Algorithms for Deterministically Constrained Stochastic Optimization

Frank E. Curtis, Lehigh University

involving joint work with

Albert Berahas, University of Michigan
Michael O'Neill, Lehigh University (soon UNC Chapel Hill)
Daniel P. Robinson, Lehigh University
Baoyu Zhou, Lehigh University (soon Chicago Booth)

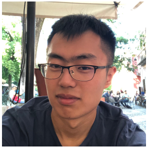
presented at

IMA Conference on Numerical Linear Algebra and Optimization

June 29, 2022

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.

Outline

Motivation

SG and SQP

Adaptive (Deterministic) SQP

Stochastic SQP

Extensions

Matrix-Free Algorithm

Conclusion

Outline

Motivation

SG and SQP

Adaptive (Deterministic) SQP

Stochastic SQP

Extensions

Matrix-Free Algorithm

Conclusion

Constrained optimization (deterministic)

Consider

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

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
- ... (distributionally) robust optimization
- ... chance-constrained optimization

Constrained optimization (stochastic objective)

Consider

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

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

- ω has probability space $(\Omega, \mathcal{F}, P)$
- $\mathbb{E}[\cdot]$ with respect to P
- Classical applications under uncertainty, constrained DNN training, etc.
- Besides cases involving a deterministic equivalent...
- ... very few algorithms so far (mostly penalty methods)

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.[†]

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

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.
- Only enforce convergence in expectation.

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

† Alternatively, see Na, Anitescu, Kolar (2021, 2022)

This talk

Consider *equality constrained* stochastic optimization:

$$\min_{x \in \mathbb{R}^n} \; f(x) \equiv \mathbb{E}[F(x,\omega)]$$
$$\text{s.t. } c_{\mathcal{E}}(x) = 0$$

- *Adaptive* SQP method for deterministic setting
- *Stochastic* SQP method for stochastic setting
- Convergence in expection (comparable to SG for unconstrained setting)
- Worst-case complexity on par with stochastic subgradient method
- Numerical experiments are very promising
- Various open questions!

Outline

Motivation

SG and SQP

Adaptive (Deterministic) SQP

Stochastic SQP

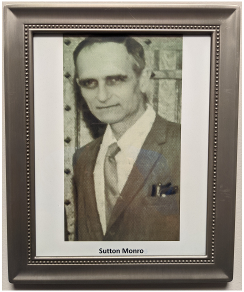
Extensions

Matrix-Free Algorithm

Conclusion

Stochastic gradient method (SG)

Invented by Herbert Robbins and Sutton Monro (1951)



Sutton Monro, former Lehigh faculty member

Stochastic gradient (*not* descent)

Consider the stochastic optimization problem

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

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

Algorithm SG : Stochastic Gradient

1: choose an initial point $x_0 \in \mathbb{R}^n$ and step sizes $\{\alpha_k\} > 0$
2: **for** $k \in \{0, 1, 2, \ldots\}$ **do**
3: set $x_{k+1} \leftarrow x_k - \alpha_k g_k$, where $\mathbb{E}_k[g_k] = \nabla f(x_k)$ and $\mathbb{E}_k[\|g_k - \nabla f(x_k)\|_2^2] \leq M$
4: **end for**

Stochastic gradient (*not* descent)

Consider the stochastic optimization problem

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

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

Algorithm SG : Stochastic Gradient

1: choose an initial point $x_0 \in \mathbb{R}^n$ and step sizes $\{\alpha_k\} > 0$
2: **for** $k \in \{0, 1, 2, \dots\}$ **do**
3: set $x_{k+1} \leftarrow x_k - \alpha_k g_k$, where $\mathbb{E}_k[g_k] = \nabla f(x_k)$ and $\mathbb{E}_k[\|g_k - \nabla f(x_k)\|_2^2] \leq M$
4: **end for**

