SQP Methods for 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

Argonne National Laboratory

July 22, 2020
References

Outline

Motivation

SG and SQP

Adaptive (Deterministic) SQP

Stochastic SQP

Conclusion
Outline

Motivation

SG and SQP

Adaptive (Deterministic) SQP

Stochastic SQP

Conclusion
Constrained optimization (deterministic)

Consider

\[
\min_{x \in \mathbb{R}^n} f(x)
\quad \text{s.t. } c_E(x) = 0
\quad c_I(x) \leq 0
\]

where \( f : \mathbb{R}^n \to \mathbb{R} \), \( c_E : \mathbb{R}^n \to \mathbb{R}^{m_E} \), and \( c_I : \mathbb{R}^n \to \mathbb{R}^{m_I} \)

- Physically-constrained, resource-constrained, PDE-constrained, etc.
- Long history of algorithms (penalty, SQP, interior-point)
- Strong theory (even with lack of constraint qualifications)
- Effective software (Ipopt, Knitro, LOQO, etc.)
Constrained optimization (stochastic constraints)

Consider

\[
\min_{x \in \mathbb{R}^n} f(x) \\
\text{s.t. } c_E(x) = 0 \\
c_I(x, \omega) \preceq 0
\]

where \( f : \mathbb{R}^n \to \mathbb{R} \), \( c_E : \mathbb{R}^n \to \mathbb{R}^{m_E} \), and \( c_I : \mathbb{R}^n \times \Omega \to \mathbb{R}^{m_I} \)

- Various modeling paradigms:
  - “Stochastic optimization”
  - “(Distributionally) robust optimization”
  - “Chance-constrained optimization”
Constrained optimization (stochastic objective)

Consider

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

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

- \( \omega \) 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:

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

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

Motivation

SG and SQP

Adaptive (Deterministic) SQP

Stochastic SQP

Conclusion
Gradient descent

\[
\min_{x \in \mathbb{R}^n} f(x)
\]

where \( \nabla f : \mathbb{R}^n \rightarrow \mathbb{R}^n \) is Lipschitz continuous with constant \( L \)

**Algorithm GD : Gradient Descent**

1: choose an initial point \( x_0 \in \mathbb{R}^n \) and stepsize \( \alpha > 0 \)
2: for \( k \in \{0, 1, 2, \ldots \} \) do
3: set \( x_{k+1} \leftarrow x_k - \alpha \nabla f(x_k) \)
4: end for
Gradient descent

\[
\min_{x \in \mathbb{R}^n} f(x)
\]

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

**Algorithm GD** : Gradient Descent

1. choose an initial point \( x_0 \in \mathbb{R}^n \) and stepsize \( \alpha > 0 \)
2. \textbf{for} \( k \in \{0, 1, 2, \ldots \} \) \textbf{do}
3. \hspace{1em} set \( x_{k+1} \leftarrow x_k - \alpha \nabla f(x_k) \)
4. \textbf{end for}
Gradient descent

\[
\min_{x \in \mathbb{R}^n} f(x)
\]

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

**Algorithm GD** : Gradient Descent

1. choose an initial point \( x_0 \in \mathbb{R}^n \) and stepsize \( \alpha > 0 \)
2. for \( k \in \{0, 1, 2, \ldots\} \) do
3. set \( x_{k+1} \leftarrow x_k - \alpha \nabla f(x_k) \)
4. end for
**GD theory**

**Theorem GD**

If $\alpha \in (0, 2/L)$, then $\sum_{k=0}^{\infty} \|\nabla f(x_k)\|^2 < \infty$, which implies $\{\nabla f(x_k)\} \to 0$.

**Proof.**

\[
f(x_{k+1}) - f(x_k) \leq \nabla f(x_k)^T(x_{k+1} - x_k) + \frac{1}{2} L \|x_{k+1} - x_k\|^2
\]

\[
= - \alpha \|\nabla f(x_k)\|^2 + \frac{1}{2} L \alpha^2 \|\nabla f(x_k)\|^2
\]

\[
\leq - \frac{1}{2} \alpha \|\nabla f(x_k)\|^2
\]
GD illustration

Figure: GD with fixed stepsize
Stochastic gradient method (SG)

Invented by Herbert Robbins and Sutton Monro (1951)

Sutton Monro, former Lehigh faculty member
Stochastic gradient (*not* descent)

