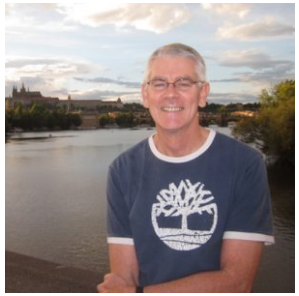
Algorithms for Deterministically Constrained Stochastic Optimization

Albert S. Berahas

Department of Industrial & Operations Engineering
University of Michigan

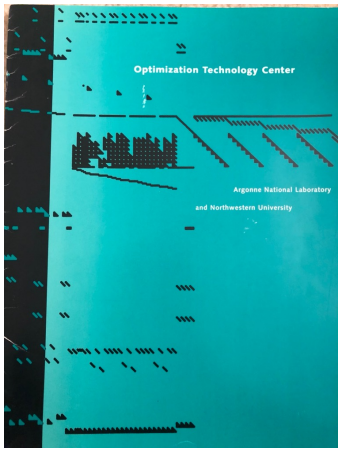
12th US-Mexico Workshop on Optimization and Its Applications
Oaxaca, Mexico. 2023

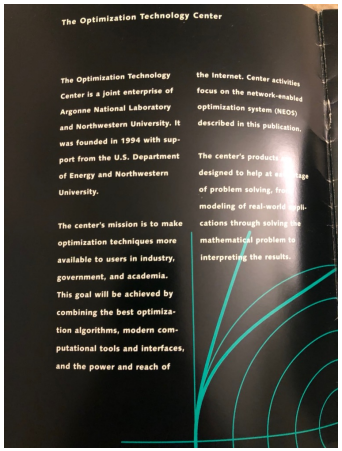
Sometimes, there's a man, well, he's the man for his time and place. He fits right in there. And that's Steve Wright, in Optimization.



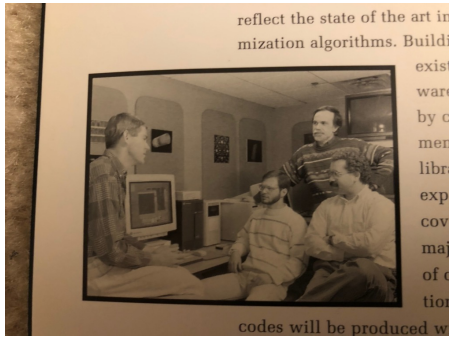
Steve 1

Steve 1


Optimization Technology Center
Argonne National Laboratory
and Northwestern University

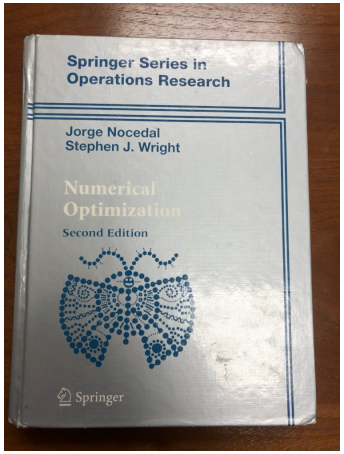

The Optimization Technology Center
The Optimization Technology Center is a joint enterprise of Argonne National Laboratory and Northwestern University. It was founded in 1994 with support from the U.S. Department of Energy and Northwestern University.
The center's mission is to make optimization techniques more available to users in industry, government, and academia. This goal will be achieved by combining the best optimization algorithms, modern computational tools and interfaces, and the power and reach of the Internet. Center activities focus on the network-enabled optimization system (NEOS) described in this publication.
The center's products
designed to help at
of problem solving, fro
modeling of real-world
cations through solving the
mathematical problem to
interpreting the results.


OTC Principals
Stephen Wright
OTC director and computer scientist, mathematics and computer science division, Argonne
Interior-point methods, nonlinear optimization
Collette Coullard
Associate professor, industrial engineering and management sciences, Northwestern
Combinatorial optimization, network optimization
Robert Fourer
Professor, industrial engineering and management sciences, Northwestern
Modeling languages and optimization environments, linear programming
Sanjay Mehrotra
Associate professor, industrial engineering and management sciences, Northwestern
Interior-point methods, optimization software
Jorge Moré
Senior computer scientist, mathematics and computer science division, Argonne
Nonlinear optimization, optimization software
Jorge Nocedal
OTC deputy director and professor, electrical engineering and computer science, Northwestern
Large-scale nonlinear optimization
David Simchi-Levi
Associate professor, industrial engineering and management sciences, Northwestern
Logistics, network design

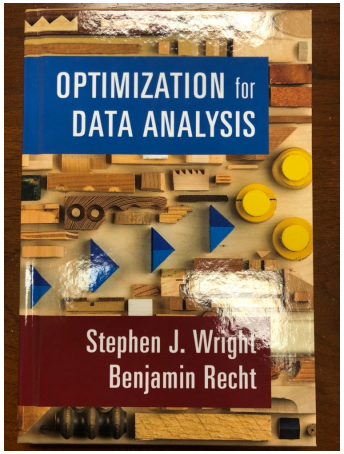

reflect the state of the art in
mization algorithms. Buildi
codes will be produced w

Steve 2

Steve 2



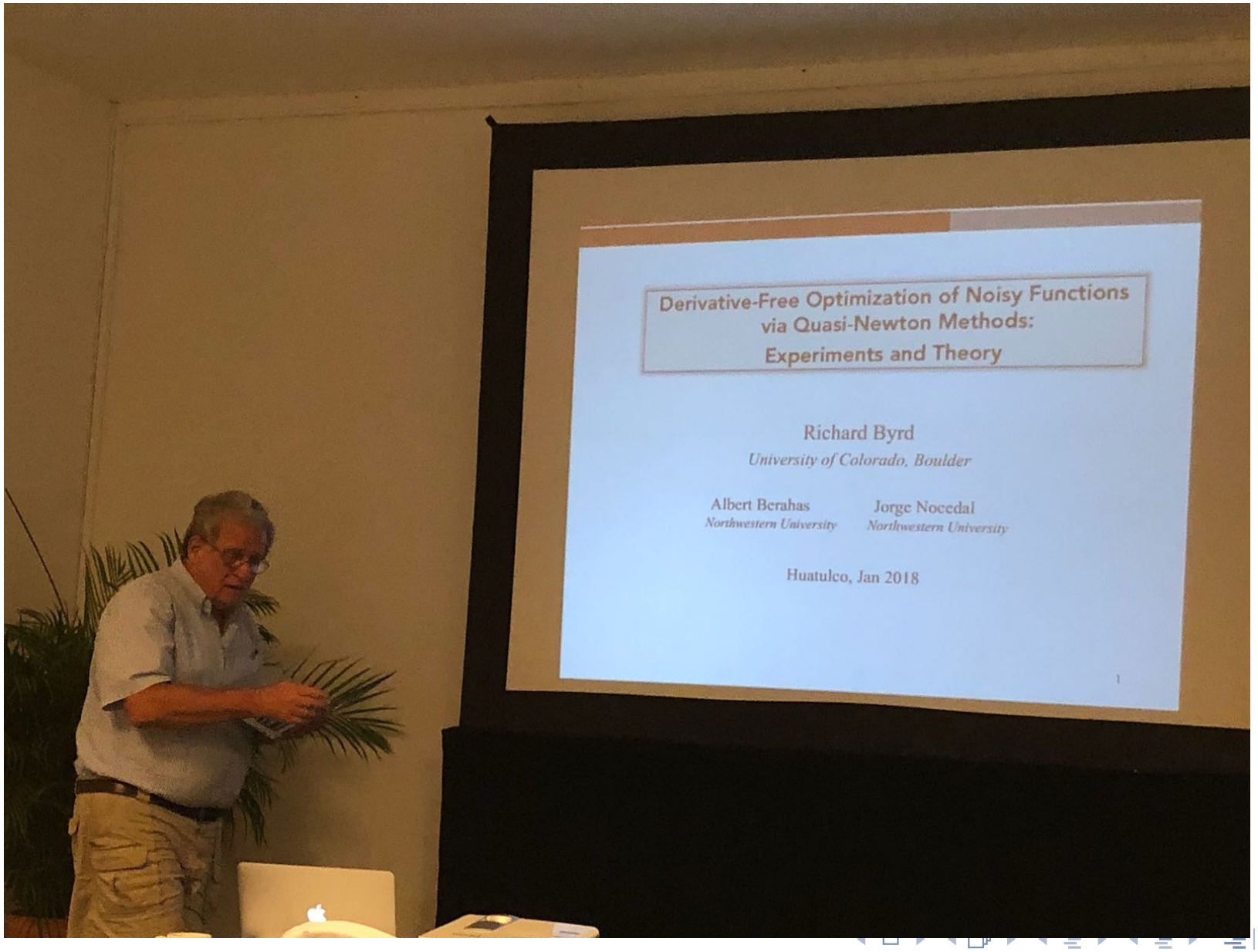
IOE 511: Continuous Optimization Methods



