_k - l \geq \epsilon(x_k - l) \quad \text{and} \quad u - (x_k + \alpha_{k,\max} d_k) \geq \epsilon(u - x_k)$$

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

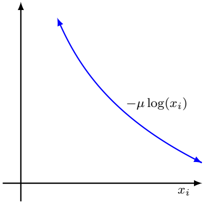
Stationarity test:

- 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:

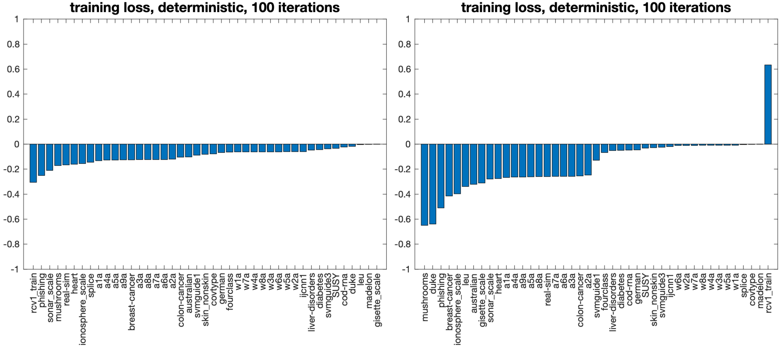
- prescribed decreasing barrier parameter sequence $\{\mu_k\} \searrow 0$ (single-loop algorithm!)
- prescribed $\{\theta_k\} \searrow 0$ and enforcing

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

"Wait! Is it worthwhile to have an algorithm like this?!"

- Our experiments say yes!

Deterministic setting



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

Proposed algorithm

Algorithm SLIP : Single-loop interior-point method

choose an initial point $x_1 \in \mathcal{N}_{[l,u]}(\theta_0)$, $\{\mu_k\} \searrow 0$, $\{\theta_k\} \searrow 0$
for all $k \in \{1, 2, \ldots\}$ **do**
  compute descent direction d_k (e.g., estimating $-\nabla\phi(x_k, \mu_k)$)
  set

$$\alpha_k \leftarrow \frac{1}{L + 2\mu_k \theta_k^{-2}}$$

  set $\gamma_k \in (0, 1]$ to ensure

$$x_{k+1} \leftarrow x_k + \gamma_k \alpha_k d_k \in \mathcal{N}_{[l,u]}(\theta_k)$$

end for

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

Key observation

Our first key observation is that the algorithm essentially acts equivalently on reducing

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

and

$$\tilde{\phi}(x,\mu) = f(x) - \mu \sum_{i=1}^{n} \log\left(\frac{x_i - l_i}{\chi}\right) - \mu \sum_{i=1}^{n} \log\left(\frac{u_i - x_i}{\chi}\right),$$

where χ is sufficiently large such that $\frac{x_i - l_i}{\chi} \in [0,1]$ and $\frac{u_i - x_i}{\chi} \in [0,1]$ for all $i \in [n]$.

The latter is simply a shifted form of the other.

- They have the same gradients! $\nabla_x \phi(x,\mu) = \nabla_x \tilde{\phi}(x,\mu)$
- For the latter, $\bar{\mu} < \mu$ implies that $\tilde{\phi}(x,\bar{\mu}) < \tilde{\phi}(x,\mu)$.

Thus, the algorithm uses ϕ, but our analysis can focus on monotonically decreasing $\{\tilde{\phi}(x_k,\mu_k)\}$.

Critical lemmas, deterministic setting

Lemma

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

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

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

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

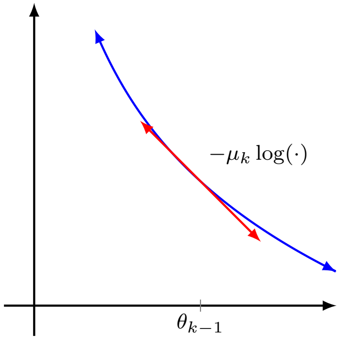
Theorem

One finds that

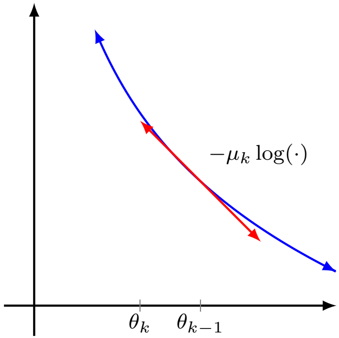
$$\liminf_{k\to\infty} \|\nabla_x \phi(x_k, \mu_k)\|_2^2 = 0,$$

and, 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).*

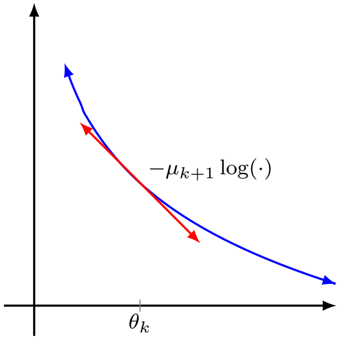
Why does it work?



Why does it work?