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

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 stepsizes \(\{\alpha_k\} > 0\)
2: for \(k \in \{0, 1, 2, \ldots\}\) do
3: \(\text{set } x_{k+1} \leftarrow x_k - \alpha_k g_k\), where \(\mathbb{E}_k[g_k] = \nabla f(x_k)\)
4: end for
Stochastic gradient (\textit{not} descent)

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

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 stepsizes \(\{\alpha_k\} > 0\)
2: \textbf{for} \(k \in \{0, 1, 2, \ldots\} \) \textbf{do}
3: \hspace{0.5cm} set \(x_{k+1} \leftarrow x_k - \alpha_k g_k\), where \(\mathbb{E}_k [g_k] = \nabla f(x_k)\)
4: \textbf{end for}

Not a descent method! \ldots but \textit{eventual descent in expectation}:

\[
f(x_{k+1}) - f(x_k) \leq \nabla f(x_k)^T (x_{k+1} - x_k) + \frac{1}{2} L \| x_{k+1} - x_k \|^2_2
\]

\[
= -\alpha_k \nabla f(x_k)^T g_k + \frac{1}{2} \alpha_k^2 L \| g_k \|^2_2
\]

\[
\implies \mathbb{E}_k [f(x_{k+1})] - f(x_k) \leq -\alpha_k \| \nabla f(x_k) \|^2_2 + \frac{1}{2} \alpha_k^2 L \mathbb{E}_k [\| g_k \|^2_2].
\]

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:

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

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

Figure: SG with fixed stepsize (left) vs. diminishing stepsizes (right)
Sequential quadratic optimization (SQP)

Consider

\[
\begin{align*}
\min_{x \in \mathbb{R}^n} & \quad f(x) \\
\text{s.t.} & \quad c(x) = 0
\end{align*}
\]

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

\[
\begin{bmatrix}
g(x) + J(x)^T y \\
\hline
c(x)
\end{bmatrix} = 0 \quad \text{or} \quad \begin{align*}
\min_{x \in \mathbb{R}^n} & \quad f(x) + g(x)^T d + \frac{1}{2} d^T H d \\
\text{s.t.} & \quad c(x) + J(x)d = 0
\end{align*}
\]

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 $\tau$, defined by

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

a model of which is defined as

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

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

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

and it is easily shown that

$$\phi'(x, \tau, d) \leq -\Delta q(x, \tau, d)$$
SQP with backtracking line search

Algorithm SQP-B

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

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

6: 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 $u \in \text{Null}(J_k)$ 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 $\tau_{\min} > 0$
- $\Delta q(x_k, \tau_k, d_k) \to 0$ implies
  \[
  \|d_k\|_2 \to 0, \quad \|c_k\|_2 \to 0, \quad \|g_k + J_k^T y_k\|_2 \to 0
  \]
Outline

Motivation

SG and SQP

Adaptive (Deterministic) SQP

Stochastic SQP

Conclusion
Toward stochastic SQP

- In a stochastic setting, line searches are (likely) intractable
- However, for $\nabla f$ and $\nabla c$, 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.

\[ \phi(x^k + \alpha_k d_k, \tau_k) - \phi(x^k, \tau_k) \leq \alpha_k \tau_k g^T d_k + |1 - \alpha_k| \|c_k\|_1 - \|c_k\|_1 + \frac{1}{2} (\tau_k L_k + \Gamma_k) \alpha_k^2 \|d_k\|_2^2 \]

where \(L_k\) and \(\Gamma_k\) are Lipschitz constant estimates for \(f\) and \(\|c\|_1\) at \(x^k\).
Primary challenge: Nonsmoothness

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

The merit function is nonsmooth! An upper bound is

$$\phi(x_k + \alpha_k d_k, \tau_k) - \phi(x_k, \tau_k) \leq \alpha_k \tau_k g_k^T d_k + |1 - \alpha_k| \|c_k\|_1 - \|c_k\|_1 + \frac{1}{2} (\tau_k L_k + \Gamma_k) \alpha_k^2 \|d_k\|^2$$

where $L_k$ and $\Gamma_k$ are Lipschitz constant estimates for $f$ and $\|c\|_1$ at $x_k$
Primary challenge: Nonsmoothness

