A Stochastic-Gradient-based Interior-Point Method for Inequality-Constrained Continuous Optimization

Frank E. Curtis, Lehigh University

presented at

"Optimization and Learning: Theory and Applications" Workshop
Centre de Recherches Mathématiques (CRM), Montreal, Canada

May 28, 2024

Collaborators and references

Published:

- 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," *SIAM Journal on Optimization*, https://doi.org/10.1137/23M1569460.

Under review:

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

Outline

Outline

Interior-point methods

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

- Ipopt, Knitro, LOQO, etc.

How about noisy or stochastic settings? The current trend is to allow "constraints" only through:

- projection-based methods
- manifold-based methods
- conditional-gradient methods
- penalization (e.g., augmented Lagrangians)

Is the current trend the end of the story? A place for IPMs in noisy and stochastic optimization?

Prior recent work

With various collaborators in recent years, I have worked on stochastic Newton/SQP methods

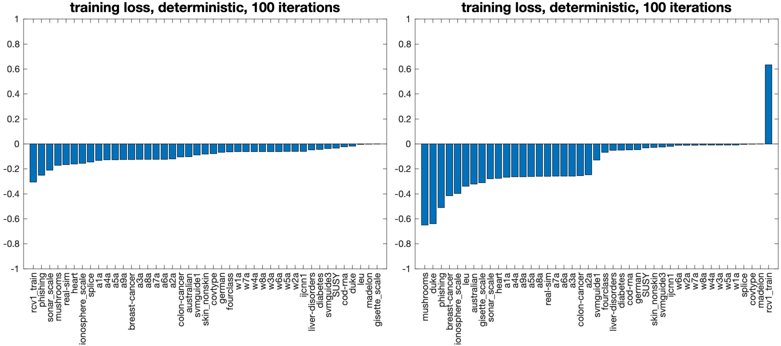
- joint work with Berahas, Jiang, O'Neill, Robinson, Wang, Zhou
- stochastic objective gradient evaluations
- deterministic equality-constraint function and derivative evaluations
- ... follow-up work by others, e.g., Na et al.
- ... some work on solving generally constrained problems (SQP, not IPM)

I have also worked on an interior-point method for constrained optimization with deterministic noise

- joint work with Dezfulian and Waechter
- noisy objective function and gradient evaluations
- noisy equality- and inequality-constraint function and derivative evaluations
- ... follows work by Nocedal et al.

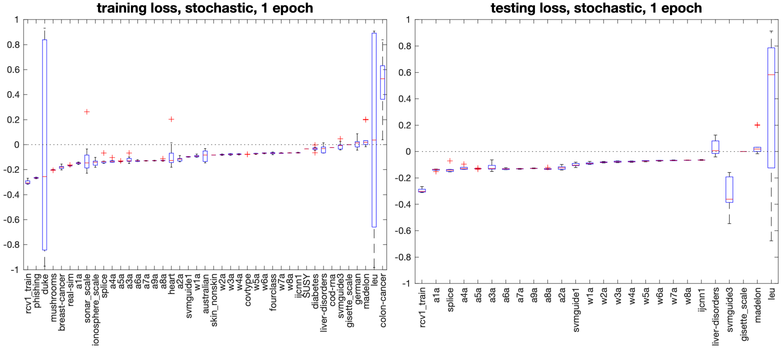
The impressive practical performance of these methods motivates stochastic-gradient-based IPMs

SLIP (this talk), deterministic setting



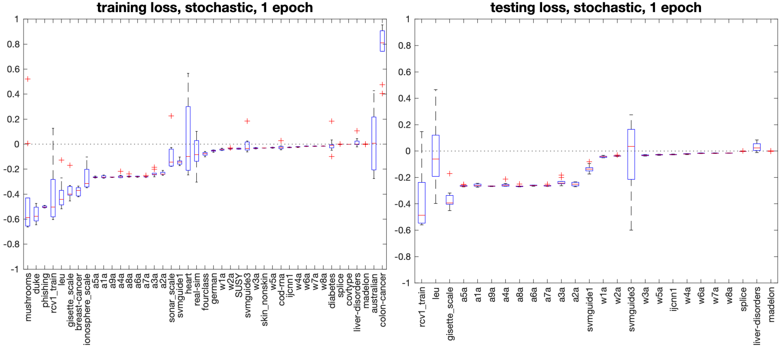
Relative performance of SLIP and PGM, deterministic setting, training logistic regression (left) and neural network models with one hidden layer with cross-entropy loss (right).