Why does it work?



Outline

Stochastic setting

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

- Notably, γ_k needs to be chosen based on knowledge of noise bound.
- For the deterministic setting, $\{\mu_k\} = \{\mu_1 k^t\}$ and $\{\theta_{k-1}\} = \{\theta_0 k^t\}$ for $t = -1$ implies

$$\{\alpha_k\} = \left\{ \frac{1}{L + 2\mu_k \theta_k^{-2}} \right\} = \Theta(k^t),$$

  but for stochastic setting, step-size sequence $\{\alpha_k\}$ can no longer decrease at same rate as $\{\mu_k\}$.
- It needs to decrease more slowly than $\{\mu_k\}$ (although rates can be arbitrarily close).

Accounting for the error

The issue arises from the following lemma.

Lemma

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

$$\begin{aligned}&\tilde{\phi}(X_{k+1}, \mu_{k+1}) - \tilde{\phi}(X_k, \mu_k) \\ \leq\ & -\Gamma_k A_k \|\nabla_x \tilde{\phi}(X_k, \mu_k)\|^2_{H_k^{-1}} + \Gamma_k A_k \nabla_x \tilde{\phi}(X_k, \mu_k)^T H_k^{-1} (\nabla_x \tilde{\phi}(X_k, \mu_k) - Q_k) \\ & + \tfrac{1}{2}\Gamma_k^2 A_k^2 \lambda_{k,\min}^{-1} \ell_{\nabla f, \mathcal{B}, k} \|Q_k\|^2_{H_k^{-1}}.\end{aligned}$$

Using $\{\mu_k\} = \{\mu_1 k^{-1}\}$ and $\{\theta_{k-1}\} = \{\theta_0 k^{-1}\}$, so $\{\alpha_k\} = \Theta(k^t)$, leaves the final term uncontrolled!

Parameter rule

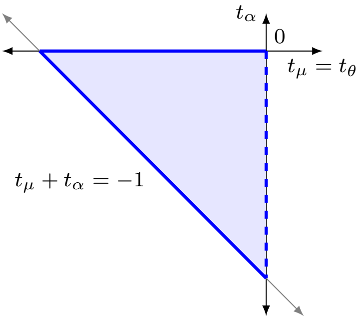
Given prescribed $(t_\mu, t_\theta, t_\alpha) \in (-\infty, -\frac{1}{2}) \times (-\infty, -\frac{1}{2}) \times (-\infty, 0)$ such that $t_\mu = t_\theta$, $t_\mu + t_\alpha \in [-1, 0)$, and $t_\mu + 2t_\alpha \in (-\infty, -1)$ along with prescribed $\alpha_{\rm buff} \in \mathbb{R}_{\geq 0}$, $\{\alpha_{k,\rm buff}\} \subset \mathbb{R}_{\geq 0}$, $\gamma_{\rm buff} \in \mathbb{R}_{\geq 0}$, and $\{\gamma_{k,\rm buff}\} \subset \mathbb{R}_{\geq 0}$ such that $\alpha_{k,\rm buff} \leq \alpha_{\rm buff} k^{2t_\mu}$ and $\gamma_{k,\rm buff} \leq \gamma_{\rm buff} k^{t_\mu}$ for all $k \in \mathbb{N}$, the algorithm employs

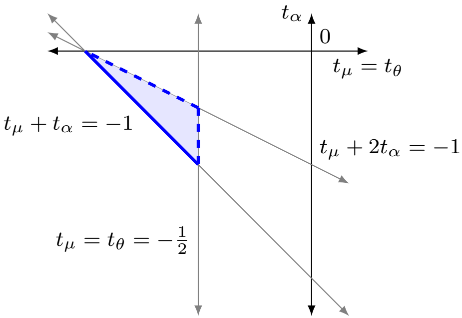
$$\alpha_{k,\min} := \frac{\lambda_{k,\min} k^{t_\alpha}}{\ell_{\nabla f,\mathcal{B}} + 2\mu_k \theta_k^{-2}}, \qquad \gamma_{k,\min} := \min\left\{1, \frac{\lambda_{k,\min}\left(\frac{\frac{1}{2}\mu_k \Delta}{\mu_k + \frac{1}{2}(\kappa_{\nabla f,\mathcal{B},\infty} + \sigma_\infty)\Delta} - \theta_k\right)}{\alpha_{k,\max}(\kappa_{\nabla f,\mathcal{B},\infty} + \sigma_\infty + \mu_k \theta_{k-1}^{-1})}\right\},$$

$$\alpha_{k,\max} := \alpha_{k,\min} + \alpha_{k,\rm buff}, \qquad \text{and} \quad \gamma_{k,\max} := \min\{1, \gamma_{k,\min} + \gamma_{k,\rm buff}\}$$

and makes a (run-and-iterate-dependent) choice $\alpha_k \in \min\left\{\frac{\lambda_{k,\min} k^{t_\alpha}}{L + 2\mu_k \theta_k^{-2}}, \alpha_{k,\max}\right\}$ for all $k \in \mathbb{N}$.