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

The merit function is nonsmooth! An upper bound is

\[ \phi(x_k + \alpha_k d_k, \tau_k) - \phi(x_k, \tau_k) \]
\[ \leq \alpha_k \tau_k ^T g_k d_k + |1 - \alpha_k| \|c_k \|_1 - \|c_k \|_1 + \|d_k \|_2 \]

where \( L_k \) and \( \Gamma_k \) are Lipschitz constant estimates for \( f \) and \( \|c\|_1 \) at \( x_k \)

Figure: Three cases for upper bound of \( \phi \)

Idea: Choose \( \alpha_k \) to minimize this upper bound
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:    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 $\tau_k$ to ensure $\Delta q(x_k, \tau_k, d_k) \gg 0$, offered by
$\tau_k \leq \frac{(1 - \sigma)\|c_k\|_1}{g_k^T d_k + \max\{d_k^T H_k d_k, 0\}}$ if $g_k^T d_k + \max\{d_k^T H_k d_k, 0\} > 0$
5:    set
$\hat{\alpha}_k \leftarrow \frac{2(1 - \eta)\Delta q(x_k, \tau_k, d_k)}{(\tau_k L_k + \Gamma_k)\|d_k\|_2^2}$
and
$\tilde{\alpha}_k \leftarrow \hat{\alpha}_k - \frac{4\|c_k\|_1}{(\tau_k L_k + \Gamma_k)\|d_k\|_2^2}$
6:    set
$\alpha_k \leftarrow \begin{cases}
\hat{\alpha}_k & \text{if } \hat{\alpha}_k < 1 \\
1 & \text{if } \tilde{\alpha}_k \leq 1 \leq \hat{\alpha}_k \\
\tilde{\alpha}_k & \text{if } \tilde{\alpha}_k > 1
\end{cases}$
7:    set $x_{k+1} \leftarrow x_k + \alpha_k d_k$ and continue or update $L_k$ and/or $\Gamma_k$ and return to step 5
8: end for
Convergence theory

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

- For SQP-A:

\[
\alpha_k = \frac{2(1 - \eta) \Delta q(x_k, \tau_k, d_k)}{(\tau_k L_k + \Gamma_k)\|d_k\|^2_2} \geq \frac{2(1 - \eta) \kappa_q \tau_{\text{min}}}{(\tau - 1 \rho L + \rho \Gamma)\kappa \Psi} > 0
\]

- For SQP-B:

\[
\alpha_k > \frac{2\nu (1 - \eta) \Delta q(x_k, \tau_k, d_k)}{(\tau_k L + \Gamma)\|d_k\|^2_2} \geq \frac{2\nu (1 - \eta) \kappa_q \tau_{\text{min}}}{(\tau - 1 L + \Gamma)\kappa \Psi} > 0
\]
Numerical experiments

**Figure:** Performance profiles for “SQP Adaptive” and “SQP Backtracking” for problems from the CUTE test set in terms of iterations (left) and function evaluations (right).
Outline

Motivation

SG and SQP

Adaptive (Deterministic) SQP

Stochastic SQP

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* \( \bar{g}_k \) *with*

\[
\mathbb{E}_k[\bar{g}_k] = g_k \\
\mathbb{E}_k[\|\bar{g}_k - g_k\|_2^2] \leq M
\]

Search directions computed by:

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

**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 $\Gamma$ 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 $k \in \{0, 1, 2, \ldots\}$ do

3: solve

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

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

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

5: set

\[
\bar{\alpha}_k \leftarrow \beta_k \Delta \bar{q}(x_k, \bar{\tau}_k, \bar{d}_k) \frac{(\bar{\tau}_k L + \Gamma)\|\bar{d}_k\|_2^2}{(\bar{\tau}_k L + \Gamma)\|\bar{d}_k\|_2^2}
\]

and

\[
\bar{\beta}_k \leftarrow \frac{4\|c_k\|_1}{(\bar{\tau}_k L + \Gamma)\|\bar{d}_k\|_2^2}
\]

6: set

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

7: set $x_{k+1} \leftarrow x_k + \tilde{\alpha}_k \bar{d}_k$

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!