IOE 499/599/699: Optimization for Data Science

US Mexico 2018

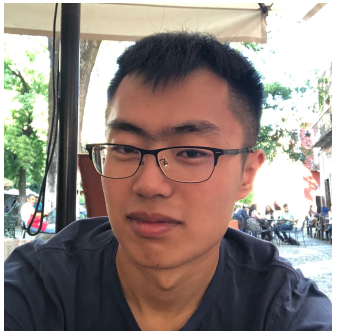
US Mexico 2018



Collaborators

Frank E. Curtis
Lehigh University

Daniel P. Robinson
Lehigh University



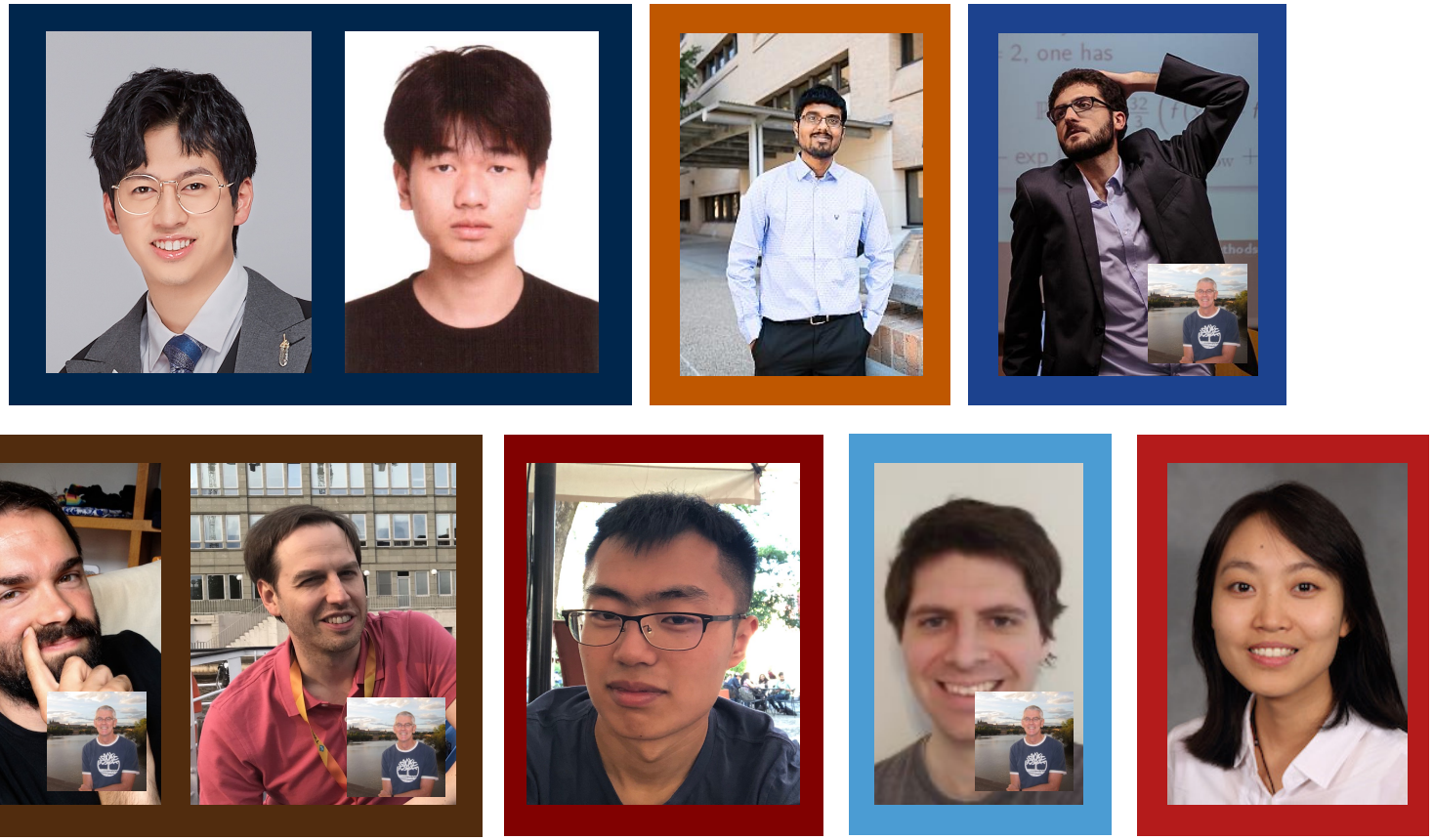
Baoyu Zhou
UChicago

- **B**, Curtis, Robinson and Zhou (2021). *Sequential Quadratic Optimization for Nonlinear Equality Constrained Stochastic Optimization*. SIAM Journal on Optimization, 31(2), 1352-1379.

  https://www.youtube.com/watch?v=HOhdpw-UX3w&ab_channel=OneWorldOptimizationSeminar

Research Team

Research Team



MIDAS MICHIGAN INSTITUTE FOR DATA SCIENCE UNIVERSITY OF MICHIGAN

MICHIGAN ENGINEERING

Outline

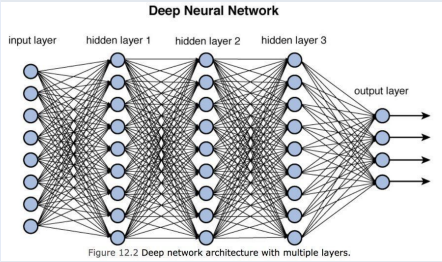
Outline

Constrained *Stochastic* Optimization

$$
\begin{array}{lll}
\min\limits_{x \in \mathbb{R}^n} & f(x) = \mathbb{E}[F(x,\omega)] & (f : \mathbb{R}^n \to \mathbb{R}) \\
\text{s.t.} & c_{\mathcal{E}}(x) = 0 & (c_{\mathcal{E}} : \mathbb{R}^n \to \mathbb{R}^{m_{\mathcal{E}}}) \\
 & c_{\mathcal{I}}(x) \leq 0 & (c_{\mathcal{I}} : \mathbb{R}^n \to \mathbb{R}^{m_{\mathcal{I}}})
\end{array}
$$

where $F : \mathbb{R}^n \times \Omega \to \mathbb{R}$, ω has a probability space $(\Omega, \mathcal{F}, P)$, $\mathbb{E}[\cdot]$ with respect to P

Applications



Constrained DNN

Optimal Power Flow

Portfolio Optimization

$$\frac{dy}{dx} = f(x)$$

$$\frac{dy}{dx} = f(x,y)$$

$$x_1 \frac{\partial y}{\partial x_1} + x_2 \frac{\partial y}{\partial x_2} = y$$

PDE constrained Optimization

This Talk

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

Assumptions:

- *Fully stochastic regime* (convergence in expectation)
- *Constraint qualifications hold* (most of the talk)
- *Feasible methods not tractable*
  - no projection methods, Frank-Wolfe, etc
- *"Two-phase" methods not effective*
  - feasibility first, then optimality (or vice-versa)

This Talk

$$\boxed{\begin{array}{ll} \min\limits_{x \in \mathbb{R}^n} & f(x) = \mathbb{E}[F(x,\omega)] \\ \text{s.t.} & c(x) = 0 \end{array}}$$

Assumptions:

- *Fully stochastic regime* (convergence in expectation)
- *Constraint qualifications hold* (most of the talk)
- *Feasible methods not tractable*
  - no projection methods, Frank-Wolfe, etc
- *"Two-phase" methods not effective*
  - feasibility first, then optimality (or vice-versa)

Goal:

1. Develop a *stochastic constrained optimization method* based on the SQP paradigm for the fully stochastic regime
2. *Extensions *(relax constraint qualifications, inexact solves, line (step) search variants, variance reduction)*

Stochastic Gradient (SG)

$$\boxed{\min_{x \in \mathbb{R}^n} \; f(x) = \mathbb{E}[F(x, \omega)]}$$

SG Algorithm

Input: x_0 (initial iterate); $\{\alpha_k\} > 0$ (stepsizes)

1: **for** $k = 0, 1, 2, \ldots$ **do**
2: Set $x_{k+1} \leftarrow x_k - \alpha_k \bar{g}_k$
3: **end for**

- **Assumptions:** *(1)* $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$ is L-Lipschitz continuous,
  (2) $\mathbb{E}_k[\bar{g}_k] = \nabla f(x_k)$ and $\mathbb{E}_k[\|\bar{g}_k - \nabla f(x_k)\|_2^2] \leq M$
- **Complications:** *Not a descent method...*

$$\mathbb{E}_k[f(x_{k+1})] - f(x_k) \leq \underbrace{-\alpha_k \|\nabla f(x_k)\|_2^2}_{\mathcal{O}(\alpha_k),\ \text{deterministic}} + \underbrace{\tfrac{1}{2}\alpha_k^2 L \mathbb{E}_k[\|\bar{g}_k\|_2^2]}_{\mathcal{O}(\alpha_k^2),\ \text{stochastic/noise}}$$

SG Theory

Theorem *(informal)*

If $\mathbb{E}_k[\bar{g}_k] = \nabla f(x_k)$ *and* $\mathbb{E}_k[\|\bar{g}_k - \nabla f(x_k)\|_2^2] \leq M$*, then, for all* k

$$\alpha_k = \alpha = \frac{1}{L} : \qquad \mathbb{E}\left[\frac{1}{K}\sum_{k=0}^{K-1}\|\nabla f(x_k)\|_2^2\right] \leq \mathcal{O}(M)$$

$$\alpha_k = \mathcal{O}\left(\frac{1}{k}\right) : \qquad \liminf_{k\to\infty} \mathbb{E}\left[\|\nabla f(x_k)\|_2^2\right] = 0$$

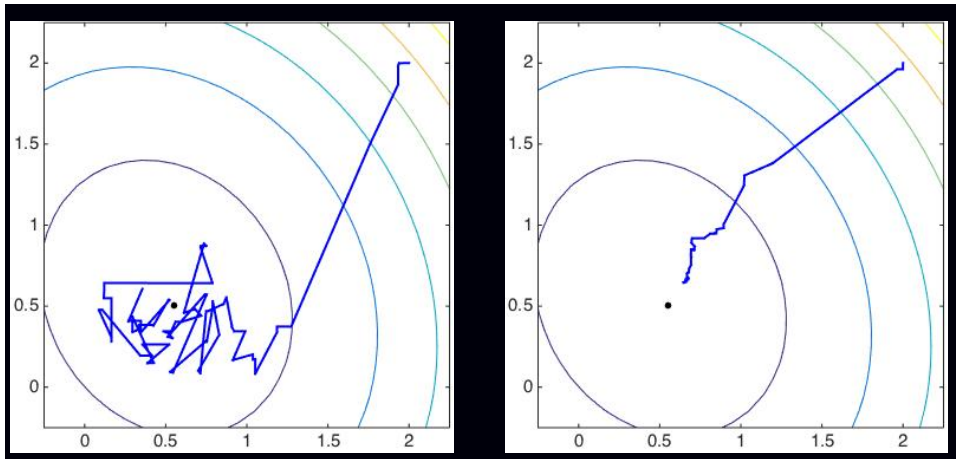


Figure: SG fixed stepsize (left) and SG diminishing stepsize (right).