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

$$\begin{aligned} f(x_{k+1}) - f(x_k) &\leq \nabla f(x_k)^T (x_{k+1} - x_k) + \tfrac{1}{2} L \|x_{k+1} - x_k\|_2^2 \\ &= -\alpha_k \nabla f(x_k)^T g_k + \tfrac{1}{2} \alpha_k^2 L \|g_k\|_2^2 \\ \Longrightarrow \mathbb{E}_k[f(x_{k+1})] - f(x_k) &\leq -\alpha_k \|\nabla f(x_k)\|_2^2 + \tfrac{1}{2} \alpha_k^2 L \mathbb{E}_k[\|g_k\|_2^2]. \end{aligned}$$

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

SG theory

Theorem SG

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

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

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

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

SG illustration

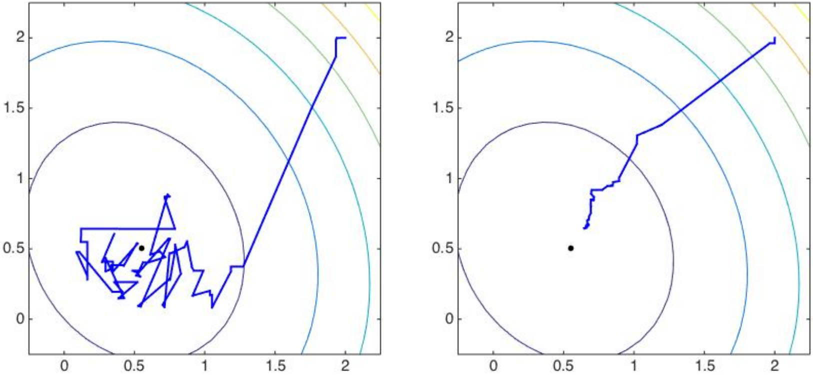


Figure: SG with fixed step size (left) vs. diminishing step sizes (right)

Sequential quadratic optimization (SQP)

Consider

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

with $J \equiv \nabla c$ and $H \succ 0$ (for simplicity), two viewpoints:

$$\begin{bmatrix} \nabla f(x) + J(x)^T y \\ c(x) \end{bmatrix} = 0$$

or

$$\min_{d\in\mathbb{R}^n} f(x) + \nabla f(x)^T d + \tfrac{1}{2} d^T H d \quad \text{s.t. } c(x) + J(x)d = 0$$

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

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

SQP illustration

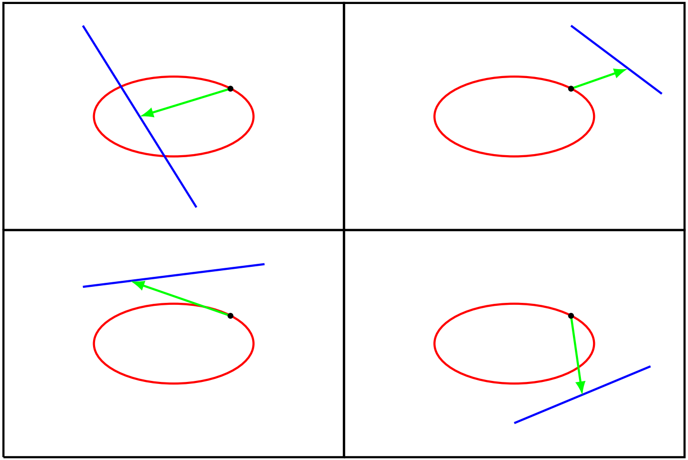


Figure: Illustrations of SQP subproblem solutions

SQP

- Algorithm guided by merit function, with adaptive parameter τ, defined by

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

  a model of which is defined as

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

- For a given $d \in \mathbb{R}^n$ satisfying $c(x) + J(x)d = 0$, the reduction in this model is

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

  and it is easily shown that

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

SQP with backtracking line search

Algorithm SQP-B

1: choose $x_0 \in \mathbb{R}^n$, $\tau_{-1} \in \mathbb{R}_{>0}$, $\sigma \in (0,1)$, $\eta \in (0,1)$