Idea: Project $\bar{\alpha}_k$ and $\tilde{\alpha}_k$ onto $\left[\beta_k \bar{\tau}_k \bar{\tau}_k L + \Gamma, \beta_k \bar{\tau}_k \bar{\tau}_k L + \Gamma + \theta \beta_k^2 \right]$ where $\theta \in \mathbb{R} > 0$ is a user-defined parameter
Motivation | SG and SQP | Adaptive (Deterministic) SQP | Stochastic SQP | Conclusion

Stepsizes control

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

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

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., \( O(1/k) \))

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 \( \tilde{\alpha}_k \) and \( \tilde{\alpha}_k \) onto

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

where \( \theta \in \mathbb{R}_{>0} \) is a user-defined parameter
Fundamental lemmas

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

$$
\phi(x_k + \bar{\alpha}_k \bar{d}_k, \bar{\tau}_k) - \phi(x_k, \bar{\tau}_k) \\
\leq -\bar{\alpha}_k \Delta q(x_k, \bar{\tau}_k, d_k) + \frac{1}{2} \bar{\alpha}_k \beta_k \Delta \bar{q}(x_k, \bar{\tau}_k, \bar{d}_k) + \bar{\alpha}_k \bar{\tau}_k g_k^T (\bar{d}_k - d_k)
$$

$O(\beta_k)$, “deterministic”  $O(\beta_k^2)$, stochastic/noise  due to adaptive $\bar{\alpha}_k$
**Fundamental lemmas**

**Lemma**

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

$$\phi(x_k + \alpha_k d_k, \tau_k) - \phi(x_k, \tau_k) \leq -\alpha_k \Delta q(x_k, \tau_k, d_k) + \frac{1}{2} \alpha_k \beta_k \Delta \bar{q}(x_k, \tau_k, d_k) + \alpha_k \tau_k g_k^T (d_k - d_k)$$

$O(\beta_k)$, “deterministic”

$O(\beta_k^2)$, stochastic/noise

due to adaptive $\alpha_k$

**Lemma**

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

$$E_k[d_k] = d_k, \quad E_k[\bar{g}_k] = y_k, \quad \text{and} \quad E_k[\|d_k - d_k\|_2] = O(\sqrt{M})$$

as well as

$$g_k^T d_k \geq E_k[\bar{g}_k^T d_k] \geq g_k^T d_k - \zeta^{-1} M \quad \text{and} \quad d_k^T H_k d_k \leq E_k[d_k^T H_k d_k]$$
Good merit parameter behavior

Lemma

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

\[
E_k[\bar{\alpha}_k \bar{\tau}_k g_k^T (\bar{d}_k - d_k)] = \beta_k^2 \tau_{\min} O(\sqrt{M})
\]
Good merit parameter behavior

**Lemma**

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

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

**Theorem**

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

\[
\beta_k = \mathcal{O}(1) \implies \alpha_k = \frac{\tau_{\text{min}}}{\tau_{\text{min}} L + \Gamma} \implies \mathbb{E} \left[ \frac{1}{k} \sum_{j=1}^{k} \Delta q(x_j, \tau_{\text{min}}, d_j) \right] \leq \mathcal{O}(M)
\]

\[
\beta_k = \mathcal{O} \left( \frac{1}{k} \right) \implies \mathbb{E} \left[ \frac{1}{(\sum_{j=1}^{k} \beta_j)} \sum_{j=1}^{k} \beta_j \Delta q(x_j, \tau_{\text{min}}, d_j) \right] \to 0
\]
Good merit parameter behavior

Lemma

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

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

Theorem

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

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

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

$\{\overline{\tau}_k\} \searrow 0$:
- cannot occur if $\|\overline{g}_k - g_k\|_2$ is bounded uniformly
- occurs with small probability if distribution of $\overline{g}_k$ has fast decay(?)
Poor merit parameter behavior

\{\bar{\tau}_k\} \searrow 0:
- cannot occur if \(\|\overline{g}_k - g_k\|_2\) is bounded uniformly
- occurs with small probability if distribution of \(\bar{g}_k\) has \textit{fast} decay(?)

\{\bar{\tau}_k\} remains too large:
- can only occur if realization of \(\{\bar{g}_k\}\) is \textit{one-sided} for all \(k\)
- if there exists \(p \in (0, 1]\) such that, for all \(k\) in infinite \(\mathcal{K}\),

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

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 $\tau$

![Box plots for feasibility errors (left) and optimality errors (right).](image)

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

Motivation

SG and SQP

Adaptive (Deterministic) SQP

Stochastic SQP

Conclusion
Summary

Consider *equality constrained* stochastic optimization:

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

- *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*
Open questions

- Under what (stronger) assumptions will the merit parameter settle \( \text{w.h.p.} \)?
- Lack of constraint qualifications?
- Inequality constraints?
- Active-set identification?
- Lagrange multiplier computation?
- Inexact SQP for large-scale problems?