Sequential Quadratic Programming (SQP)

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

Main Idea:

- **Interpretation:** @ x_k

$$\begin{aligned} \min_{d \in \mathbb{R}^n} \quad & f(x_k) + \nabla f(x_k)^T d + \tfrac{1}{2} d^T H_k d \\ \text{s.t.} \quad & c(x_k) + \nabla c(x_k)^T d = 0 \end{aligned}$$

- **Step computation "Newton-SQP system":**

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

(y_k Langrange multiplier; H_k assumed to be positive definite on $\text{Null}(\nabla c)$)

SQP in Practice

$$\min_{d \in \mathbb{R}^n} \; f(x_k) + \nabla f(x_k)^T d + \tfrac{1}{2} d^T H_k d$$
$$\text{s.t.} \; c(x_k) + \nabla c(x_k)^T d = 0$$

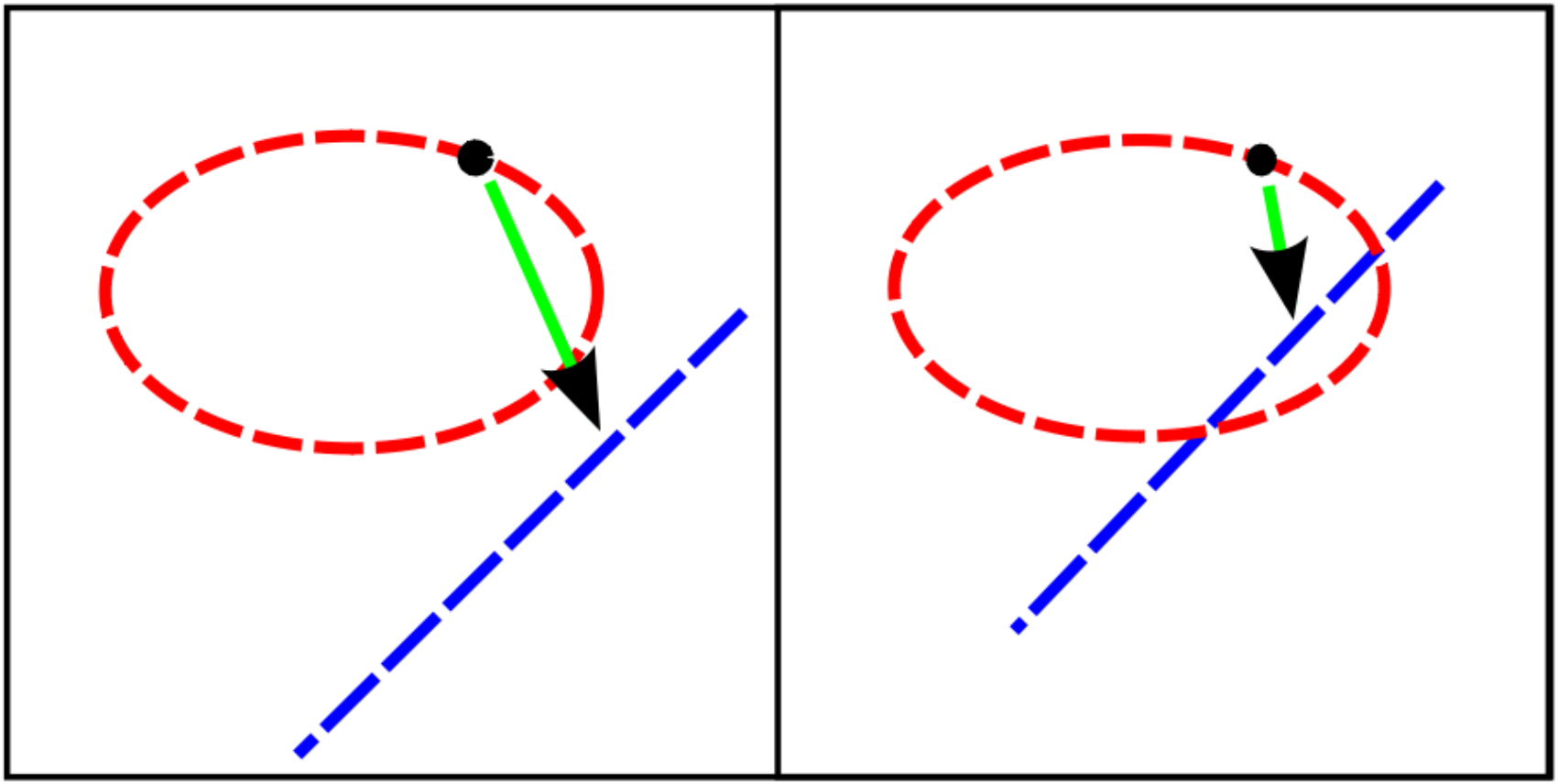


Figure: SQP subproblem solutions.

SQP–Merit Function

- **Merit Function:** guides algorithm, $\tau > 0$ (merit parameter)

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

- **Model of Merit Function:**

$$q(x, \tau, d) = \tau(f(x) + \nabla f(x)^T d + \tfrac{1}{2}\max\{d^T H d, 0\}) + \|c(x) + \nabla c(x) d\|_1$$

- **Reduction in Model of Merit Function:** for any $d \in \mathbb{R}^n$ satisfying $c(x) + \nabla c(x)^T d = 0$

$$\begin{aligned}\Delta q(x, \tau, d) &= q(x, \tau, 0) - q(x, \tau, d) \\ &= -\tau(\nabla f(x)^T d + \tfrac{1}{2}\max\{d^T H d, 0\}) + \|c(x)\|_1\end{aligned}$$

Lemma *(informal)*

For any $d \in \mathbb{R}^n$ satisfying $c(x) + \nabla c(x)^T d = 0$, we have

$$\phi'(x, \tau, d) \leq -\Delta q(x, \tau, d)$$

Line Search-SQP

SQP Backtracking

Input: x_0 (initial iterate); $\tau_{-1} > 0$ (initial penalty parameter)

1: **for** $k = 0, 1, 2, \ldots$ **do**

2: Compute step: solve

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

3: Update merit parameter: update $\tau_k > 0$ to ensure $\Delta q(x_k, \tau_k, d_k) \gg 0$

$$\tau_k \leftarrow \frac{\frac{1}{2}\|c(x_k)\|_1}{\nabla f(x)^T d_k + \max\{d_k^T H_k d_k, 0\}} \quad \text{if} \quad \nabla f(x)^T d_k + \max\{d_k^T H_k d_k, 0\} > 0$$

4: Line Search: find α_k such that $x_{k+1} \leftarrow x_k + \alpha_k d_k$ yields

$$\phi(x_k + \alpha_k d_k, \tau_k) \leq \phi(x_k, \tau_k) - \tfrac{1}{2}\alpha_k \Delta q(x_k, \tau_k, d_k)$$

5: **end for**

SQP Theory

Assumptions

- f, c, ∇f, ∇c bounded and Lipschitz
- singular values of ∇c bounded below (e.g., LICQ)

Theorem *(informal)*

We have (for the SQP Backtracking algorithm) that:

- $\{\alpha_k\} \geq \alpha_{\min}$, *for some* $\alpha_{\min} > 0$
- $\{\tau_k\} \geq \tau_{\min}$, *for some* $\tau_{\min} > 0$
- $\Delta q(x_k, \tau_k, d_k) \to 0$, *which implies*

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

Outline

Steps Towards *Stochsatic Adaptive* SQP: Step Size Selection

- In **Line Search-SQP**, step size chosen based on reducing the merit function

Steps Towards *Stochsatic Adaptive* SQP: Step Size Selection

- In **Line Search-SQP**, step size chosen based on reducing the merit function
- **Challenge:** Merit function is **nonsmooth**; upper bound (based on assumptions)

$$\begin{aligned}&\phi(x_k + \alpha_k d_k, \tau_k) - \phi(x_k, \tau_k)\\ &\leq \alpha_k \tau_k \nabla f(x_k)^T d_k + |1 - \alpha_k| \|c(x_k)\|_1 - \|c(x_k)\|_1 + \tfrac{1}{2}(\tau_k L_k + \Gamma_k)\alpha_k^2 \|d_k\|_2^2\end{aligned}$$

(L_k, Γ_k Lipschitz constants estimates for f and $\|c\|_1$ @ x_k)

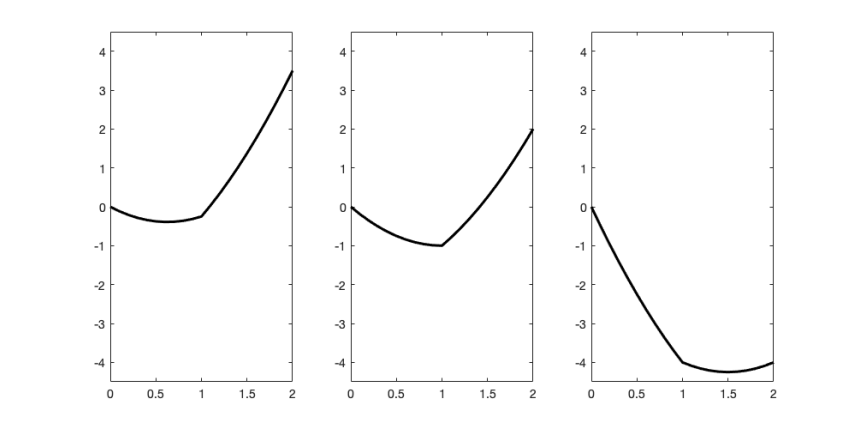


Figure: Cases of upper bound of merit function.