2: **for** $k \in \{0, 1, 2, \dots\}$ **do**

3: 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: set τ_k to ensure $\Delta q(x_k, \tau_k, \nabla f(x_k), d_k) \gg 0$, offered by

$$\tau_k \leq \frac{(1-\sigma)\|c_k\|_1}{\nabla f(x_k)^T d_k + d_k^T H_k d_k} \quad \text{if} \quad \nabla f(x_k)^T d_k + d_k^T H_k d_k > 0$$

5: 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 SQP-B

- $\{\alpha_k\} \geq \alpha_{\min}$ *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*

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

Outline

Motivation

SG and SQP

Adaptive (Deterministic) SQP

Stochastic SQP

Extensions

Matrix-Free Algorithm

Conclusion

Toward stochastic SQP

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

  step sizes determined by Lipschitz constant estimates

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

Primary challenge: Nonsmoothness

In SQP-B, step size is chosen based on reducing the merit function.

The merit function is nonsmooth! An upper bound is

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

where L_k and Γ_k are Lipschitz constant estimates for f and $\|c\|_1$ at x_k

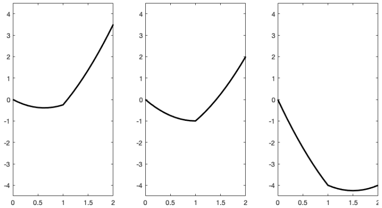


Figure: Three cases for upper bound of ϕ

Idea: Choose α_k to ensure sufficient decrease using this bound

SQP with adaptive step sizes

Algorithm SQP-A

1: choose $x_0 \in \mathbb{R}^n$, $\tau_{-1} \in \mathbb{R}_{>0}$, $\sigma \in (0,1)$, $\eta \in (0,1)$
2: **for** $k \in \{0,1,2,\dots\}$ **do**
3: 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: set τ_k to ensure $\Delta q(x_k, \tau_k, \nabla f(x_k), d_k) \gg 0$, offered by

$$\tau_k \leq \frac{(1-\sigma)\|c_k\|_1}{\nabla f(x_k)^T d_k + d_k^T H_k d_k} \quad \text{if} \quad \nabla f(x_k)^T d_k + d_k^T H_k d_k > 0$$

5: set

$$\widehat{\alpha}_k \leftarrow \frac{2(1-\eta)\Delta q(x_k, \tau_k, \nabla f(x_k), d_k)}{(\tau_k L_k + \Gamma_k)\|d_k\|_2^2} \quad \text{and}$$

$$\widetilde{\alpha}_k \leftarrow \widehat{\alpha}_k - \frac{4\|c_k\|_1}{(\tau_k L_k + \Gamma_k)\|d_k\|_2^2}$$

6: set

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

7: set $x_{k+1} \leftarrow x_k + \alpha_k d_k$ and continue or update L_k and/or Γ_k and return to step 5
8: **end for**

Convergence theory

Exactly the same as for SQP-B, except different step size lower bound

- For SQP-A:

$$\alpha_k = \frac{2(1-\eta)\Delta q(x_k, \tau_k, \nabla f(x_k), d_k)}{(\tau_k L_k + \Gamma_k)\|d_k\|_2^2} \geq \frac{2(1-\eta)\kappa_q \tau_{\min}}{(\tau_{-1}\rho L + \rho\Gamma)\kappa_\Psi} > 0$$

- For SQP-B:

$$\alpha_k > \frac{2\nu(1-\eta)\Delta q(x_k, \tau_k, \nabla f(x_k), d_k)}{(\tau_k L + \Gamma)\|d_k\|_2^2} \geq \frac{2\nu(1-\eta)\kappa_q \tau_{\min}}{(\tau_{-1} L + \Gamma)\kappa_\Psi} > 0$$

Outline

Motivation

SG and SQP

Adaptive (Deterministic) SQP

Stochastic SQP

Extensions

Matrix-Free Algorithm

Conclusion

Stochastic setting

Consider the stochastic problem:

$$\min_{x \in \mathbb{R}^n} \; f(x) \equiv \mathbb{E}[F(x,\omega)]$$
$$\text{s.t. } c(x) = 0$$

Let us assume only the following:

Assumption

For all $k \in \mathbb{N}$, one can compute g_k with

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

Search directions computed by:

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

Important: Given x_k, the values (c_k, J_k, H_k) are determined

Stochastic SQP with adaptive step sizes

(For simplicity, assume Lipschitz constants L and Γ are known.)

