Stochastic Algorithms for Solving Constrained Optimization Problems

Frank E. Curtis, Lehigh University

involving joint work with

Albert Berahas (University of Michigan), **Michael O'Neill** (UNC Chapel Hill), **Suyun Liu** (Amazon), **Daniel P. Robinson** (Lehigh University), **Qi Wang** (Lehigh University), **Baoyu Zhou** (Chicago Booth)

presented at

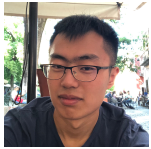
Department of Industrial and Systems Engineering, North Carolina State University

March 6, 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," https://arxiv.org/abs/2112.14799.
- F. E. Curtis, S. Liu, and D. P. Robinson, "Fair Machine Learning through Constrained Stochastic Optimization and an ϵ-Constraint Method."
- F. E. Curtis, D. P. Robinson, and B. Zhou, "Stochastic SQP for Inequality-Constrained Optimization."
- F. E. Curtis, D. P. Robinson, and Q. Wang, "Stochastic Interior-Point Methods."

Outline

Outline

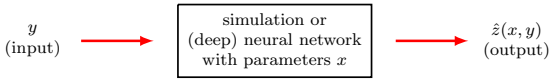
Stochastic optimization

Consider

$$\min_{x \in \mathbb{R}^n} f(x), \quad \text{where} \quad f(x) = \mathbb{E}_\omega[F(x, \omega)],$$

where ω has probability space $(\Omega, \mathcal{F}, P)$, $F : \mathbb{R}^n \times \Omega \to \mathbb{R}$, and $f : \mathbb{R}^n \to \mathbb{R}$

- Simulation-based optimization, machine learning, etc.



- Long history of algorithms (stochastic approximation, stochastic average approximation, etc.)
- Explosion in interest and number of new algorithms
- Various convergence guarantees about algorithms and statistical guarantees about solutions
- Algorithms generally "easy" to implement (but tuning is expensive)

Constrained optimization (deterministic)

Consider

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

where $f : \mathbb{R}^n \to \mathbb{R}$, $c_{\mathcal{E}} : \mathbb{R}^n \to \mathbb{R}^{m_{\mathcal{E}}}$, and $c_{\mathcal{I}} : \mathbb{R}^n \to \mathbb{R}^{m_{\mathcal{I}}}$ are smooth

- Physics-constrained, resource-constrained, etc.
- Long history of algorithms (penalty, SQP, interior-point, etc.)
- Comprehensive theory (even with lack of constraint qualifications)
- Effective software (Ipopt, Knitro, LOQO, etc.)

Constrained optimization (stochastic constraints)

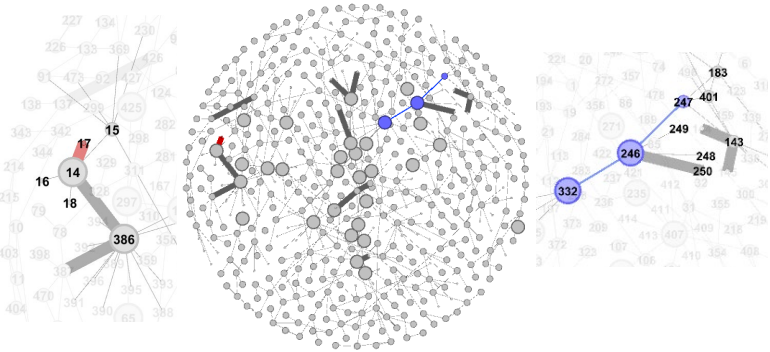
Consider

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

where $f : \mathbb{R}^n \to \mathbb{R}$, $c_{\mathcal{E}} : \mathbb{R}^n \to \mathbb{R}^{m_{\mathcal{E}}}$, and $c_{\mathcal{I}} : \mathbb{R}^n \times \Omega \to \mathbb{R}^{m_{\mathcal{I}}}$

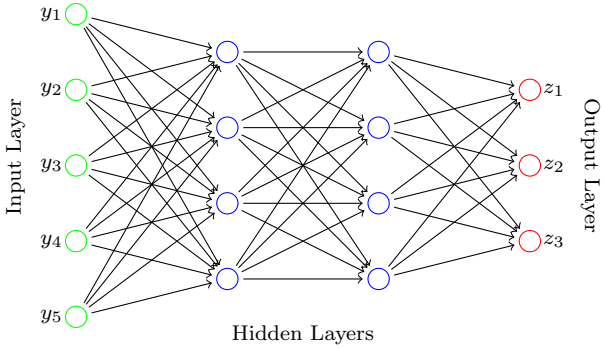
- Various modeling paradigms:
- . . . stochastic optimization (i.e., constrain an expectation)
- . . . (distributionally) robust optimization
- . . . chance-constrained optimization
- Algorithms (e.g., based on integer programming techniques) can be expensive

