SQP for Equality Constrained Stochastic Optimization

Frank E. Curtis, Lehigh University

joint work with

Albert Berahas, University of Michigan
Daniel P. Robinson, Lehigh University
Baoyu Zhou, Lehigh University

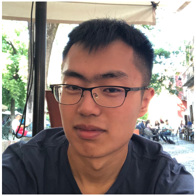
presented at

INFORMS Annual Meeting

November 13, 2020

References





- "Sequential Quadratic Optimization for Nonlinear Equality Constrained Stochastic Optimization" https://arxiv.org/abs/2007.10525.

Outline

Outline

Constrained stochastic optimization

Consider

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

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 with objective uncertainty, *constrained* DNNs, etc.
- Very few algorithms so far (mostly penalty methods)

Contributions

Consider *equality constrained* stochastic optimization:

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

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

Outline

Stochastic gradient (*not* descent)

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

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

Algorithm invented by Herbert Robbins and Sutton Monro (1951):

Algorithm SG : Stochastic Gradient

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

Stochastic gradient (*not* descent)

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

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

Algorithm invented by Herbert Robbins and Sutton Monro (1951):

Algorithm SG : Stochastic Gradient

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

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

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

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

SG theory

Theorem SG

If $\mathbb{E}_k[\|g_k - \nabla f(x_k)\|_2^2] \leq M$, *then:*

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

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

SG illustration

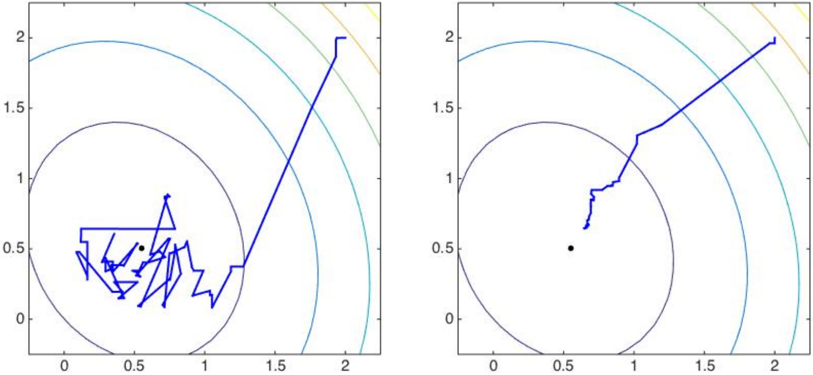


Figure: SG with fixed stepsize (left) vs. diminishing stepsizes (right)

Sequential quadratic optimization (SQP)

Consider

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

with $g \equiv \nabla f$, $J \equiv \nabla c$, and H (positive definite on Null(J)), two viewpoints:

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

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

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

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

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

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

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

and it is easily shown that

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

SQP with backtracking line search

Algorithm SQP-B

choose $x_0 \in \mathbb{R}^n$, $\tau_{-1} \in \mathbb{R}_{>0}$, $\sigma \in (0,1)$, $\eta \in (0,1)$
for $k \in \{0, 1, 2, \dots\}$ **do**
  Compute step: solve

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

  Update parameter: set τ_k to ensure $\Delta q(x_k, \tau_k, d_k) \gg 0$, offered by

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

  Line search: backtracking line search to ensure $x_{k+1} \leftarrow x_k + \alpha_k d_k$ yields

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

end for

Convergence theory

Assumption

- *f, c, g, 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* $u \in \mathrm{Null}(J_k)$ *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* $\tau_{\min} > 0$
- $\Delta q(x_k, \tau_k, d_k) \to 0$ *implies*

$$\|d_k\|_2 \to 0, \quad \|c_k\|_2 \to 0, \quad \|g_k + J_k^T y_k\|_2 \to 0$$

Outline

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

  stepsizes determined by Lipschitz constant estimates

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

Primary challenge: Nonsmoothness

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

Primary challenge: Nonsmoothness

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

The merit function is nonsmooth! An upper bound is

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

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

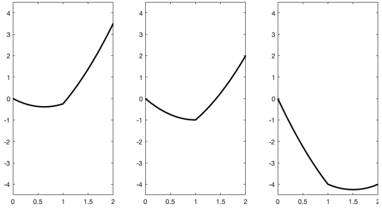


Figure: Three cases for upper bound of ϕ

SQP with adaptive stepsizes

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** $k \in \{0, 1, 2, \ldots\}$ **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 parameter: set τ_k to ensure $\Delta q(x_k, \tau_k, d_k) \gg 0$, offered by

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

5: Compute stepsize: set

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

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

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**

Approximately the same theory and similar empirical performance as SQP-B

Outline

Stochastic setting

Consider the stochastic problem:

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

Let us assume only the following:

Assumption

For all $k \in \mathbb{N}$, one can compute $\bar{g}_k$ with

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

Search directions computed by:

$$\begin{bmatrix} H_k & J_k^T \\ J_k & 0 \end{bmatrix} \begin{bmatrix} \bar{d}_k \\ \bar{y}_k \end{bmatrix} = - \begin{bmatrix} \bar{g}_k \\ c_k \end{bmatrix}$$