- **Idea:** Choose α_k to ensure sufficient decrease using this bound

Adaptive SQP Algorithm

SQP Adaptive

Input: x_0 (initial iterate); $\tau_{-1} > 0$ (initial penalty parameter)

1: **for** $k = 0, 1, 2, \ldots$ **do**

2: Compute step: solve

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

3: Update merit parameter: update $\tau_k > 0$ to ensure $\Delta q(x_k, \tau_k, d_k) \gg 0$

$$\tau_k \leftarrow \frac{\frac{1}{2}\|c(x_k)\|_1}{\nabla f(x)^T d_k + \max\{d_k^T H_k d_k, 0\}} \quad \text{if} \quad \nabla f(x)^T d_k + \max\{d_k^T H d_k, 0\} > 0$$

4: Compute step size & update iterate: set

$$\hat{\alpha}_k \leftarrow \frac{\Delta q(x_k, \tau_k, d_k)}{(\tau_k L_k + \Gamma_k)} \quad \text{and} \quad \tilde{\alpha}_k \leftarrow \hat{\alpha}_k - \frac{4\|c(x_k)\|_1}{(\tau_k L_k + \Gamma_k)}$$

$$\alpha_k \leftarrow \begin{cases} \hat{\alpha}_k & \text{if } \hat{\alpha}_k < 1 \\ 1 & \text{if } \tilde{\alpha}_k \leq 1 \leq \hat{\alpha}_k \\ \tilde{\alpha}_k & \text{if } \tilde{\alpha}_k > 1 \end{cases},$$

and, update iterate $x_{k+1} \leftarrow x_k + \alpha_k d_k$ and **continue** or update L_k and/or Γ_k and **return to step 4**

5: **end for**

Adaptive SQP–Theory & Practice

- Exactly the same theory and similar performance to **SQP Backtracking**

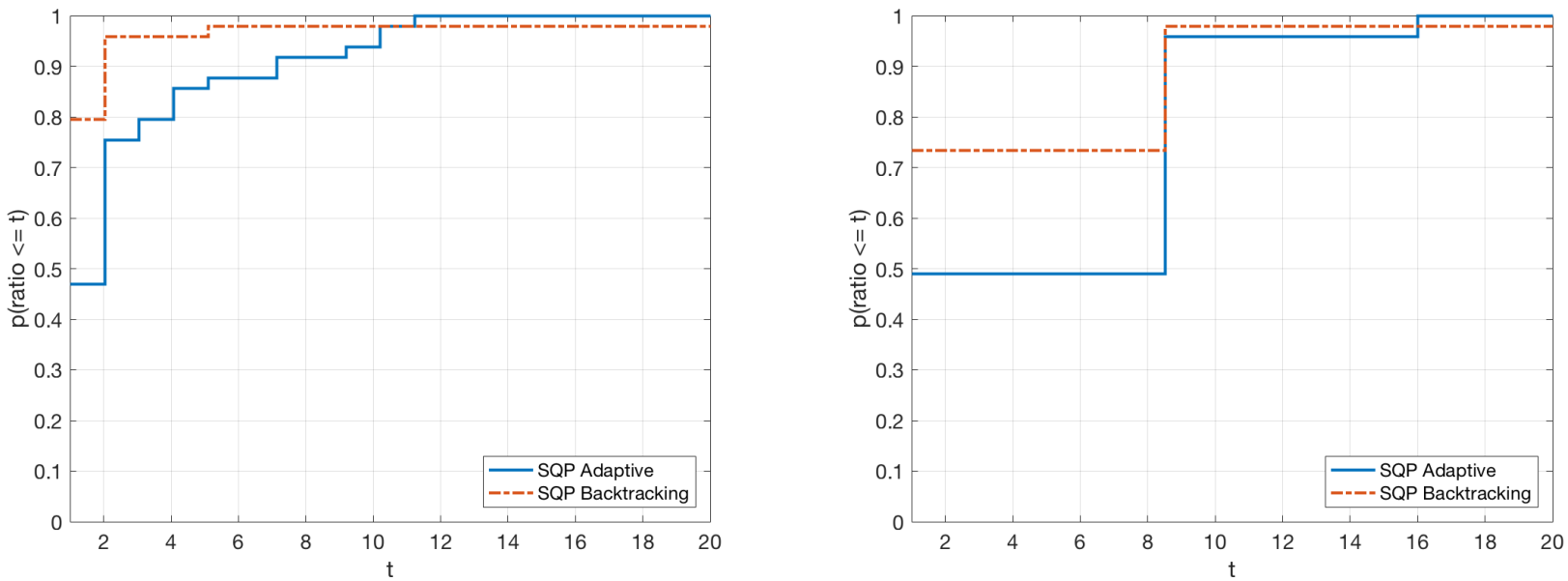


Figure: Performance profiles for SQP Adaptive and SQP Backtracking on CUTEtest set in terms of iterations (left) and function evaluations (right).

Equality Constrained *Stochastic* Optimization and *Stochastic* SQP

$$\boxed{\begin{array}{ll} \min\limits_{x \in \mathbb{R}^n} & f(x) = \mathbb{E}[F(x,\omega)] \\ \text{s.t.} & c(x) = 0 \end{array}}$$

Assumption:

- For all k, one can compute $\bar{g}_k$, such that

$$\mathbb{E}_k[\bar{g}_k] = \nabla f(x_k) \quad \text{and} \quad \mathbb{E}_k[\|\bar{g}_k - \nabla f(x_k)\|_2^2] \leq M$$

Step Computation:

- Search direction $\bar{d}_k$ is computed via:

$$\begin{bmatrix} H_k & \nabla c(x_k) \\ \nabla c(x_k)^T & 0 \end{bmatrix} \begin{bmatrix} \bar{d}_k \\ \bar{y}_k \end{bmatrix} = - \begin{bmatrix} \bar{g}_k \\ c(x_k) \end{bmatrix}$$

- Note: Given x_k, the quantities $(c(x_k), \nabla c(x_k), H_k)$ are **deterministic**

Adaptive Stochastic SQP Algorithm

Stochastic SQP

Input: x_0 (initial iterate); $\bar{\tau}_{-1} > 0$ (initial penalty parameter), $L > 0$ and $\Gamma > 0$ (Lipschitz constant estimates); $\beta_k \in (0, 1]$

1: **for** $k = 0, 1, 2, \ldots$ **do**

2: Compute step: solve

$$\begin{bmatrix} H_k & \nabla c(x_k) \\ \nabla c(x_k)^T & 0 \end{bmatrix} \begin{bmatrix} \bar{d}_k \\ \bar{y}_k \end{bmatrix} = - \begin{bmatrix} \bar{g}_k \\ c(x_k) \end{bmatrix}$$

3: Update parameter: update $\bar{\tau}_k > 0$ to ensure $\Delta\bar{q}(x_k, \bar{\tau}_k, \bar{d}_k) \gg 0$

$$\bar{\tau}_k \leftarrow \frac{\frac{1}{2}\|c(x_k)\|_1}{\bar{g}_k{}^T \bar{d}_k + \max\{\bar{d}_k{}^T H_k \bar{d}_k, 0\}} \quad \text{if} \quad \bar{g}_k{}^T \bar{d}_k + \max\{\bar{d}_k{}^T H_k \bar{d}_k, 0\} > 0$$

4: Compute step size & update iterate: set

$$\hat{\alpha}_k \leftarrow \frac{\beta_k \Delta\bar{q}(x_k, \bar{\tau}_k, \bar{d}_k)}{(\bar{\tau}_k L + \Gamma)} \quad \text{and} \quad \tilde{\alpha}_k \leftarrow \hat{\alpha}_k - \frac{4\|c(x_k)\|_1}{(\bar{\tau}_k L + \Gamma)}$$

$$\bar{\alpha}_k \leftarrow \begin{cases} \hat{\alpha}_k & \text{if } \hat{\alpha}_k < 1 \\ 1 & \text{if } \tilde{\alpha}_k \leq 1 \leq \hat{\alpha}_k \\ \tilde{\alpha}_k & \text{if } \tilde{\alpha}_k > 1 \end{cases},$$

and, update iterate $x_{k+1} \leftarrow x_k + \bar{\alpha}_k \bar{d}_k$

5: **end for**

Step Size Selection (*Adaptive Stochastic* SQP)

- The sequence β_k allows us to consider (as for the SG method)
  - fixed stepsizes
  - diminishing stepsizes