Motivation #1: Network optimization



Motivation #2: Physics-informed learning (e.g., PINNs)

Deep learning models have proved to be powerful tools in certain contexts:



Motivation #2: Physics-informed learning (e.g., PINNs)

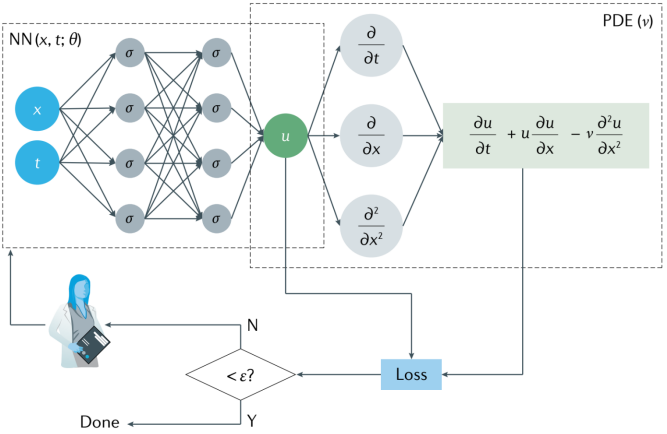


Photo: Karniadakis et al.

Motivation #3: Fair learning

Let

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

Given a model function ϕ and loss ℓ, an optimization problem arising in machine learning has the form:

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

However, the resulting "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 (a constraint!)

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

Regularized optimization

The typical approach for "informed optimization" is regularization (to avoid constraints)

$$\min_{x \in \mathbb{R}^n} \; f(x) + r(x), \quad \text{where} \quad f(x) = \mathbb{E}_\omega[F(x, \omega)],$$

where $r : \mathbb{R}^n \times \mathbb{R}$ is often convex and potentially nonsmooth, but this can be computationally expensive (due to need to tune hyperparameters) and does not guarantee *exact* satisfaction

An idealized approach might be to consider a problem formulation such as

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \; & f(x), \quad \text{where} \quad f(x) = \mathbb{E}_\omega[F(x, \omega)] \\ \text{s.t. } & c_{\mathcal{E}}(x) = 0 \\ & c_{\mathcal{I}}(x, \omega) \lesssim 0 \end{aligned}$$

but this leads to serious computational tractability issues! (At least for now....)

Constrained optimization (stochastic objective)

Our approach (as a stepping stone to tackling more difficult settings) is to consider

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \; & f(x), \quad \text{where} \quad f(x) = \mathbb{E}_\omega[F(x, \omega)] \\ \text{s.t.} \; & c_{\mathcal{E}}(x) = 0 \\ & c_{\mathcal{I}}(x) \leq 0 \end{aligned}$$

- Classical applications under uncertainty, constrained DNN training, etc.
- Besides cases involving a deterministic equivalent...
- ... very few algorithms so far (mostly penalty methods)

Outline

Equality-constrained setting (to start)

Consider the equality-constrained optimization problem:

$$\begin{aligned} &\min_{x \in \mathbb{R}^n} \ f(x), \quad \text{where} \quad f(x) = \mathbb{E}_\omega[F(x, \omega)] \\ &\text{s.t. } c(x) = 0 \end{aligned}$$

- Adaptive SQP method for deterministic setting
- Stochastic SQP method for stochastic setting
- Convergence guarantees
- Worst-case complexity guarantees
- Promising numerical experiments
- Various extensions

What kind of algorithm do we want?

Need to establish what we want/expect from an algorithm.

Note: We are interested in the fully stochastic regime.[†]

We assume:

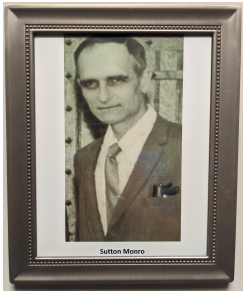
- Feasible methods are not tractable
- ... so no projection methods, Frank-Wolfe, etc.
- "Two-phase" methods are not effective
- ... so should not search for feasibility, then optimize.

Finally, want to use techniques that can generalize to diverse settings.