Algorithm : Stochastic SQP

1: choose $x_0 \in \mathbb{R}^n$, $\tau_{-1} \in \mathbb{R}_{>0}$, $\sigma \in (0,1)$, $\{\beta_k\} \in (0,1]$
2: **for** $k \in \{0,1,2,\dots\}$ **do**
3: 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: set τ_k to ensure $\Delta q(x_k, \tau_k, g_k, d_k) \gg 0$, offered by

$$\tau_k \leq \frac{(1-\sigma)\|c_k\|_1}{g_k^T d_k + d_k^T H_k d_k} \quad \text{if} \quad g_k^T d_k + d_k^T H_k d_k > 0$$

5: set

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

$$\widetilde{\alpha}_k \leftarrow \widehat{\alpha}_k - \frac{4\|c_k\|_1}{(\tau_k L + \Gamma)\|d_k\|_2^2}$$

6: set

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

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

step size control

The sequence $\{\beta_k\}$ allows us to consider, like for SG,

- a fixed step size
- diminishing step sizes (e.g., $\Theta(1/k)$)

step size control

The sequence $\{\beta_k\}$ allows us to consider, like for SG,

- a fixed step size
- diminishing step sizes (e.g., $\Theta(1/k)$)

Unfortunately, additional control on the step size is needed

- too small: insufficient progress
- too large: ruins progress toward feasibility / optimality

We never know when the step size is too small or too large!

step size control

The sequence $\{\beta_k\}$ allows us to consider, like for SG,

- a fixed step size
- diminishing step sizes (e.g., $\Theta(1/k)$)

Unfortunately, additional control on the step size is needed

- too small: insufficient progress
- too large: ruins progress toward feasibility / optimality

We never know when the step size is too small or too large!

Idea: Project $\widehat{\alpha}_k$ and $\widetilde{\alpha}_k$ onto

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

where $\theta \in \mathbb{R}_{>0}$ is a user-defined parameter

Fundamental lemmas

Lemma

For all $k \in \mathbb{N}$, for any realization of g_k, one finds

$$\begin{aligned} &\phi(x_k + \alpha_k d_k, \tau_k) - \phi(x_k, \tau_k) \\ \leq\ &\underbrace{-\alpha_k \Delta q(x_k, \tau_k, \nabla f(x_k), d_k^{\text{true}})}_{\mathcal{O}(\beta_k),\text{ “deterministic”}} + \underbrace{\tfrac{1}{2}\alpha_k \beta_k \Delta q(x_k, \tau_k, g_k, d_k)}_{\mathcal{O}(\beta_k^2),\text{ stochastic/noise}} + \underbrace{\alpha_k \tau_k \nabla f(x_k)^T (d_k - d_k^{\text{true}})}_{\text{due to adaptive } \alpha_k} \end{aligned}$$

Fundamental lemmas

Lemma

For all $k \in \mathbb{N}$, for any realization of g_k, one finds

$$\begin{aligned} &\phi(x_k + \alpha_k d_k, \tau_k) - \phi(x_k, \tau_k) \\ \leq\ &\underbrace{-\alpha_k \Delta q(x_k, \tau_k, \nabla f(x_k), d_k^{\text{true}})}_{\mathcal{O}(\beta_k),\ \text{“deterministic”}} + \underbrace{\tfrac{1}{2}\alpha_k \beta_k \Delta q(x_k, \tau_k, g_k, d_k)}_{\mathcal{O}(\beta_k^2),\ \text{stochastic/noise}} + \underbrace{\alpha_k \tau_k \nabla f(x_k)^T (d_k - d_k^{\text{true}})}_{\text{due to adaptive } \alpha_k} \end{aligned}$$