Step Size Selection (*Adaptive Stochastic* SQP)

- The sequence β_k allows us to consider (as for the SG method)
  - fixed stepsizes
  - diminishing stepsizes

- Unfortunately, **additional control needed**
  - too small: slow or insufficient progress
  - too big: may ruin progress towards optimality/feasibility
- **Idea:** Project $\hat{\alpha}_k$ and $\tilde{\alpha}_k$ onto

$$\left[\frac{\beta_k \bar{\tau}_k}{\bar{\tau}_k L + \Gamma}, \frac{\beta_k \bar{\tau}_k}{\bar{\tau}_k L + \Gamma} + \theta \beta_k^2\right]$$

($\theta > 0$ user-defined parameter), then follow rule in algorithm

Merit Parameter Behavior

Deterministic setting

- $\{\tau_k\} \geq \tau_{\min}$

Stochastic setting

- No necessarily true...
  - "good" case: $\{\bar{\tau}_k\}$ becomes sufficiently small (compared to $\tau_{\min}$)
  - "poor" case: $\{\bar{\tau}_k\} \searrow 0$
  - "poor" case: $\{\bar{\tau}_k\}$ remains too large

Note: "poor" cases exists is solely due to the use of stochastic gradients

Main Theoretical Results

Theorem (Good Merit Parameter Behavior)

If $\{\bar{\tau}_k\}$ eventually remains fixed at sufficient small $\tau_{\min}$, the for large k

$$\beta_k = \beta = \mathcal{O}(1): \qquad \mathbb{E}\left[\frac{1}{K}\sum_{k=0}^{K-1} \Delta q(x_k, \tau_{\min}, d_k)\right] \leq \mathcal{O}(M)$$

$$\beta_k = \mathcal{O}\left(\frac{1}{k}\right): \qquad \liminf_{k\to\infty} \mathbb{E}\left[\Delta q(x_k, \tau_{\min}, d_k)\right] = 0$$

Main Theoretical Results

Theorem (Good Merit Parameter Behavior)

If $\{\bar{\tau}_k\}$ *eventually remains fixed at sufficient small* $\tau_{\min}$*, the for large* k

$$\beta_k = \beta = \mathcal{O}(1): \qquad \mathbb{E}\left[\frac{1}{K}\sum_{k=0}^{K-1}\left(\|\nabla f(x_k) + \nabla c(x_k)^T y_k\|_2^2 + \|c(x_k)\|_2\right)\right] \leq \mathcal{O}(M)$$

$$\beta_k = \mathcal{O}\left(\frac{1}{k}\right): \qquad \liminf_{k\to\infty} \mathbb{E}\left[\left(\|\nabla f(x_k) + \nabla c(x_k)^T y_k\|_2^2 + \|c(x_k)\|_2\right)\right] = 0$$

Main Theoretical Results

Theorem (Good Merit Parameter Behavior)

If $\{\bar{\tau}_k\}$ *eventually remains fixed at sufficient small* $\tau_{\min}$*, the for large* k

$$\beta_k = \beta = \mathcal{O}(1): \qquad \mathbb{E}\left[\frac{1}{K}\sum_{k=0}^{K-1}\left(\|\nabla f(x_k) + \nabla c(x_k)^T y_k\|_2^2 + \|c(x_k)\|_2\right)\right] \leq \mathcal{O}(M)$$

$$\beta_k = \mathcal{O}\left(\frac{1}{k}\right): \qquad \liminf_{k\to\infty} \mathbb{E}\left[\left(\|\nabla f(x_k) + \nabla c(x_k)^T y_k\|_2^2 + \|c(x_k)\|_2\right)\right] = 0$$

Poor Merit Parameter Behavior

- $\{\bar{\tau}_k\}$ remains too large: *if distribution symmetric* **probability zero**
- $\{\bar{\tau}_k\} \searrow 0$: *under reasonable assumptions* **cannot occur**

Numerical Experiments

- CUTE problems with noise added to the gradient (different noise levels)
- 49 problems ($n + m \leq 1000$ and LICQ holds), 10 different random seeds
- Stochastic SQP (10^3 iterations)
- Stochastic Subgradient (10^4 iterations and tuned over 11 values of τ);

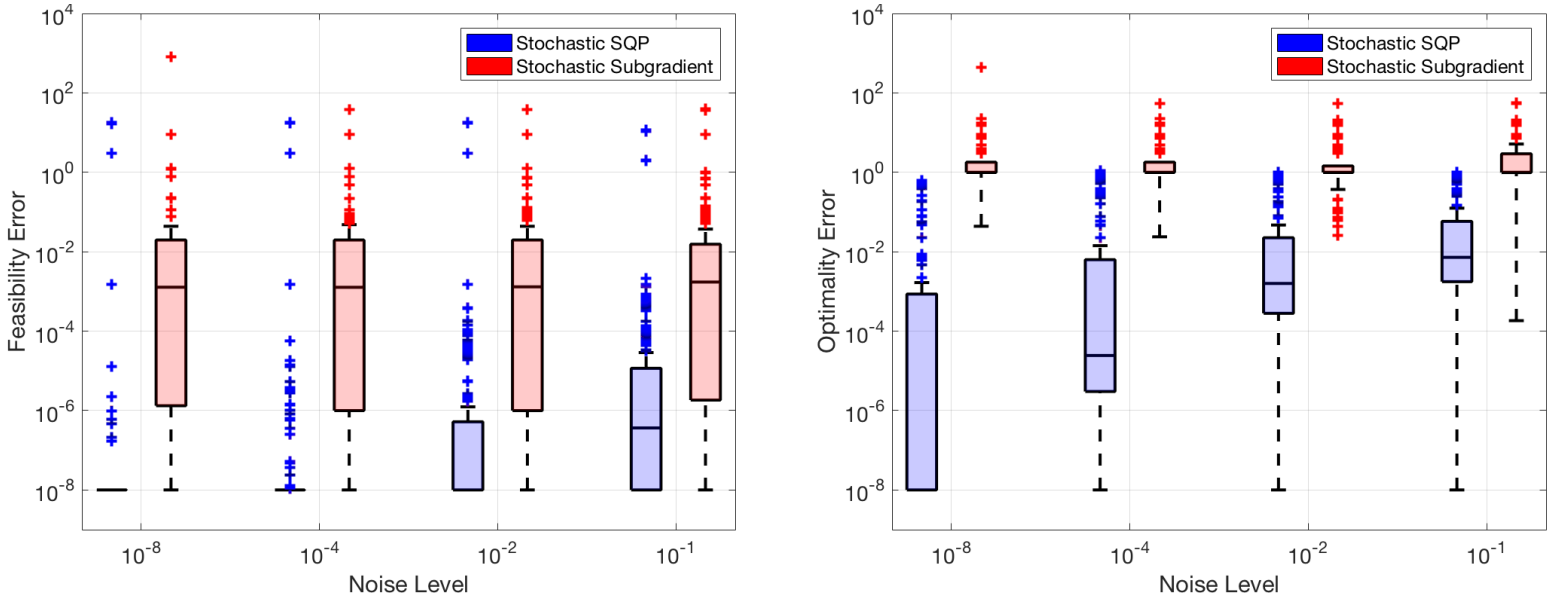


Figure: Box plots for feasibility errors (left) and optimality errors (right).

Outline

Extension 1: Relaxing constraint qualifications

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

Assumptions

- f, c, ∇f, ∇c bounded and Lipschitz
- ~~singular values of ∇c bounded below (i.e., LICQ)~~

Challenges

- primal stationarity **not necessary**
- constraints may be **infeasible**

- **B**, Curtis, O'Neil and Robinson. *A Stochastic Sequential Quadratic Optimization Algorithm for Nonlinear Equality Constrained Optimization with Rank-Deficient Jacobians*. arXiv preprint arXiv:2106.13015. *(submitted)*

Extension 1: Relaxing constraint qualifications

$$\begin{array}{ll} \min\limits_{x \in \mathbb{R}^n} & f(x) = \mathbb{E}[F(x,\omega)] \\ \text{s.t.} & c(x) = 0 \end{array}$$

Assumptions

- f, c, ∇f, ∇c bounded and Lipschitz
- ~~singular values of ∇c bounded below (i.e., LICQ)~~

Challenges

- primal stationarity **not necessary**
- constraints may be **infeasible**

What changes...