[†] Alternatively, see Na, Anitescu, Kolar (2021, 2022) and others

Stochastic gradient method (SG)

Stochastic approximation by Herbert Robbins and Sutton Monro (1951)



Sutton Monro, former Lehigh faculty member

Stochastic gradient (*not* descent)

Suppose $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$ is Lipschitz continuous with constant L

Algorithm SG : Stochastic Gradient

1: choose an initial point $x_1 \in \mathbb{R}^n$ and step sizes $\{\alpha_k\} > 0$
2: **for** $k \in \{1, 2, \dots\}$ **do**
3: set $x_{k+1} \leftarrow x_k - \alpha_k g_k$, where $\mathbb{E}[G_k|\mathcal{F}_k] = \nabla f(X_k)$ and $\mathbb{E}[\|G_k - \nabla f(X_k)\|_2^2|\mathcal{F}_k] \leq M$
4: **end for**

Notation: $\{(x_k, g_k)\}$ is a realization of the stochastic process $\{(X_k, G_k)\}$ with filtration $\{\mathcal{F}_k\}$

Not a descent method! ... but *eventual descent in expectation*:

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

Markovian: In any run, x_{k+1} depends only on x_k and random choice at iteration k.

SG theory

Theorem SG

Since $\mathbb{E}[G_k|\mathcal{F}_k] = \nabla f(X_k)$ *and* $\mathbb{E}[\|G_k - \nabla f(X_k)\|_2^2|\mathcal{F}_k] \leq M$ *for all* $k \in \mathbb{N}$*:*

$$\alpha_k = \frac{1}{L} \quad \Longrightarrow \mathbb{E}\left[\frac{1}{k}\sum_{j=1}^{k}\|\nabla f(X_j)\|_2^2\right] = \mathcal{O}(M)$$

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

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

SG illustration

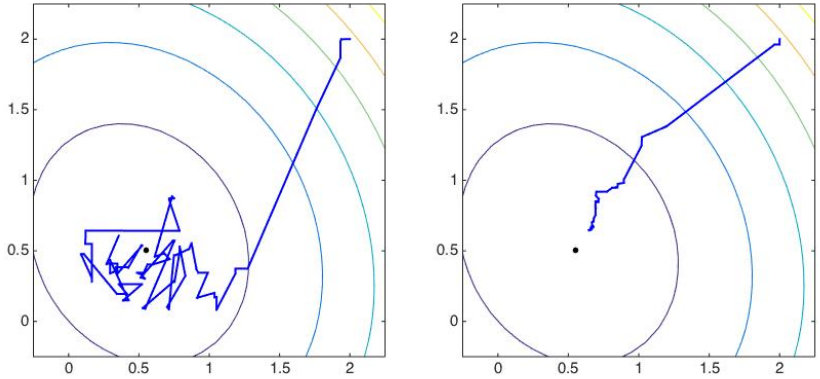


Figure: SG with fixed step size (left) vs. diminishing step sizes (right)

Sequential quadratic optimization (SQP)

Consider

$$\boxed{\begin{array}{l}\min\limits_{x\in\mathbb{R}^n} f(x)\\ \quad \text{s.t. } c(x)=0\end{array}}$$

with $J \equiv \nabla c$ and H positive definite over $\text{Null}(J)$, two viewpoints:

$$\boxed{\begin{bmatrix}\nabla f(x) + J(x)^T y\\ c(x)\end{bmatrix} = 0} \quad \text{or} \quad \boxed{\begin{array}{l}\min\limits_{d\in\mathbb{R}^n} f(x) + \nabla f(x)^T d + \frac{1}{2} d^T H d\\ \quad \text{s.t. } c(x) + J(x)d = 0\end{array}}$$

both leading to the same "Newton-SQP system":

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

SQP illustration

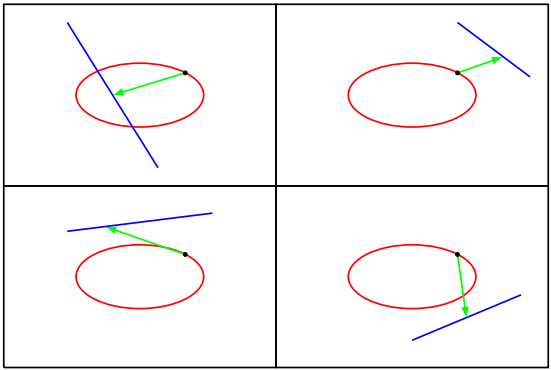


Figure: Illustrations of SQP subproblem solutions

SQP with backtracking line search

Algorithm guided by merit function with adaptive parameter τ defined by

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

Algorithm SQP w/ line search

1: choose $x_1 \in \mathbb{R}^n$, $\tau_0 \in \mathbb{R}_{>0}$, $\eta \in (0, 1)$

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

3: compute step: solve

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

4: update merit parameter: set τ_k to ensure

$$\phi'(x_k, \tau_k, d_k) \leq -\Delta q(x_k, \tau_k, \nabla f(x_k), d_k) \ll 0$$

5: compute step size: backtracking line search to ensure $x_{k+1} \leftarrow x_k + \alpha_k d_k$ yields

$$\phi(x_{k+1}, \tau_k) \leq \phi(x_k, \tau_k) - \eta \alpha_k \Delta q(x_k, \tau_k, \nabla f(x_k), d_k)$$