Lemma

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

$$\mathbb{E}_k[d_k] = d_k^{\text{true}}, \quad \mathbb{E}_k[y_k] = y_k^{\text{true}}, \quad \textit{and} \quad \mathbb{E}_k[\|d_k - d_k^{\text{true}}\|_2] = \mathcal{O}(\sqrt{M})$$

as well as

$$\nabla f(x_k)^T d_k^{\text{true}} \geq \mathbb{E}_k[g_k^T d_k] \geq (\nabla f(x_k)^T d_k)^{\text{true}} - \zeta^{-1} M \quad \textit{and}$$
$$\mathbb{E}_k[d_k^T H_k d_k] \geq d_k^{\text{true}\,T} H_k d_k^{\text{true}}$$

Good merit parameter behavior

Lemma

If $\{\tau_k\}$ eventually remains fixed at sufficiently small $\tau_{\min} > 0$, then for large k

$$\mathbb{E}_k[\alpha_k \tau_k \nabla f(x_k)^T (d_k - d_k^{\text{true}})] = \beta_k^2 \tau_{\min} \mathcal{O}(\sqrt{M})$$

Theorem

If $\{\tau_k\}$ eventually remains fixed at sufficiently small $\tau_{\min} > 0$, then for large k

$$\beta_k = \Theta(1) \implies \mathbb{E}\left[\frac{1}{k}\sum_{j=1}^{k} \Delta q(x_j, \tau_{\min}, \nabla f(x_j), d_j^{\text{true}})\right] \leq \mathcal{O}(M)$$

$$\beta_k = \Theta\left(\frac{1}{k}\right) \implies \mathbb{E}\left[\frac{1}{\left(\sum_{j=1}^{k} \beta_j\right)} \sum_{j=1}^{k} \beta_j \Delta q(x_j, \tau_{\min}, \nabla f(x_j), d_j^{\text{true}})\right] \to 0$$

Good merit parameter behavior

Lemma

If $\{\tau_k\}$ eventually remains fixed at sufficiently small $\tau_{\min} > 0$, then for large k

$$\mathbb{E}_k[\alpha_k \tau_k \nabla f(x_k)^T (d_k - d_k^{\text{true}})] = \beta_k^2 \tau_{\min} \mathcal{O}(\sqrt{M})$$

Theorem

If $\{\tau_k\}$ eventually remains fixed at sufficiently small $\tau_{\min} > 0$, then for large k

$$\beta_k = \Theta(1) \implies \mathbb{E}\left[\frac{1}{k}\sum_{j=1}^{k}(\|g_j + J_j^T y_j^{\text{true}}\|_2 + \|c_j\|_2)\right] \leq \mathcal{O}(M)$$

$$\beta_k = \Theta\left(\frac{1}{k}\right) \implies \mathbb{E}\left[\frac{1}{\left(\sum_{j=1}^{k}\beta_j\right)}\sum_{j=1}^{k}\beta_j(\|g_j + J_j^T y_j^{\text{true}}\|_2 + \|c_j\|_2)\right] \to 0$$

Poor merit parameter behavior

$\{\tau_k\} \searrow 0$:

- cannot occur if $\|g_k - \nabla f(x_k)\|_2$ is bounded uniformly
- occurs with small probability if distribution of g_k has *fast* decay

Poor merit parameter behavior

$\{\tau_k\} \searrow 0$:

- cannot occur if $\|g_k - \nabla f(x_k)\|_2$ is bounded uniformly
- occurs with small probability if distribution of g_k has *fast* decay

$\{\tau_k\}$ remains too large:

- if there exists $p \in (0,1]$ such that, for all k in infinite $\mathcal{K}$,

$$\mathbb{P}_k\left[g_k^T d_k + \max\{d_k^T H_k d_k, 0\} \geq \nabla f(x_k)^T d_k^{\text{true}} + \max\{(d_k^{\text{true}})^T H_k d_k^{\text{true}}, 0\}\right] \geq p$$

then occurs with probability zero

Numerical results

Matlab software: 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 τ

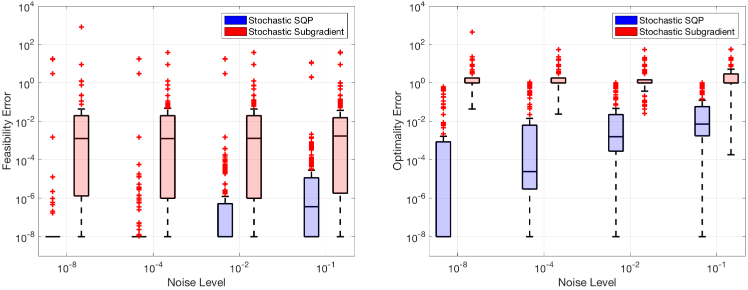


Figure: Box plots for feasibility errors (left) and optimality errors (right).