- **Algorithm** (search direction computation, stepsize update)
- **Theory** (convergence to infeasible stationary point possible)

- **B**, Curtis, O'Neil and Robinson. *A Stochastic Sequential Quadratic Optimization Algorithm for Nonlinear Equality Constrained Optimization with Rank-Deficient Jacobians*. arXiv preprint arXiv:2106.13015. *(submitted)*

Extension 1: Relaxing constraint qualifications

Search direction computation

LICQ:

$$\begin{bmatrix} H_k & \nabla c(x_k) \\ \nabla c(x_k)^T & 0 \end{bmatrix} \begin{bmatrix} \bar{d}_k \\ \bar{y}_k \end{bmatrix} = - \begin{bmatrix} \bar{g}_k \\ c(x_k) \end{bmatrix}$$

NO LICQ:

- **Step decomposition (Byrd-Omojokun):**

$$\bar{d}_k := v_k + \bar{u}_k$$

- Normal component v_k: reduce linearized constraint violation

$$\min_v \tfrac{1}{2}\|c(x_k) + \nabla c(x_k)^T v\|_2^2 \quad \text{s.t.} \quad \|v\|_2 \leq \omega \|\nabla c(x_k) c(x_k)\|_2$$

- Tangential component $\bar{u}_k$: minimize model of objective function s.t. remaining in null space of ∇c^T

$$\begin{bmatrix} H_k & \nabla c(x_k) \\ \nabla c(x_k)^T & 0 \end{bmatrix} \begin{bmatrix} \bar{u}_k \\ \bar{y}_k \end{bmatrix} = - \begin{bmatrix} \bar{g}_k + H_k v_k \\ 0 \end{bmatrix}$$

Extension 1: Relaxing constraint qualifications

Stepsize Update

LICQ:

- Adaptive selection, β_k control, projection $\left[\frac{\beta_k \bar{\tau}_k}{\bar{\tau}_k L+\Gamma}, \frac{\beta_k \bar{\tau}_k}{\bar{\tau}_k L+\Gamma} + \theta {\beta_k}^2\right]$

NO LICQ:

- **Similar** adaptive rule–more careful selection (required both for theory and practice)
- Iteration (search direction) dependent projection

$$\begin{cases} \left[\frac{\beta_k \bar{\tau}_k}{\bar{\tau}_k L+\Gamma}, \frac{\beta_k \bar{\tau}_k}{\bar{\tau}_k L+\Gamma} + \theta {\beta_k}^2\right], & \text{if } \|u_k\|_2^2 \geq \chi_k \|v_k\|_2^2 \\ \left[\frac{\beta_k}{\bar{\tau}_k L+\Gamma}, \frac{\beta_k}{\bar{\tau}_k L+\Gamma} + \theta {\beta_k}^2\right], & \text{otherwise} \end{cases}$$

where χ_{-1} is user defined, then adaptive

Extension 1: Relaxing constraint qualifications

Merit Parameter Behavior

Deterministic setting

- $\{\tau_k\} \geq \tau_{\min}$

Stochastic setting

- No necessarily true...
  - "good" case: $\{\bar{\tau}_k\}$ becomes sufficiently small (as compared to $\tau_{\min}$)
  -
  - "poor" case: $\{\bar{\tau}_k\} \searrow 0$
  - "poor" case: $\{\bar{\tau}_k\}$ remains too large
-

- Note: "poor" cases exists is solely due to the use of stochastic gradients

Extension 1: Relaxing constraint qualifications

Merit Parameter Behavior

Deterministic setting **(no LICQ)**

- $\cancel{\{\tau_k\} \geq \tau_{\min}}$

Stochastic setting **(no LICQ)**

- No necessarily true...
  - "good" case: $\{\bar{\tau}_k\}$ becomes sufficiently small (as compared to $\tau_{\min}$)
  - "good" case: $\{\bar{\tau}_k\} \searrow 0$ (due to rank deficiency)
  - "poor" case: $\{\bar{\tau}_k\} \searrow 0$
  - "poor" case: $\{\bar{\tau}_k\}$ remains too large
- **Difficulty:** distinguish between the $\{\bar{\tau}_k\} \searrow 0$ cases

- Note: "poor" cases exists is solely due to the use of stochastic gradients

Extension 1: Relaxing constraint qualifications

Main Results

Theorem

If $\{\bar{\tau}_k\}$ *eventually remains fixed at sufficient small* $\tau_{\min}$ *, then for large* k

$$\beta_k = \beta = \mathcal{O}(1): \quad \mathbb{E}\left[\frac{1}{K}\sum_{k=0}^{K-1}\left(\|\nabla f(x_k) + \nabla c(x_k) y_k\|_2^2 + \|\nabla c(x_k) c(x_k)\|_2\right)\right] \leq \mathcal{O}(M)$$

$$\beta_k = \mathcal{O}\left(\frac{1}{k}\right): \quad \liminf_{k\to\infty} \mathbb{E}\left[\left(\|\nabla f(x_k) + \nabla c(x_k) y_k\|_2^2 + \|\nabla c(x_k) c(x_k)\|_2\right)\right] = 0$$

Extension 1: Relaxing constraint qualifications

Main Results

Theorem

If $\{\bar{\tau}_k\}$ eventually remains fixed at sufficient small $\tau_{\min}$, and in addition the singular values of ∇c are bounded, then for large k

$$\beta_k = \beta = \mathcal{O}(1): \qquad \mathbb{E}\left[\frac{1}{K}\sum_{k=0}^{K-1}\left(\|\nabla f(x_k) + \nabla c(x_k) y_k\|_2^2 + \kappa_c \|c(x_k)\|_2\right)\right] \leq \mathcal{O}(M)$$

$$\beta_k = \mathcal{O}\left(\frac{1}{k}\right): \qquad \liminf_{k\to\infty} \mathbb{E}\left[\left(\|\nabla f(x_k) + \nabla c(x_k) y_k\|_2^2 + \kappa_c \|c(x_k)\|_2\right)\right] = 0$$

Extension 1: Relaxing constraint qualifications

Main Results

Theorem

If $\{\bar{\tau}_k\}$ eventually remains fixed at sufficient small $\tau_{\min}$, and in addition the singular values of ∇c are bounded, then for large k

$$\beta_k = \beta = \mathcal{O}(1): \qquad \mathbb{E}\left[\frac{1}{K}\sum_{k=0}^{K-1}\left(\|\nabla f(x_k) + \nabla c(x_k) y_k\|_2^2 + \kappa_c \|c(x_k)\|_2\right)\right] \leq \mathcal{O}(M)$$

$$\beta_k = \mathcal{O}\left(\frac{1}{k}\right): \qquad \liminf_{k\to\infty} \mathbb{E}\left[\left(\|\nabla f(x_k) + \nabla c(x_k) y_k\|_2^2 + \kappa_c \|c(x_k)\|_2\right)\right] = 0$$

Theorem

If $\{\bar{\tau}_k\} \searrow 0$ and $\|\bar{g}_k - \nabla f(x_k)\|_2^2 \leq M$, then for large k

$$\beta_k = \beta = \mathcal{O}(1): \qquad \liminf_{k\to\infty} \|\nabla c(x_k) c(x_k)\|_2 = 0$$

Extension 1: Relaxing constraint qualifications

Numerical Experiment 1: CUTE problems

- CUTE problems with noise added to the gradient (different noise levels)
- ...and additional redundant constraint

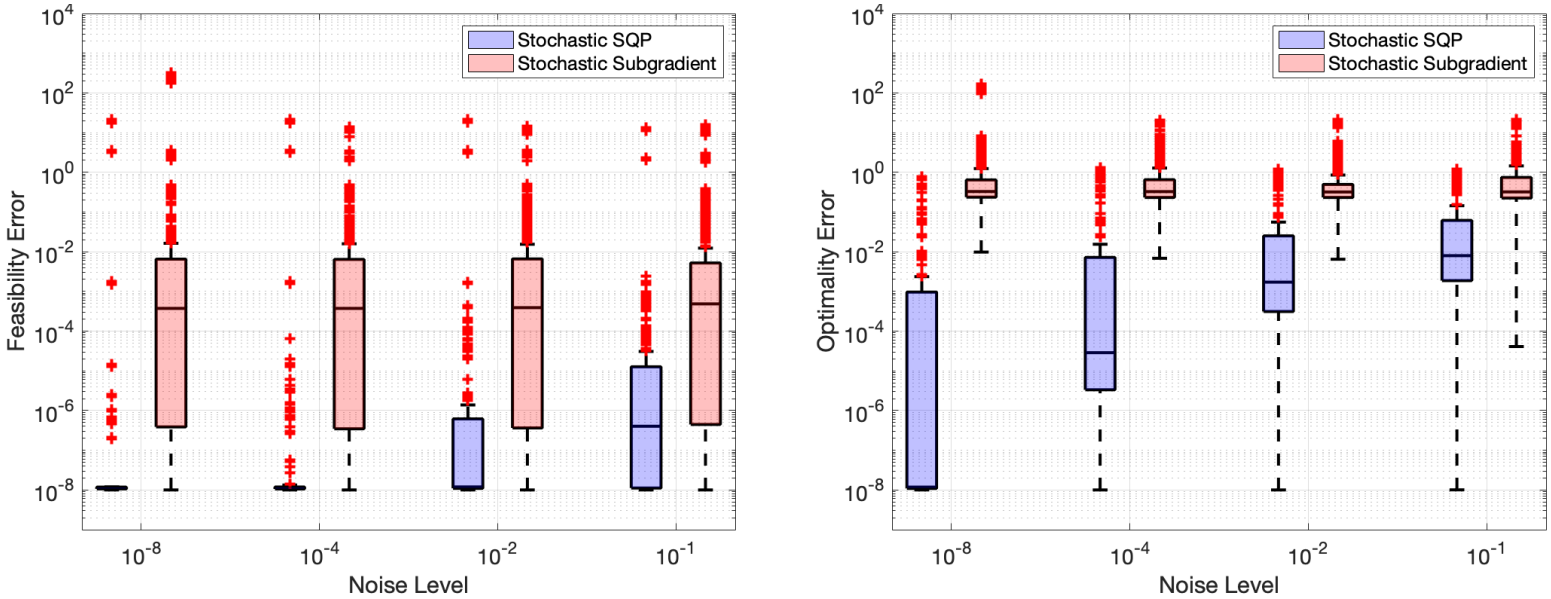


Figure: Box plots for feasibility errors (left) and optimality errors (right).