6: **end for**

Convergence theory

Assumption

- f, c, ∇f, *and* J *bounded and Lipschitz*
- *singular values of* J *bounded below* (*i.e., the LICQ*)
- $u^T H_k u \geq \zeta \|u\|_2^2$ *for all* $u \in \text{Null}(J_k)$ *for all* $k \in \mathbb{N}$

Theorem

- $\{\alpha_k\} \geq \alpha_{\min}$ *for some* $\alpha_{\min} > 0$
- $\{\tau_k\} \geq \tau_{\min}$ *for some* $\tau_{\min} > 0$
- $\Delta q(x_k, \tau_k, \nabla f(x_k), d_k) \to 0$ *implies optimality error vanishes, specifically,*

$$\|d_k\|_2 \to 0, \quad \|c_k\|_2 \to 0, \quad \|\nabla f(x_k) + J_k^T y_k\|_2 \to 0$$

Toward stochastic SQP

- In a stochastic setting, line searches are (likely) intractable
- However, for ∇f and ∇c, may have Lipschitz constants L and Γ
- Step #1: Design an adaptive SQP method with

$$\underline{\textit{step sizes determined by Lipschitz constants}}$$

- Step #2: Design a stochastic SQP method based on this approach

SQP with adaptive step sizes

Algorithm SQP w/o line search

1: choose $x_1 \in \mathbb{R}^n$, $\tau_0 \in \mathbb{R}_{>0}$, $\eta \in (0,1)$
2: **for** $k \in \{1, 2, \dots\}$ **do**
3: compute step: solve

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

4: update merit parameter: set τ_k to ensure

$$\phi'(x_k, \tau_k, d_k) \leq -\Delta q(x_k, \tau_k, \nabla f(x_k), d_k) \ll 0$$

5: compute step size: set

$$\widehat{\alpha}_k \leftarrow \frac{2(1-\eta)\Delta q(x_k, \tau_k, \nabla f(x_k), d_k)}{(\tau_k L + \Gamma)\|d_k\|_2^2} \quad \text{and} \quad \widetilde{\alpha}_k \leftarrow \widehat{\alpha}_k - \frac{4\|c_k\|_1}{(\tau_k L + \Gamma)\|d_k\|_2^2}$$

6: then

$$\alpha_k \leftarrow \begin{cases} \widehat{\alpha}_k & \text{if } \widehat{\alpha}_k < 1 \\ 1 & \text{if } \widetilde{\alpha}_k \leq 1 \leq \widehat{\alpha}_k \\ \widetilde{\alpha}_k & \text{if } \widetilde{\alpha}_k > 1 \end{cases}$$

7: then set $x_{k+1} \leftarrow x_k + \alpha_k d_k$
8: **end for**

Convergence theory: Nearly identical as for SQP w/ line search.

Stochastic SQP with adaptive step sizes

Algorithm : Stochastic SQP

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

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

4: update merit parameter: set τ_k to ensure

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

5: compute step size: set

$$\widehat{\alpha}_k \leftarrow \frac{\beta_k \Delta q(x_k, \tau_k, g_k, d_k)}{(\tau_k L + \Gamma)\|d_k\|_2^2} \quad \text{and} \quad \widetilde{\alpha}_k \leftarrow \widehat{\alpha}_k - \frac{4\|c_k\|_1}{(\tau_k L + \Gamma)\|d_k\|_2^2}$$

6: then

$$\alpha_k \leftarrow \begin{cases} \widehat{\alpha}_k & \text{if } \widehat{\alpha}_k < 1 \\ 1 & \text{if } \widetilde{\alpha}_k \leq 1 \leq \widehat{\alpha}_k \\ \widetilde{\alpha}_k & \text{if } \widetilde{\alpha}_k > 1 \end{cases}$$

7: then $x_{k+1} \leftarrow x_k + \alpha_k d_k$
8: **end for**

Assume $\{g_k\}$ is a realization of $\{G_k\}$ with $\mathbb{E}[G_k|\mathcal{F}_k] = \nabla f(X_k)$ and $\mathbb{E}[\|G_k - \nabla f(X_k)\|_2^2|\mathcal{F}_k] \leq M$

Fundamental lemma

Recall in the unconstrained setting that

$$\mathbb{E}_\omega[f(X_{k+1})|\mathcal{F}_k] - f(X_k) \leq -\alpha_k \|\nabla f(X_k)\|_2^2 + \tfrac{1}{2}\alpha_k^2 L \mathbb{E}_\omega[\|G_k\|_2^2|\mathcal{F}_k]$$