Outline

Motivation

SG and SQP

Adaptive (Deterministic) SQP

Stochastic SQP

Extensions

Matrix-Free Algorithm

Conclusion

Complexity of deterministic algorithm

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

Lemma 7

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 8

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

$$\left(\frac{\tau_{-1}(f_0 - f_{\inf}) + \|c_0\|_1}{\min\{\kappa_1, \kappa_2 \tau_{\min}\}}\right)\epsilon^{-2}.$$

Worst-case iteration complexity of $\widetilde{\mathcal{O}}(\epsilon^{-4})$

Theorem 9

Suppose the algorithm is run

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

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

$$\mathbb{E}[\|g_{k_*} + J_{k_*}^T y_{k_*}^{\text{true}}\|_2^2 + \|c_{k_*}\|_1] \leq \frac{\tau_{-1}(f_0 - f_{\inf}) + \|c_0\|_1 + M}{\sqrt{k_{\max}+1}} + \frac{(\tau_{-1} - \tau_{\min})(s_{\max}\log(k_{\max}) + \log(1/\delta))}{\sqrt{k_{\max}+1}}$$

Theorem 10

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

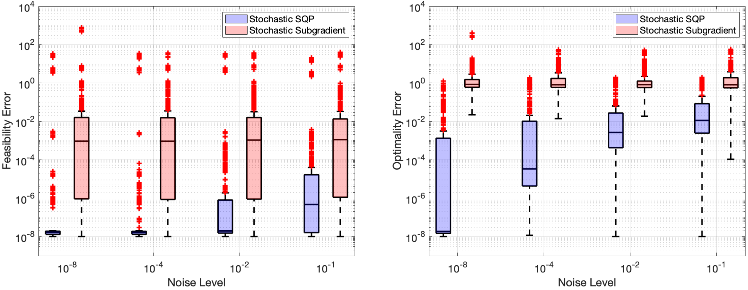


Figure: Box plots for feasibility errors (left) and optimality errors (right).

Outline

Motivation

SG and SQP

Adaptive (Deterministic) SQP

Stochastic SQP

Extensions

Matrix-Free Algorithm

Conclusion

Motivation

Solving for the search directions can be expensive:

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

To avoid direct+exact solves,

- aim to use iterative solver(s) and
- allow inexactness in the subproblem solves.

Algorithm now involves both *stochasticity* and *inexactness*.

Normal search direction computation

As is standard, compute normal search direction by approximately solving

$$\min_{v \in \mathrm{Range}(J_k^T)} \tfrac{1}{2}\|c_k + J_k v\|_2^2$$

e.g., by the conjugate gradient (CG) method, satisfying at least

$$\|c_k\|_2 - \|c_k + J_k v_k\|_2 \geq \epsilon_c(\|c_k\|_2 - \|c_k + J_k v_k^C\|_2),$$

i.e., Cauchy decrease.

Tangential search direction computation

Exact tangential direction using the true gradient:

$$\begin{bmatrix} H_k & J_k^T \\ J_k & 0 \end{bmatrix} \begin{bmatrix} u_k^{\text{true}} \\ \delta_k^{\text{true}} \end{bmatrix} = - \begin{bmatrix} \nabla f_k + H_k v_k + J_k^T y_k \\ 0 \end{bmatrix}$$

Exact tangential direction using the stochastic gradient:

$$\begin{bmatrix} H_k & J_k^T \\ J_k & 0 \end{bmatrix} \begin{bmatrix} u_{k,*} \\ \delta_{k,*} \end{bmatrix} = - \begin{bmatrix} g_k + H_k v_k + J_k^T y_k \\ 0 \end{bmatrix}$$