Extension 1: Relaxing constraint qualifications

Numerical Experiment 2: Constrained Logistic Regression (with redundant constraints)

$$\min_{x \in \mathbb{R}^n} \; f(x) = \frac{1}{N} \sum_{i=1}^{N} \log\left(1 + e^{-y_i (X_i^T x)}\right) \quad \text{s.t.} \quad Ax = b,$$

- Stochastic Subgradient, Stochastic Projected Gradient, Stochastic SQP

Table 2: Average feasibility and stationarity errors, along with 95% confidence intervals, when our stochastic SQP method, a stochastic subgradient method, and a stochastic projected gradient method are employed to solve logistic regression problems with linear constraints (only). The results for the best-performing algorithm are shown in bold.

		Stochastic Subgradient		Stochastic Projected Gradient	Stochastic SQP	
dataset	batch	Feasibility	Stationarity	Stationarity	Feasibility	Stationarity
a9a	16	$8.30e-03 \pm 2.32e-03$	$1.64e-01 \pm 3.55e-03$	$3.64e-02 \pm 2.95e-03$	$\mathbf{1.22e-15 \pm 2.18e-16}$	$\mathbf{9.99e-03 \pm 6.92e-03}$
a9a	128	$1.16e-02 \pm 4.60e-05$	$1.69e-01 \pm 2.51e-02$	$1.69e-02 \pm 2.79e-03$	$\mathbf{1.64e-15 \pm 4.00e-16}$	$\mathbf{7.33e-03 \pm 4.68e-05}$
australian	16	$7.94e-02 \pm 1.60e-05$	$7.94e-02 \pm 1.60e-05$	$9.17e-02 \pm 4.32e-04$	$\mathbf{5.72e-06 \pm 1.56e-06}$	$\mathbf{2.67e-02 \pm 6.43e-04}$
australian	128	$5.02e-01 \pm 7.04e-05$	$5.02e-01 \pm 7.04e-05$	$1.11e-02 \pm 7.19e-05$	$\mathbf{6.58e-05 \pm 7.90e-07}$	$\mathbf{5.50e-02 \pm 1.08e-03}$
heart	16	$3.66e-01 \pm 4.37e-03$	$3.28e+01 \pm 7.02e+00$	$\mathbf{3.17e+01 \pm 6.72e+00}$	$\mathbf{8.83e-03 \pm 2.77e-03}$	$3.39e+01 \pm 9.85e+00$
heart	128	$1.52e+00 \pm 4.96e-02$	$\mathbf{1.23e+01 \pm 1.40e+01}$	$3.29e+01 \pm 3.21e+00$	$\mathbf{1.26e-01 \pm 7.86e-04}$	$3.24e+01 \pm 1.76e+00$
ijccn1	16	$3.58e-03 \pm 2.00e-05$	$4.70e-02 \pm 6.45e-07$	$7.41e-02 \pm 3.33e-07$	$\mathbf{3.03e-15 \pm 6.20e-16}$	$\mathbf{1.93e-03 \pm 4.07e-06}$
ijccn1	128	$3.90e-02 \pm 4.01e-06$	$5.17e-02 \pm 1.65e-07$	$3.88e-02 \pm 6.15e-07$	$\mathbf{2.16e-09 \pm 2.62e-09}$	$\mathbf{1.70e-02 \pm 5.19e-05}$
ionosphere	16	$5.41e-01 \pm 8.80e-05$	$5.41e-01 \pm 8.80e-05$	$9.77e-01 \pm 8.55e-03$	$\mathbf{9.61e-07 \pm 2.77e-09}$	$\mathbf{4.17e-02 \pm 1.08e-03}$
ionosphere	128	$5.76e+00 \pm 3.76e-05$	$5.76e+00 \pm 3.76e-05$	$5.98e+00 \pm 3.21e-03$	$\mathbf{1.31e-05 \pm 1.14e-09}$	$\mathbf{1.55e-01 \pm 2.61e-03}$
madelon	16	$3.06e-02 \pm 1.85e-02$	$5.46e+01 \pm 1.25e+01$	$2.11e+01 \pm 2.72e+00$	$\mathbf{2.88e-08 \pm 5.51e-08}$	$\mathbf{1.09e+01 \pm 3.00e+00}$
madelon	128	$1.87e+00 \pm 7.62e-01$	$2.21e+01 \pm 1.55e+01$	$\mathbf{2.16e+01 \pm 4.17e+00}$	$\mathbf{5.81e-01 \pm 1.63e-02}$	$4.81e+01 \pm 4.75e+00$
mushrooms	16	$2.19e-01 \pm 6.55e-04$	$2.19e-01 \pm 6.55e-04$	$7.31e-03 \pm 3.21e-06$	$\mathbf{2.08e-15 \pm 3.28e-16}$	$\mathbf{5.95e-03 \pm 3.21e-05}$
mushrooms	128	$4.73e-01 \pm 4.37e-05$	$4.73e-01 \pm 4.37e-05$	$3.31e-02 \pm 7.13e-05$	$\mathbf{1.66e-09 \pm 6.20e-14}$	$\mathbf{3.28e-02 \pm 9.15e-04}$
phishing	16	$2.67e-02 \pm 2.76e-07$	$3.47e-02 \pm 1.39e-09$	$\mathbf{2.20e-05 \pm 9.29e-06}$	$\mathbf{4.26e-15 \pm 1.27e-15}$	$3.37e-03 \pm 1.27e-06$
phishing	128	$3.06e-01 \pm 1.13e-06$	$3.06e-01 \pm 1.13e-06$	$2.29e-01 \pm 8.88e-03$	$\mathbf{1.83e-15 \pm 4.99e-16}$	$\mathbf{2.20e-02 \pm 7.29e-03}$
sonar	16	$1.33e+00 \pm 1.08e-04$	$1.33e+00 \pm 1.08e-04$	$6.13e-01 \pm 2.22e-03$	$\mathbf{7.02e-07 \pm 1.60e-07}$	$\mathbf{2.34e-02 \pm 2.03e-04}$
sonar	128	$1.33e+01 \pm 1.48e-04$	$1.33e+01 \pm 1.48e-04$	$6.46e-02 \pm 4.73e-03$	$\mathbf{2.07e-06 \pm 6.70e-10}$	$\mathbf{2.98e-02 \pm 1.71e-03}$
splice	16	$2.56e-03 \pm 3.39e-04$	$4.56e-01 \pm 3.55e-02$	$9.65e-01 \pm 3.19e-03$	$\mathbf{7.49e-14 \pm 1.03e-13}$	$\mathbf{2.19e-02 \pm 4.33e-03}$
splice	128	$3.14e-01 \pm 1.09e-04$	$4.83e-01 \pm 4.65e-05$	$1.23e+00 \pm 9.44e-05$	$\mathbf{3.54e-08 \pm 5.74e-09}$	$\mathbf{1.07e-02 \pm 3.16e-04}$
w8a	16	$2.38e-02 \pm 1.75e-03$	$1.47e-01 \pm 1.89e-06$	$9.85e-04 \pm 3.31e-05$	$\mathbf{7.35e-15 \pm 6.98e-16}$	$\mathbf{6.07e-05 \pm 6.46e-05}$
w8a	128	$1.79e-02 \pm 1.25e-03$	$1.49e-01 \pm 4.64e-03$	$3.41e-02 \pm 7.43e-03$	$\mathbf{5.96e-15 \pm 5.67e-16}$	$\mathbf{1.20e-03 \pm 1.85e-03}$

Extension 2: Line (Step) Search Stochastic SQP

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

Assumptions

- f, c, ∇f, ∇c bounded and Lipschitz
- singular values of ∇c bounded below (i.e., LICQ)

Oracles

- **Zeroth-order:** bounded noise (ϵ_f)
- **Probabilistic First-order:** $\mathbb{P}[\|\bar{g}_k - \nabla f(x)\| \leq \max\{\epsilon_g, \kappa_{FO}\alpha\Delta\bar{q}(x_k, \bar{\tau}_k, \bar{d}_k)\}] \geq 1 - \delta$

- **B**, Xie and Zhou. *A Sequential Quadratic Programming Method with High Probability Complexity Bounds for Nonlinear Equality Constrained Stochastic Optimization.* *(submitted)*

Extension 2: Line (Step) Search Stochastic SQP

$$\begin{array}{ll} \min\limits_{x \in \mathbb{R}^n} & f(x) = \mathbb{E}[F(x,\omega)] \\ \text{s.t.} & c(x) = 0 \end{array}$$

Assumptions

- f, c, ∇f, ∇c bounded and Lipschitz
- singular values of ∇c bounded below (i.e., LICQ)

Oracles

- **Zeroth-order:** bounded noise (ϵ_f)
- **Probabilistic First-order:** $\mathbb{P}[\|\bar{g}_k - \nabla f(x)\| \leq \max\{\epsilon_g, \kappa_{FO}\alpha\Delta\bar{q}(x_k, \bar{\tau}_k, \bar{d}_k)\}] \geq 1 - \delta$

Goal

- **Algorithm:** *relaxed* step search condition (relaxation proportional to ϵ_f)
- **Theory:** high-probability bound on the iteration complexity of the algorithm

- **B**, Xie and Zhou. *A Sequential Quadratic Programming Method with High Probability Complexity Bounds for Nonlinear Equality Constrained Stochastic Optimization.* *(submitted)*

Extension 2: Line (Step) Search Stochastic SQP

- CUTE problems with noise added to the gradient and function (different noise levels)
- 49 problems ($n + m \leq 1000$ and LICQ holds), 10 different random seeds
- Stochastic SQP (AS-SQP) [part 1]; Stochastic Step Search SQP (SS-SQP)