Lemma

For all $k \in \mathbb{N}$ one finds (before taking expectations)

$$\begin{aligned}
&\phi(X_{k+1}, \mathcal{T}_{k+1}) - \phi(X_k, \mathcal{T}_k) \\
\leq\ & \underbrace{-\mathcal{A}_k \Delta q(X_k, \mathcal{T}_k, \nabla f(X_k), D_k^{\text{true}})}_{\mathcal{O}(\beta_k),\ \text{“deterministic”}} + \underbrace{\tfrac{1}{2}\mathcal{A}_k \beta_k \Delta q(X_k, \mathcal{T}_k, G_k, D_k)}_{\mathcal{O}(\beta_k^2),\ \text{stochastic/noise}} + \underbrace{\mathcal{A}_k \mathcal{T}_k \nabla f(X_k)^T (D_k - D_k^{\text{true}})}_{\text{due to adaptive } \mathcal{A}_k}
\end{aligned}$$

Good merit parameter behavior

Lemma

Let $\mathcal{E} :=$ event that $\{\mathcal{T}_k\}$ eventually remains constant at $\mathcal{T}' \geq \tau_{\min} > 0$. Then, for large k,

$$\mathbb{E}_\omega[\mathcal{A}_k \mathcal{T}_k \nabla f(X_k)^T (D_k - D_k^{\text{true}}) | \mathcal{F}_k \cap \mathcal{E}] = \beta_k^2 \mathcal{T}' \mathcal{O}(\sqrt{M})$$

Theorem

Conditioned on $\mathcal{E}$, for large k, one finds

$$\beta_k = \Theta(1) \implies \mathbb{E}\left[\frac{1}{k}\sum_{j=1}^{k} \Delta q(X_j, \mathcal{T}', \nabla f(X_j), D_j^{\text{true}})\right] = \mathcal{O}(M)$$

$$\beta_k = \Theta\left(\frac{1}{k}\right) \implies \mathbb{E}\left[\frac{1}{\left(\sum_{j=1}^{k} \beta_j\right)}\sum_{j=1}^{k} \beta_j \Delta q(X_j, \mathcal{T}', \nabla f(X_j), D_j^{\text{true}})\right] \to 0$$

Good merit parameter behavior

Lemma

Let $\mathcal{E} :=$ event that $\{\mathcal{T}_k\}$ eventually remains constant at $\mathcal{T}' \geq \tau_{\min} > 0$. Then, for large k,

$$\mathbb{E}_\omega[\mathcal{A}_k \mathcal{T}_k \nabla f(X_k)^T (D_k - D_k^{\text{true}}) | \mathcal{F}_k \cap \mathcal{E}] = \beta_k^2 \mathcal{T}' \mathcal{O}(\sqrt{M})$$

Theorem

Conditioned on $\mathcal{E}$, for large k, one finds

$$\beta_k = \Theta(1) \implies \mathbb{E}\left[\frac{1}{k}\sum_{j=1}^{k} (\|\nabla f(X_j) + \nabla c(X_j)^T Y_j^{\text{true}}\|_2 + \|c(X_j)\|_2)\right] = \mathcal{O}(M)$$

$$\beta_k = \Theta\left(\frac{1}{k}\right) \implies \mathbb{E}\left[\frac{1}{\left(\sum_{j=1}^{k} \beta_j\right)}\sum_{j=1}^{k} \beta_j (\|\nabla f(X_j) + \nabla c(X_j)^T Y_j^{\text{true}}\|_2 + \|c(X_j)\|_2)\right] \to 0$$

Poor merit parameter behavior

$\{\mathcal{T}_k\} \searrow 0$:

- cannot occur if $\|G_k - \nabla f(X_k)\|_2$ is bounded uniformly
- occurs with small probability if distribution of G_k has "small tails"

$\{\mathcal{T}_k\}$ remains too large:

- under a modest assumption, occurs with probability zero

Numerical results: (Matlab) https://github.com/frankecurtis/StochasticSQP

CUTE problems with noise added to gradients with different noise levels

- Stochastic SQP: 10^3 iterations
- Stochastic Subgradient: 10^4 iterations and tuned over 11 values of penalty parameter

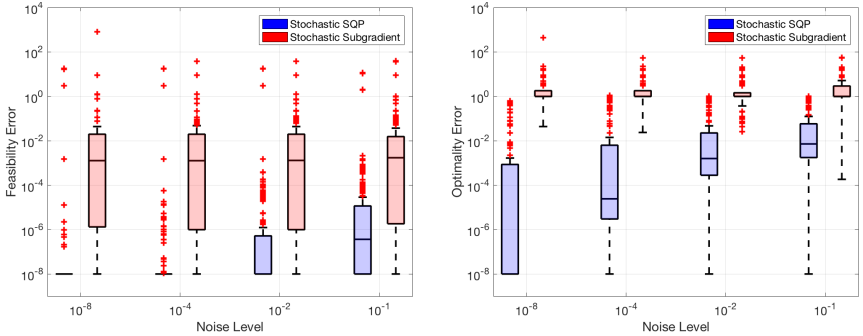


Figure: Box plots for feasibility errors (left) and optimality errors (right).