Tangential direction actually computed, with corresponding residual:

$$\begin{bmatrix} \rho_k \\ r_k \end{bmatrix} := \begin{bmatrix} H_k & J_k^T \\ J_k & 0 \end{bmatrix} \begin{bmatrix} u_k \\ \delta_k \end{bmatrix} + \begin{bmatrix} g_k + H_k v_k + J_k^T y_k \\ 0 \end{bmatrix}$$

Termination tests

Termination test 1 (simplified):

$$\Delta l(x_k, \tau_k, g_k, d_k) \geq \sigma_u \tau \epsilon \|u_k\|_2^2 + \sigma_c(\|c_k\|_2 - \|c_k + J_k v_k\|_2)$$

$$\|\rho_k\|_2 \leq \kappa \left\| \begin{bmatrix} g_k + J_k^T(y_k + \delta_k) \\ c_k \end{bmatrix} \right\|$$

$$\|\rho_k\|_2 \leq \kappa\beta_k \quad \text{and} \quad \|r_k\|_2 \leq \kappa\beta_k$$

Termination test 2 (simplified):

same *residual* conditions as above and

$$\|c_k\|_2 - \|c_k + J_k v_k + r_k\|_2 \geq \epsilon(\|c_k\|_2 - \|c_k + J_k v_k\|_2) > 0$$

Results on CUTEst problems

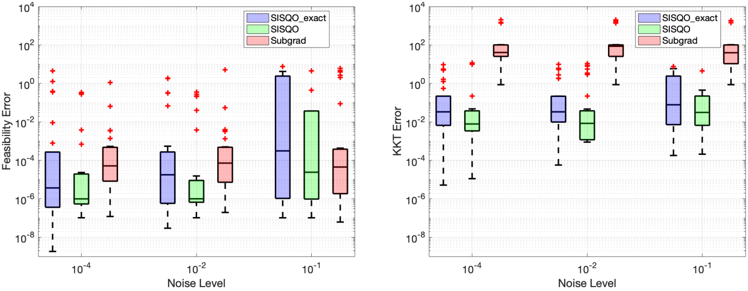


Figure: Box plots for feasibility errors (left) and optimality errors (right).

Optimal control problems

Given domain $\Xi \in \mathbb{R}^2$, constant $N \in \mathbb{N}_{>0}$, reference functions $\overline{w}_{ij} \in L^2(\Xi)$ and $\bar{z}_{ij} \in L^2(\Xi)$ for $(i,j) \in \{1,\dots,N\}^2$, and regularization $\lambda \in \mathbb{R}_{>0}$, consider

$$\begin{aligned} &\min_{w,z} \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \left(\tfrac{1}{2} \|w - \overline{w}_{ij}\|^2_{L^2(\Xi)} + \tfrac{\lambda}{2} \|z - \bar{z}_{ij}\|^2_{L^2(\Xi)} \right) \\ &\text{s.t. } -\Delta w = z \text{ in } \Xi \text{ and } w = 0 \text{ on } \partial\Xi, \end{aligned} \tag{1}$$

and, with the same notation but $\bar{z}_{ij} \in L^2(\partial\Xi)$, consider

$$\begin{aligned} &\min_{w,z} \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \left(\tfrac{1}{2} \|w - \overline{w}_{ij}\|^2_{L^2(\Xi)} + \tfrac{\lambda}{2} \|z - \bar{z}_{ij}\|^2_{L^2(\partial\Xi)} \right) \\ &\text{s.t. } -\Delta w + w = 0 \text{ in } \Xi \text{ and } \tfrac{\partial w}{\partial p} = z \text{ on } \partial\Xi, \end{aligned} \tag{2}$$

where p represents the unit outer normal to Ξ along $\partial\Xi$.