SLIP (this talk), 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.

SLIP (this talk), 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.

Challenges

... so the motivation to develop stochastic-gradient-based interior-point methods seems sufficient

What are the challenges?

- Stochastic-gradient-based algorithms require gradients to be bounded and Lipschitz continuous
- ... but barrier functions (e.g., logarithmic barrier) have neither property
- Standard interior-point methods have a two-loop structure
- ... but the stationarity test in the inner loop is problematic in stochastic settings

I will begin the talk by introducing the generally constrained case

- However, for a detailed look at the analysis, I will focus on the bound-constrained case, for simplicity
- I will end with our conclusions for the generally constrained case

Outline

Problem formulation

Let's begin by supposing that we can handle the generally constrained case:

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

To allow for an *infeasible method*, the typical approach is to introduce slack variables:

$$\boxed{\begin{aligned} \min_{(x,s)\in\mathbb{R}^n\times\mathbb{R}^{|\mathcal{I}|}} \ & f(x) \\ \text{s.t. } & c_{\mathcal{E}}(x) = 0, \\ & c_{\mathcal{I}}(x) + s = 0, \quad s \geq 0 \end{aligned}}$$

Barrier subproblem

The barrier can now be applied to the slack variables to allow infeasibility of the original constraints:

$$\boxed{\begin{aligned} \min_{(x,s)\in\mathbb{R}^n\times\mathbb{R}^{|\mathcal{I}|}} \ & f(x) - \sum_{i\in\mathcal{I}} \log(s_i) \\ \text{s.t.} \ & c_{\mathcal{E}}(x) = 0, \\ & c_{\mathcal{I}}(x) + s = 0, \quad s > 0 \end{aligned}}$$

This is essentially an equality-constrained subproblem, so use simple (stochastic) Newton/SQP, right?

- *Sorry, no!* Waechter and Biegler taught this lesson 25 years ago...

Waechter and Biegler example

Linearization of constraints + fraction-to-the-boundary rule (for slacks) $\Longrightarrow$ failure + degeneracy

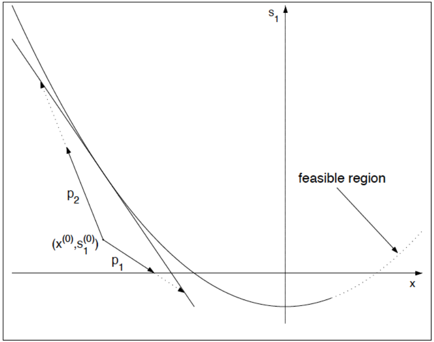


image from Nocedal & Wright (2006)

What can be done?

Some options from the literature include:

- relax constraints
- feasibility restoration
- step decomposition with scaling

Each of these approaches come with major complications for the stochastic setting. Thus, we consider:

- *feasible method*, so need to limit the scope somewhat.

Feasible interior-point method

Given $f : \mathbb{R}^n \to \mathbb{R}$, $A \in \mathbb{R}^{l\times n}$, $b \in \mathbb{R}^l$, and $c : \mathbb{R}^n \to \mathbb{R}^m$, consider

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

If x is a minimizer and a constraint qualification (e.g., the MFCQ) holds, then for some (y, z) one has

$$\nabla f(x) + A^T y + \nabla c(x) z = 0,\quad Ax = b,\quad c(x) \leq 0,\quad z \geq 0,\quad \text{and}\quad -z \circ c(x) = 0.$$

Textbook algorithm

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

$$\phi(x, \mu) = f(x) - \mu \sum_{i=1}^{m} \log(-c_i(x)) \qquad \text{and} \qquad \boxed{\begin{aligned} \min_{x \in \mathbb{R}^n} \ & \phi(x, \mu) \\ \text{s.t.} \ & Ax = b \end{aligned}}$$