Complexity of $\mathcal{O}(\epsilon^{-2})$ for deterministic algorithm

All reductions in the merit function can be cast in terms of smallest τ.

Lemma

If $\{\tau_k\}$ eventually remains fixed at sufficiently small $\tau_{\min}$, then for any $\epsilon \in (0,1)$ there exists $(\kappa_1, \kappa_2) \in (0,\infty) \times (0,\infty)$ such that, for all k,

$$\|\nabla f(x_k) + J_k^T y_k\| > \epsilon \text{ or } \sqrt{\|c_k\|_1} > \epsilon \implies \Delta q(x_k, \tau_k, d_k) \geq \min\{\kappa_1, \kappa_2 \tau_{\min}\}\epsilon.$$

Since $\tau_{\min}$ is determined by the initial point, *it will be reached.*

Theorem

For any $\epsilon \in (0,1)$, there exists $(\kappa_1, \kappa_2) \in (0,\infty) \times (0,\infty)$ such that

$$\|\nabla f(x_k) + J_k^T y_k\| \leq \epsilon \text{ and } \sqrt{\|c_k\|_1} \leq \epsilon$$

in a number of iterations no more than

$$\left(\frac{\tau_0(f_1 - f_{\text{inf}}) + \|c_1\|_1}{\min\{\kappa_1, \kappa_2 \tau_{\min}\}} \right) \epsilon^{-2}.$$

Complexity of $\widetilde{\mathcal{O}}(\epsilon^{-4})$ for stochastic algorithm

Theorem

Suppose the algorithm is run $k_{\max}$ iterations with

- $\beta_k = \gamma/\sqrt{k_{\max}+1}$ *and*
- *the merit parameter is reduced at most $s_{\max} \in \{0, 1, \ldots, k_{\max}\}$ times.*

Let k_* be sampled uniformly over $\{1, \ldots, k_{\max}\}$. Then, with probability $1-\delta$,

$$\mathbb{E}[\|\nabla f(x_{k_*}) + J_{k_*}^T y_{k_*}\|_2^2 + \|c_{k_*}\|_1] \leq \frac{\tau_0(f_1 - f_{\inf}) + \|c_1\|_1 + M}{\sqrt{k_{\max}+1}} + \frac{(\tau_{-1} - \tau_{\min})(s_{\max}\log(k_{\max}) + \log(1/\delta))}{\sqrt{k_{\max}+1}}$$

Theorem

If the stochastic gradient estimates are sub-Gaussian, then w.p. $1-\bar{\delta}$

$$s_{\max} = \mathcal{O}\left(\log\left(\log\left(\frac{k_{\max}}{\bar{\delta}}\right)\right)\right).$$

Recent work (under review): No LICQ

Remove constraint qualification

- infeasible and/or degenerate problems
- step decomposition method

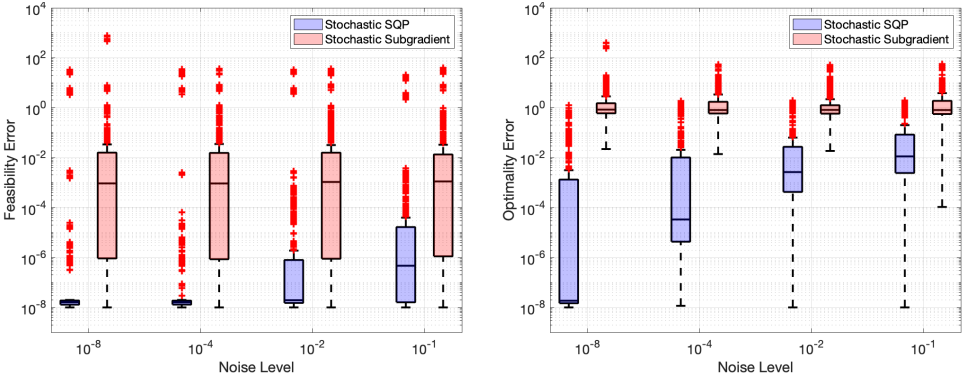


Figure: Box plots for feasibility errors (left) and optimality errors (right).