For all $(i,j) \in \{1,\dots,N\}^2$ for some $(\epsilon_S, \epsilon_N) \in \mathbb{R}^2_{>0}$, we chose $\bar{z}_{ij} = 0$ and

$$\overline{w}_{ij}(x_1,x_2) = \sin((4 + \tfrac{\epsilon_N}{\epsilon_S}(i - \tfrac{N+1}{2}))x_1) + \cos((3 + \tfrac{\epsilon_N}{\epsilon_S}(j - \tfrac{N+1}{2}))x_2).$$

Also, $N = 3$, $\lambda = 10^{-5}$, $\epsilon_S = 50$, and $\epsilon_N \in \{10^{-4}, 10^{-2}, 10^{-1}\}$.

Numerical results

		problem (1)			problem (2)		
strategy	ϵ_N	feasibility error	KKT error	C+M iter. (iter.)	feasibility error	KKT error	C+M iter. (iter.)
SISQO	10^{-4}	6.30×10^{-7}	2.08×10^{-6}	61225.8 (6)	7.96×10^{-7}	7.72×10^{-6}	96684.4 (9)
SISQO_exact	10^{-4}	5.90×10^{-7}	1.76×10^{0}	61225.8 (6)	3.91×10^{-6}	8.29×10^{-1}	96684.4 (8)
Subgrad	10^{-4}	4.98×10^{1}	4.98×10^{1}	0 (61225.8)	$1.00 \times 10^{+2}$	$1.00 \times 10^{+2}$	0 (96684.4)
SISQO	10^{-2}	6.37×10^{-7}	2.10×10^{-4}	60113 (6)	7.80×10^{-7}	1.86×10^{-4}	96103.4 (9)
SISQO_exact	10^{-2}	5.82×10^{-7}	1.76×10^{0}	60113 (6)	1.44×10^{-6}	8.29×10^{-1}	96103.4 (8.8)
Subgrad	10^{-2}	4.98×10^{1}	4.98×10^{1}	0 (60113)	$1.00 \times 10^{+2}$	$1.00 \times 10^{+2}$	0 (96103.4)
SISQO	10^{-1}	6.81×10^{-7}	2.09×10^{-3}	58901.2 (6)	8.12×10^{-7}	1.68×10^{-3}	96914.6 (9.2)
SISQO_exact	10^{-1}	5.85×10^{-7}	1.76×10^{0}	58901.2 (6)	1.33×10^{-6}	8.29×10^{-1}	96914.6 (8.8)
Subgrad	10^{-1}	4.98×10^{1}	4.98×10^{1}	0 (58901.2)	$1.00 \times 10^{+2}$	$1.00 \times 10^{+2}$	0 (96914.6)

Table: Numerical results for problems (1) and (2) averaged over ten independent runs.

Outline

Motivation

SG and SQP

Adaptive (Deterministic) SQP

Stochastic SQP

Extensions

Matrix-Free Algorithm

Conclusion

Summary

Consider *equality constrained* stochastic optimization:

$$\min_{x\in\mathbb{R}^n} f(x) \equiv \mathbb{E}[F(x,\omega)]$$
$$\text{s.t. } c_{\mathcal{E}}(x) = 0$$

- *Adaptive* SQP method for deterministic setting
- *Stochastic* SQP method for stochastic setting
- Convergence in expection (comparable to SG for unconstrained setting)
- Worst-case complexity on par with stochastic subgradient method
- Numerical experiments are very promising
- Various extensions (on-going)

Current work: Inequality constraints

Inequality constraints

- SQP
- interior-point

Main challenge: For *equality* constraints only, subproblem solution on linearized constraints remains unbiased:

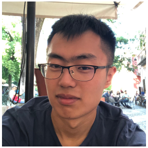
$$c_k + J_k \bar{d}_k = 0 \iff \bar{d}_k = v_k + \bar{u}_k$$

$$\text{with } v_k \in \text{Range}(J_k^T) \text{ and } \bar{u}_k \in \text{Null}(J_k)$$

$$\text{has } E_k[\bar{u}_k] = u_k.$$

However, when *inequalities* are present, subproblem solution is biased.

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.

ICCOPT 2022

International Conference on Continuous Optimization

July 23–28, 2022