Algorithm IPM : Interior-point method (textbook version)

choose x_1 with $Ax_1 = b$ and $c(x_1) < 0$; choose y_1; choose barrier parameter $\mu \in \mathbb{R}_{>0}$
for all $k \in \{1, 2, \dots\}$ **do**
    set $z_{k,i} \leftarrow -\mu(c_i(x_k))^{-1}$ for all $i \in [m]$
    if $\|(\nabla f(x_k) + A^T y_k + \nabla c(x_k) z_k, Ax_k - b, z_k \circ c(x_k) + \mu \mathbb{1})\|_2 = \mathcal{O}(\mu)$ **then** set $\mu \leftarrow 10^{-1}\mu$
    compute $d_k \in \text{Null}(A)$ that is a descent direction for $\phi(\cdot, \mu)$ at x_k
    set $\alpha_{k,\max} \in (0, 1]$ by fraction-to-the-boundary rule to ensure $c(x_k + \alpha_{k,\max} d_k) \leq 10^{-2} c(x_k)$
    set $\alpha_k \in (0, \alpha_{k,\max}]$ to ensure sufficient decrease $\phi(x_k + \alpha_k d_k, \mu) \ll \phi(x_k, \mu)$
    set $x_{k+1} \leftarrow x_k + \alpha_k d_k$
end for

Major challenges for the stochastic setting

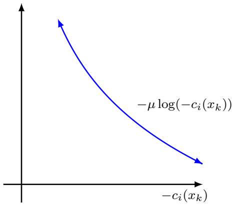
Stationarity test:

- Computing $\|\nabla f(x_k) + A^T y_k + \nabla c(x_k) z_k\|_2$ 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 through $c(x_k)$ 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 enforcement of

$$x_{k+1} \in \mathcal{N}(\theta_k) := \{x \in \mathbb{R}^n : c(x) \leq -\theta_k \mathbb{1}\}$$

Will it converge? Will it work well in practice?

- Our work says yes to both!

Outline

Proposed algorithm

$$\boxed{\min_{x\in\mathbb{R}^n} \; f(x) \quad \text{s.t.} \quad l \le x \le u} \qquad \phi(x,\mu) = f(x) - \mu \sum_{i=1}^{n} \log(x_i - l_i) + \sum_{i=1}^{n} \log(u_i - x_i)$$

Algorithm SLIP : Single-loop interior-point method

1: choose an initial point $x_1 \in \mathcal{N}_{[l,u]}(\theta_0)$, $\{\mu_k\} \searrow 0$, $\{\theta_k\} \searrow 0$

2: **for** all $k \in \{1,2,\dots\}$ **do**

3: compute descent direction d_k (e.g., estimating $-\nabla\phi(x_k,\mu_k)$)

4: set

$$\alpha_k \leftarrow \frac{1}{L + 2\mu_k\theta_k^{-2}} \quad \text{where} \quad L = \text{Lipschitz constant for } \nabla f$$

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**

Note: Our paper considers a more general framework; this is a simplified instance

Key observation

Our first key observation is that the algorithm essentially acts equivalently to minimize

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

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 former.

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

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

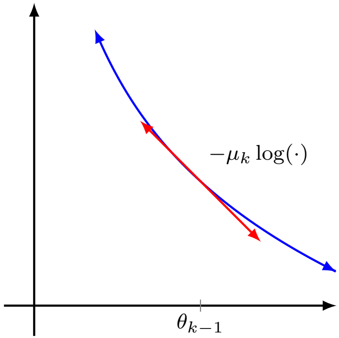
Theorem

One finds that

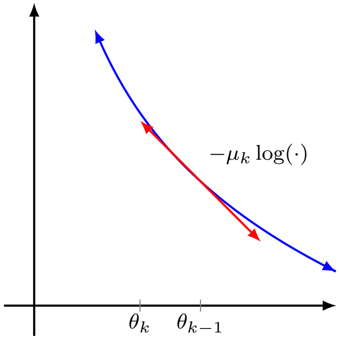
$$\liminf_{k\to\infty} \|\nabla_x \phi(x_k, \mu_k)\|_2^2 = 0,$$

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).*

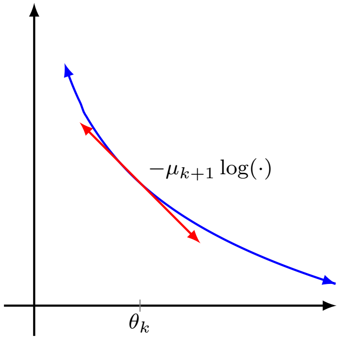
Why does it work?



Why does it work?