Recent work (under review): Matrix-free methods

Inexact subproblem solves

- stochasticity and inexactness(!)
- applicable for large-scale, e.g., PDE-constrained

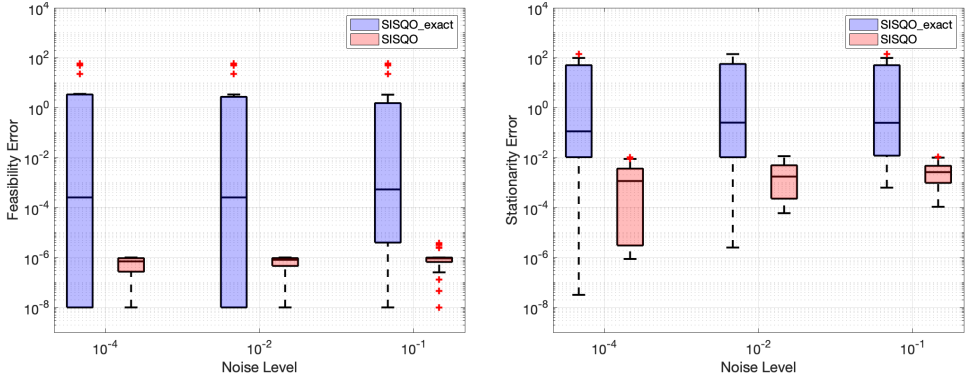
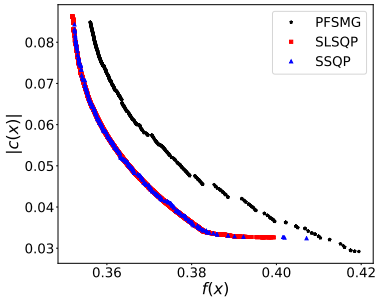


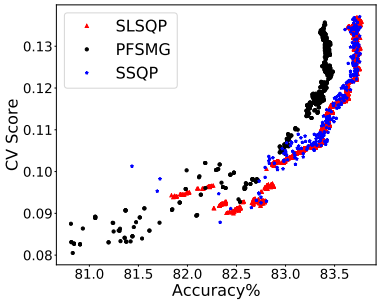
Figure: Box plots for feasibility errors (left) and optimality errors (right).

Recent work (under review): Inequality-constrained optimization

Theory and application papers:

- stochastic SQP for inequality constrained problems
- employed in an ϵ-constraint method for fair machine learning





$$\min_{x\in\mathbb{R}^n} \frac{1}{N^o} \sum_{(v_i,y_i)\in D_o} \ell(x, v_i, y_i) \quad \text{s.t.} \quad -\epsilon \leq \frac{1}{N^c} \sum_{(v_i,a_i)\in D_c} (a_i - \overline{a}) x^T v_i \leq \epsilon$$

Outline

Motivation

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

- Ipopt, Knitro, LOQO, etc.

As far as we are aware, there exist 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 simple constraints (and use projections).
- It is difficult! Stochastic algorithms require gradients to be bounded and Lipschitz continuous
- ... but the typical (e.g., logarithmic) barrier function has neither property.

Bound-constrained setting

Consider

$$\boxed{\begin{array}{l}\min\limits_{x\in\mathbb{R}^n} \; f(x) \\ \quad \text{s.t. } l \le x \le u\end{array}}$$

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

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

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

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

Deterministic setting

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

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

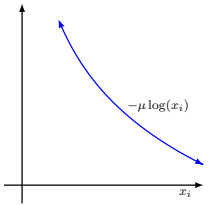
Stationarity test:

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

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 $\{\mu_k\} \searrow 0$
- 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?

Proposed algorithm

Algorithm SIPM : Stochastic interior-point method

1: choose an initial point $x_1 \in \mathbb{R}^n$, $\{\mu_k\} = \{\mu_1 k^{-1}\}$, $\{\theta_k\} = \{\theta_1 (k+1)^{-1}\}$
2: **for** $k \in \{1, 2, \dots\}$ **do**
3: compute descent direction d_k (e.g., $-\nabla\phi(x_k, \mu_k)$)
4: set $\alpha_k \in (0, 1]$ with (see paper)

$$\alpha_k = \mathcal{O}\left(\frac{1}{\ell_{\nabla f} + 2\mu_k \theta_k^{-2}}\right)$$