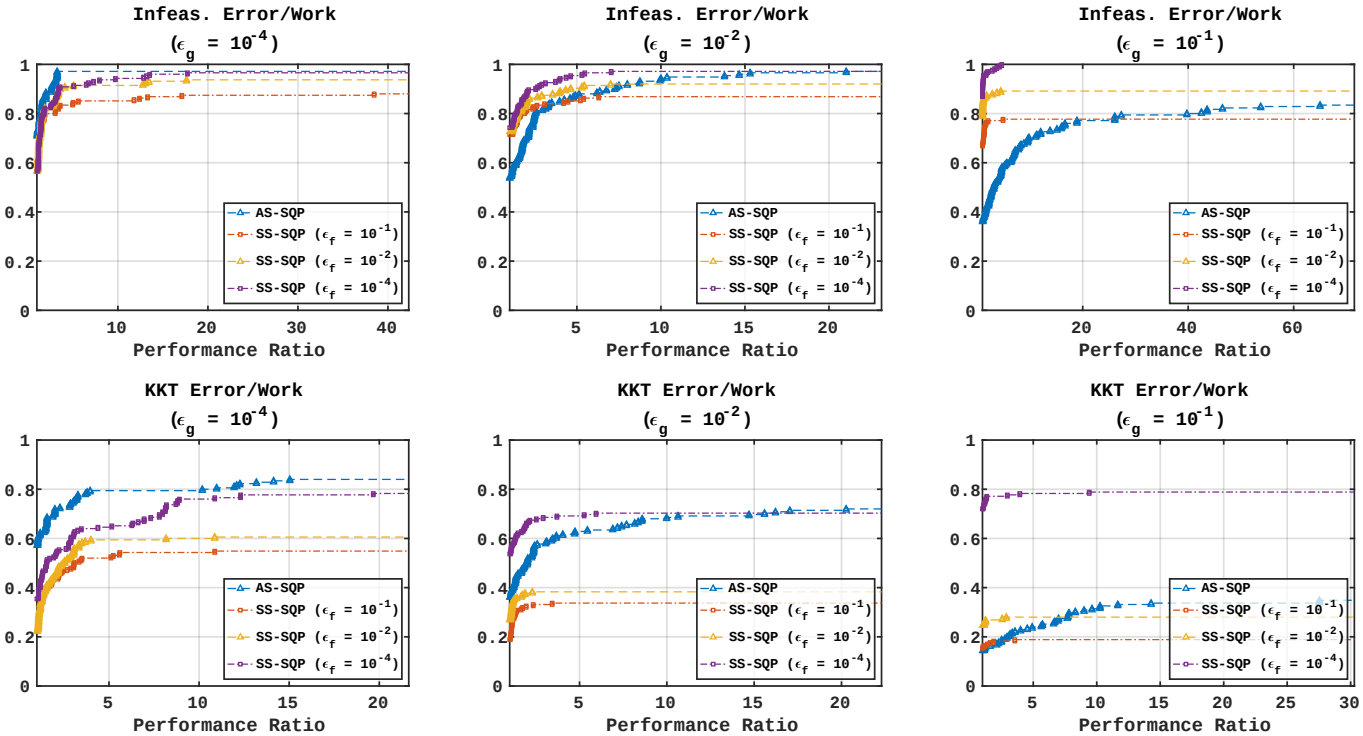


Figure: Box plots for feasibility errors (top) and optimality errors (bottom).

Extension 3: Variance Reduced Stochastic SQP

$$\min_{x \in \mathbb{R}^n} \quad f(x) = \frac{1}{N} \sum_{i=1}^{N} f_i(x)$$
$$\text{s.t.} \quad c(x) = 0$$

Idea & Motivation

- Replace $\bar{g}_k$ with variance reduced gradient approximation $\hat{g}_k$ (e.g., SVRG, SARAH)

Setting	Method	Step size		
		Diminishing	Constant	Adaptive
Unconstrained	SG	exact	neighborhood	-
Finite Sum	SVRG	-	exact	-
Equality Constrained	Stoch. SQP	exact	neighborhood	neighborhood
Finite Sum	SVRSQP	-	exact	exact

- **B**, Shi, Yi and Zhou. *Accelerating Stochastic Sequential Quadratic Programming for Equality Constrained Optimization using Predictive Variance Reduction.* *(submitted)*

Extension 3: Variance Reduced Stochastic SQP

$$\min_{x \in \mathbb{R}^n} \quad f(x) = \frac{1}{N} \sum_{i=1}^{N} f_i(x)$$
$$\text{s.t.} \quad c(x) = 0$$

Idea & Motivation

- Replace $\bar{g}_k$ with variance reduced gradient approximation $\hat{g}_k$ (e.g., SVRG, SARAH)

Setting	Method	Step size		
		Diminishing	Constant	Adaptive
Unconstrained	SG	exact	neighborhood	-
Finite Sum	SVRG	-	exact	-
Equality Constrained	Stoch. SQP	exact	neighborhood	neighborhood
Finite Sum	SVRSQP	-	exact	exact

Results

- Strong theoretical results (convergence with **fixed** β)
- Very promising preliminary numerical results; constrained binary classification

- **B**, Shi, Yi and Zhou. *Accelerating Stochastic Sequential Quadratic Programming for Equality Constrained Optimization using Predictive Variance Reduction.* *(submitted)*

Extension 3: Variance Reduced Stochastic SQP

- Logistic regression with constraints
- Stochastic SQP (StochSQP) [part 1]; Stochastic SVRG SQP (SVR-SQP); Stochastic subgradient with variance reduction (StochSubVR)

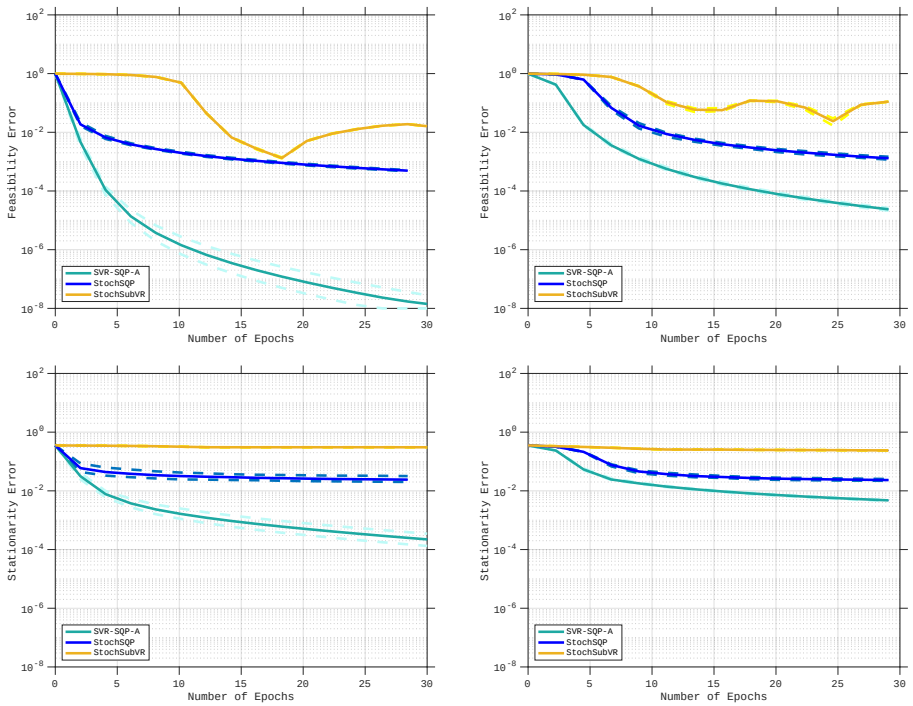


Figure: Box plots for feasibility errors (top) and optimality errors (bottom).

Other Extensions

- **Relax LICQ:** Stochastic SQP with step decomposition (**B**, Curtis, O'Neil & Robinson)
- **Variance reduction:** Stochastic SQP with SVRG gradients (**B**, Shi, Yi & Zhou)
- **Adaptive sampling:** Adaptive sampling SQP with inexact solves (**B**, Bollapragada & Zhou)
- **Line search variants:** Stochastic line search SQP (**B**, O'Neil & Royer)
- **DFO variants:** Line search/adaptive derivative-free SQP (**B**, Shi & Zhou)
- **2nd-order (exact & inexact):** Subsampled methods (**B**, Bollapragada, Shi & Zhou)

- *(Lehigh ISE friends)* **Inexact algorithms:** (Curtis, Robinson & Zhou)
- *(Lehigh ISE friends)* **Inequality constraints:** (Curtis, Li & Robinson)
- *(Lehigh ISE friends)* **Complexity Analysis:** (Curtis, O'Neil & Robinson)

Outline

Final Remarks

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

Summary:

- **Algorithms:** Adaptive SQP and Adaptive stochastic SQP
- **Theory:** Convergence in expectation (similar to SG for unconstrained optimization)
- **Practice:** Promising numerical results

Extensions:

- inexact solves, inequality constraints, complexity, adaptive sampling methods, line search variants, variance reduced variants, ...

Final Remarks

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

Summary:

- **Algorithms:** Adaptive SQP and Adaptive stochastic SQP
- **Theory:** Convergence in expectation (similar to SG for unconstrained optimization)
- **Practice:** Promising numerical results

Extensions:

- inexact solves, inequality constraints, complexity, adaptive sampling methods, line search variants, variance reduced variants, ...

Many many Fundamental Open Questions:

- *Behavior of merit parameter? Inequality constraints? Active-set identification? Lagrange multiplier computation?*

Thank You!

Questions?