Acceptable rate values





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

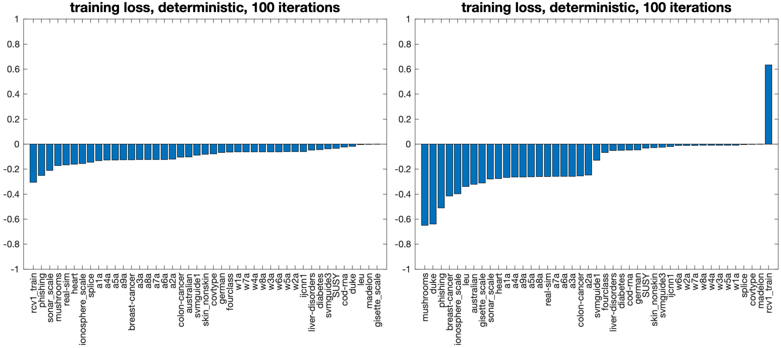
$$x_{k+1} \leftarrow \text{Proj}_{[l,u]}(x_k + \alpha_k d_k).$$

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

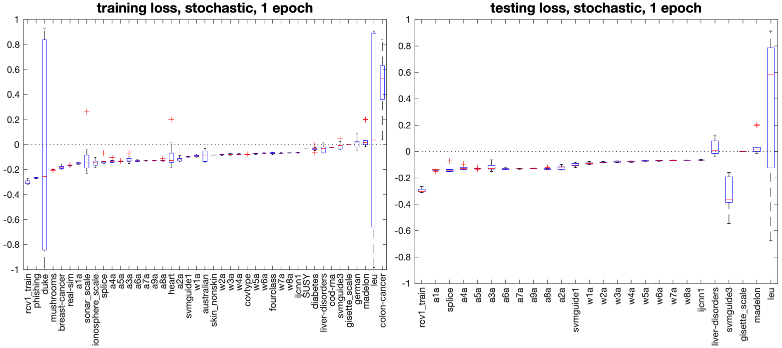
$$\frac{f(x_{\text{end}}^{\text{SLIP}}) - f(x_{\text{end}}^{\text{PSGM}})}{\max\{f(x_{\text{end}}^{\text{SLIP}}), f(x_{\text{end}}^{\text{PSGM}}), 1\}} \in (-1, 1)$$

Deterministic setting



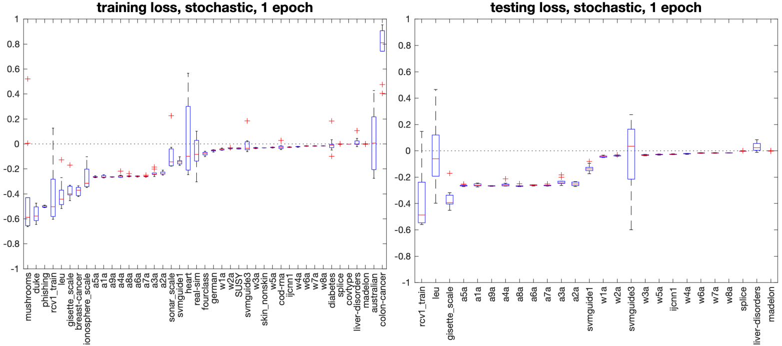
Relative performance of SLIP and PGM, deterministic setting, training logistic regression (left) and neural network models with one hidden layer with cross-entropy loss (right).

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

SLIP algorithm

Algorithm SLIP : Single-loop interior-point method

choose an initial point $x_1 \in \mathcal{N}_{[l,u]}(\theta_0)$, $\{\mu_k\} \searrow 0$, $\{\theta_k\} \searrow 0$
for all $k \in \{1, 2, \ldots\}$ **do**
  compute descent direction d_k (e.g., estimating $-\nabla\phi(x_k, \mu_k)$)
  set
$$\alpha_k \leftarrow \frac{1}{L + 2\mu_k\theta_k^{-2}}$$
  set $\gamma_k \in (0, 1]$ to ensure
$$x_{k+1} \leftarrow x_k + \gamma_k\alpha_k d_k \in \mathcal{N}_{[l,u]}(\theta_k)$$
end for

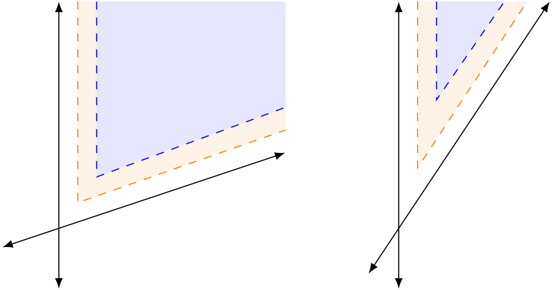
How can this be extended for the generally constrained setting?

- This is a feasible algorithm; this can be maintained, at least for inequalities only.
- Neighborhood enforcement is the real issue! Constraint value depends nonlinearly on γ_k.

Main challenge

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

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



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're doing it!
- 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."