Why does it work?



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 $\{\alpha_k\} = \Theta(k^{-1})$, leaves the final term uncontrolled!

Parameter rule

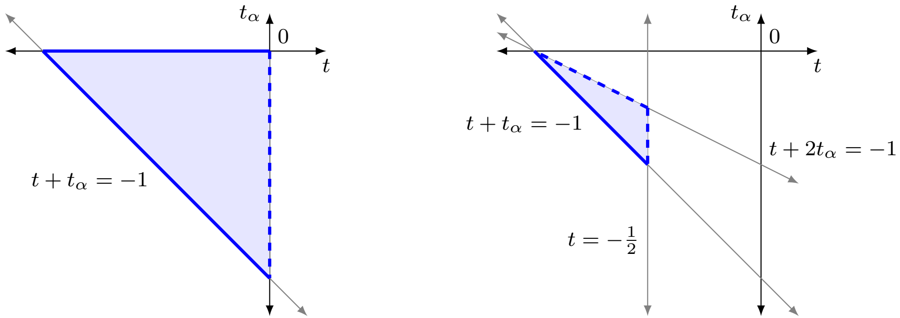
Given prescribed $(t, t_\alpha) \in (-\infty, -\frac{1}{2}) \times (-\infty, 0)$ such that $t + t_\alpha \in [-1, 0)$, and $t + 2t_\alpha \in (-\infty, -1)$ along with prescribed $\alpha_{\rm buff} \in \mathbb{R}_{\geq 0}$, $\{\alpha_{k,\rm buff}\} \subset \mathbb{R}_{\geq 0}$, $\gamma_{\rm buff} \in \mathbb{R}_{\geq 0}$, and $\{\gamma_{k,\rm buff}\} \subset \mathbb{R}_{\geq 0}$ such that $\alpha_{k,\rm buff} \leq \alpha_{\rm buff} k^{2t}$ and $\gamma_{k,\rm buff} \leq \gamma_{\rm buff} k^t$ 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

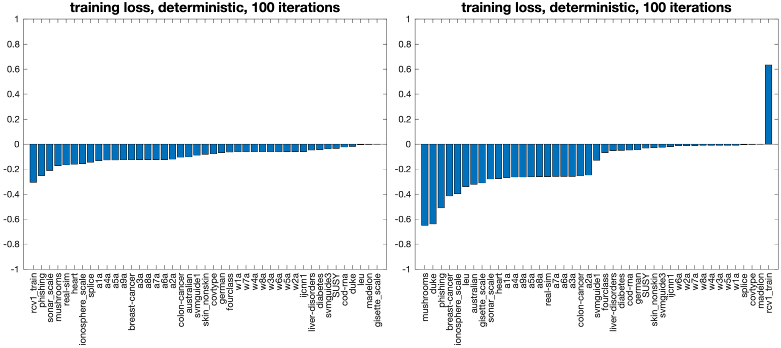
$$x_{k+1} \leftarrow \text{Proj}_{[l,u]}(x_k + \alpha_k d_k).$$

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

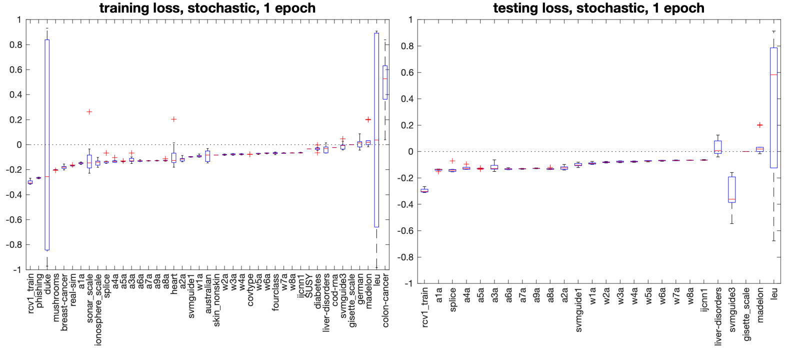
$$\frac{f(x_{\text{end}}^{\text{SLIP}}) - f(x_{\text{end}}^{\text{PSGM}})}{\max\{f(x_{\text{end}}^{\text{SLIP}}), f(x_{\text{end}}^{\text{PSGM}}), 1\}} \in (-1, 1)$$