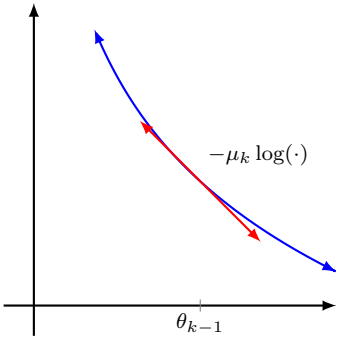
5: set $\gamma_k \in (0, 1]$ to ensure

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

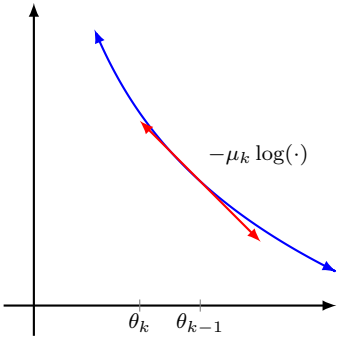
6: **end for**

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

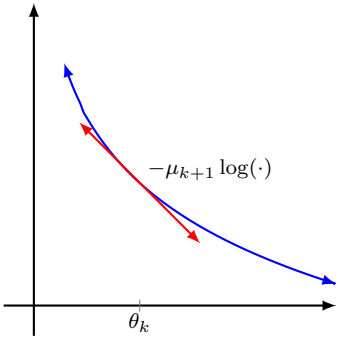
Why does it work?



Why does it work?



Why does it work?



Convergence guarantee

Theorem

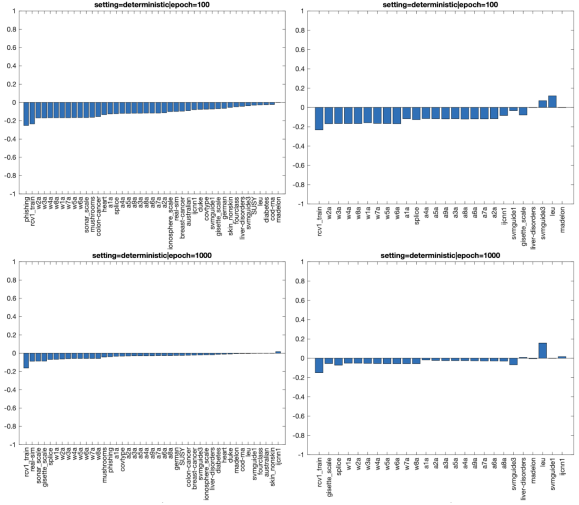
Suppose that f is bounded below, ∇f is Lipschtiz continuous, $\mu_k = \mu_1 k^{-1}$ for some sufficiently large $\mu_1 \in \mathbb{R}_{>0}$ for all $k \in \mathbb{N}$, $\theta_k = \theta_1 (k+1)^{-1}$ for some sufficiently small $\theta_1 \in \mathbb{R}_{>0}$ for all $k \in \mathbb{N}$,

$$\mathbb{E}[G_k | \mathcal{F}_k] = \nabla f(X_k), \quad \textit{and} \quad \|G_k - \nabla f(X_k)\|_2 \leq \sigma \in \mathbb{R}_{>0}.$$

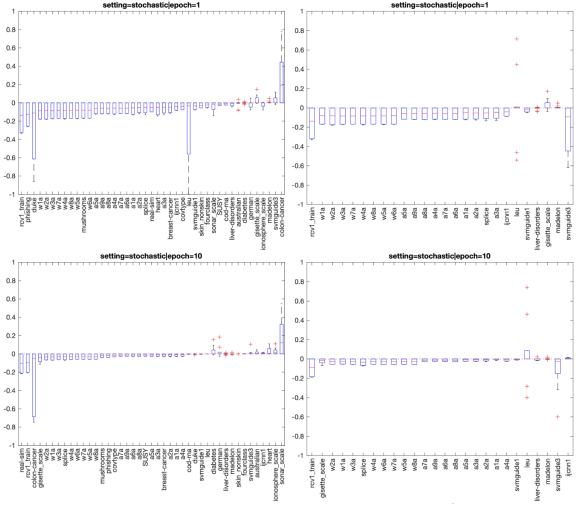
Then,

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

Preliminary results: Deterministic



Preliminary results: Stochastic



Outline

Summary

Consider stochastic-gradient-based algorithms for solving problems of the form:

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \; & f(x), \quad \text{where} \quad f(x) = \mathbb{E}_\omega[F(x, \omega)] \\ \text{s.t.} \; & c_{\mathcal{E}}(x) = 0 \\ & c_{\mathcal{I}}(x) \leq 0 \end{aligned}$$

Stochastic SQP methods

- equality-constraints only and inequality-constrained settings
- various extensions for solving large-scale problems
- convergence and complexity guarantees
- very promising experimental results

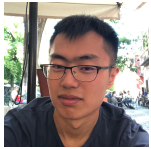
Stochastic interior-point methods

- new frontier with promising results so far

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," https://arxiv.org/abs/2112.14799.
- F. E. Curtis, S. Liu, and D. P. Robinson, "Fair Machine Learning through Constrained Stochastic Optimization and an ϵ-Constraint Method."
- F. E. Curtis, D. P. Robinson, and B. Zhou, "Stochastic SQP for Inequality-Constrained Optimization."
- F. E. Curtis, D. P. Robinson, and Q. Wang, "Stochastic Interior-Point Methods."