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

Stochastic SQP with adaptive stepsizes

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

Algorithm : Stochastic SQP

1: choose $x_0 \in \mathbb{R}^n$, $\bar{\tau}_{-1} \in \mathbb{R}_{>0}$, $\sigma \in (0,1)$, $\{\beta_k\} \in (0,1]$
2: **for** $k \in \{0,1,2,\dots\}$ **do**
3: Compute step: solve

$$\begin{bmatrix} H_k & J_k^T \\ J_k & 0 \end{bmatrix} \begin{bmatrix} \bar{d}_k \\ \bar{y}_k \end{bmatrix} = - \begin{bmatrix} \bar{g}_k \\ c_k \end{bmatrix}$$

4: Update parameter: set $\bar{\tau}_k$ to ensure $\Delta \bar{q}(x_k, \bar{\tau}_k, \bar{d}_k) \gg 0$, offered by

$$\bar{\tau}_k \leq \frac{(1-\sigma)\|c_k\|_1}{\bar{g}_k^T \bar{d}_k + \max\{\bar{d}_k^T H_k \bar{d}_k, 0\}} \quad \text{if} \quad \bar{g}_k^T \bar{d}_k + \max\{\bar{d}_k^T H_k \bar{d}_k, 0\} > 0$$

5: Compute stepsize: set

$$\bar{\hat{\alpha}}_k \leftarrow \frac{\beta_k \Delta \bar{q}(x_k, \bar{\tau}_k, \bar{d}_k)}{(\bar{\tau}_k L + \Gamma)\|\bar{d}_k\|_2^2} \quad \text{and}$$

$$\bar{\tilde{\alpha}}_k \leftarrow \bar{\hat{\alpha}}_k - \frac{4\|c_k\|_1}{(\bar{\tau}_k L + \Gamma)\|\bar{d}_k\|_2^2}$$

6: set

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

7: set $x_{k+1} \leftarrow x_k + \bar{\alpha}_k \bar{d}_k$
8: **end for**

Stepsize control

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

- a fixed stepsize
- diminishing stepsizes (e.g., $\mathcal{O}(1/k)$)

Unfortunately, additional control on the stepsize is needed

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

We never know when the stepsize is too small or too large!

Stepsize control

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

- a fixed stepsize
- diminishing stepsizes (e.g., $\mathcal{O}(1/k)$)

Unfortunately, additional control on the stepsize is needed

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

We never know when the stepsize is too small or too large!

Idea: Project $\hat{\bar{\alpha}}_k$ and $\tilde{\bar{\alpha}}_k$ onto

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

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

Fundamental lemma

Lemma

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

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

Good merit parameter behavior

Lemma

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

$$\mathbb{E}_k[\bar{\alpha}_k \bar{\tau}_k g_k^T(\bar{d}_k - d_k)] = \beta_k^2 \tau_{\min} \mathcal{O}(\sqrt{M})$$

Theorem

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

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

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

Poor merit parameter behavior

$\{\bar{\tau}_k\} \searrow 0$:

- cannot occur if $\|\bar{g}_k - g_k\|_2$ is bounded uniformly
- occurs with small probability if distribution of $\bar{g}_k$ has *fast* decay(?)

$\{\bar{\tau}_k\}$ remains too large:

- can only occur if realization of $\{\bar{g}_k\}$ is *one-sided for all k*
- if there exists $p \in (0,1]$ such that, for all k in infinite $\mathcal{K}$,

$$\mathbb{P}_k\left[\bar{g}_k^T\bar{d}_k + \max\{\bar{d}_k^T H_k \bar{d}_k, 0\} \geq g_k^T d_k + \max\{d_k^T H_k d_k, 0\}\right] \geq p$$

  then occurs with probability zero

Neither occurred in our experiments

Numerical results

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

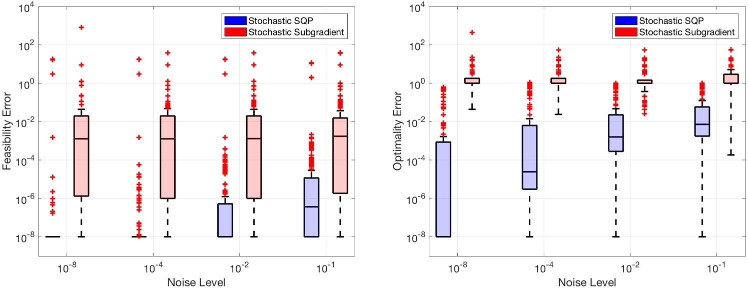


Figure: Box plots for feasibility errors (left) and optimality errors (right).

Outline

Summary

Consider *equality constrained* stochastic optimization:

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

- *Adaptive* SQP method for deterministic setting
- *Stochastic* SQP method for stochastic setting
- Convergence in expection (comparable to SG for unconstrained setting)
- Numerical experiments are *very promising*