Deterministic setting



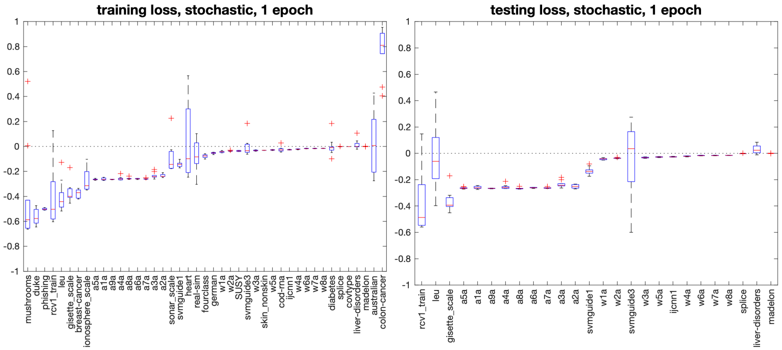
Relative performance of SLIP and PGM, deterministic setting, training logistic regression (left) and neural network models with one hidden layer with cross-entropy loss (right).

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

Search direction conditions

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

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

Need an initial point $x_1 \in \mathbb{R}^n$ satisfying

$$Ax_1 = b \ \text{ and } \ c(x_1) < 0,$$

and, with $P := I - A^T(AA^T)^{-1}A$, to ensure/assume that, for all $k \in \mathbb{N}$, one can compute d_k satisfying

$$\begin{aligned} Ad_k &= 0 \\ \underline{\zeta}\|Pq_k\|_2 \leq \|d_k\|_2 &\leq \overline{\zeta}\|Pq_k\|_2 \\ -(Pq_k)^T d_k &\geq \zeta\|Pq_k\|_2\|d_k\|_2 \\ \nabla c_i(x_k)^T d_k &\leq -\tfrac{1}{2}\overline{\eta}\|d_k\|_2 \ \text{ for all } \ i \in \{j \in [m] : -\eta\mu_k < c_i(x_k)\}. \end{aligned}$$

Convergence guarantees

Theorem

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

$$\sum_{k=1}^{\infty} \gamma_k \alpha_k = \infty \quad \textit{and} \quad \liminf_{k \to \infty} \|P \nabla_x \phi(x_k, \mu_k)\|_2^2 = 0.$$

In addition, for any infinite-cardinality set $\mathcal{K} \subseteq \mathbb{N}$ such that

- $\{P \nabla_x \phi(x_k, \mu_k)\}_{k \in \mathcal{K}} \to 0$,
- $\{x_k\}_{k \in \mathcal{K}} \to \bar{x}$, *and*
- *the LICQ holds at* $\bar{x}$,

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).

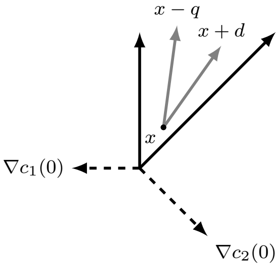
Stochastic setting: Similar parameter rule as bound-constrained case yields

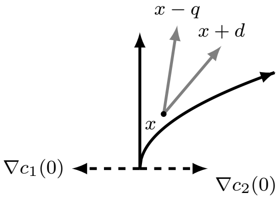
$$\liminf_{k \to \infty} \|P \nabla_x \phi(x_k, \mu_k)\|_2^2 = 0 \ \text{ almost surely}$$

Main challenge

Recall the assumption that, for all $k \in \mathbb{N}$, the algorithm computes d_k satisfying

$$\begin{aligned} Ad_k &= 0 \\ \underline{\zeta}\|Pq_k\|_2 \leq \|d_k\|_2 &\leq \overline{\zeta}\|Pq_k\|_2 \\ -(Pq_k)^T d_k &\geq \zeta\|Pq_k\|_2\|d_k\|_2 \\ \nabla c_i(x_k)^T d_k &\leq -\tfrac{1}{2}\overline{\eta}\|d_k\|_2 \ \text{ for all } \ i \in \{j \in [m] : -\eta\mu_k < c_i(x_k)\}. \end{aligned}$$





Outline

Summary

Presented single-loop interior-point methods for solving inequality-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!

Collaborators and references




Published:

- 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," *SIAM Journal on Optimization*, https://doi.org/10.1137/23M1569460.

Under review:

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