Adaptive Stochastic Algorithms for Nonlinearly Constrained Optimization

Frank E. Curtis, Lehigh University

involving joint work with

Albert S. Berahas (U. of Michigan), **Xin Jiang** (Lehigh), **Vyacheslav Kungurtsev** (Czech TU),
Suyun Liu (Amazon), **Michael O'Neill** (UNC Chapel Hill), **Daniel P. Robinson** (Lehigh),
Qi Wang (Lehigh), **Baoyu Zhou** (Chicago Booth)

presented at

SIAM Conference on Optimization (OP23)

June 3, 2023

Outline

Outline

Optimization problem formulations

$$\boxed{\min_{x\in\mathbb{R}^n} \; f(x)}$$

with $f : \mathbb{R}^n \to \mathbb{R}$ where

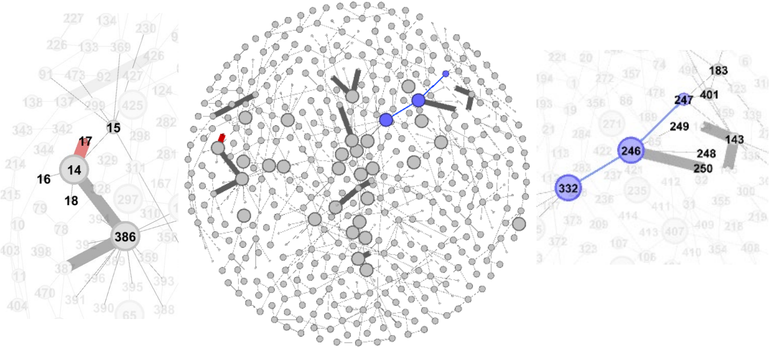
- $f(x) = \mathbb{E}_\omega[F(x,\omega)]$
- ω has probability space $(\Omega, \mathcal{F}_\omega, \mathbb{P}_\omega)$
- $F : \mathbb{R}^n \times \Omega \to \mathbb{R}$
- $\mathbb{E}_\omega[\cdot]$ denotes expectation w.r.t. $\mathbb{P}_\omega$

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

with $f : \mathbb{R}^n \to \mathbb{R}$ and $c : \mathbb{R}^n \to \mathbb{R}^m$ where

- ξ has probability space $(\Xi, \mathcal{F}_\xi, \mathbb{P}_\xi)$ and
- $c(x) = \mathbb{E}_\xi[C(x,\xi)]$ or
- $c(x) = \alpha - \mathbb{P}_\xi[C(x,\xi) \leq 0]$ or
- $c(x) = [C(x,\xi)]_{\xi\in\mathcal{D}}$

Motivation #1: Network optimization



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

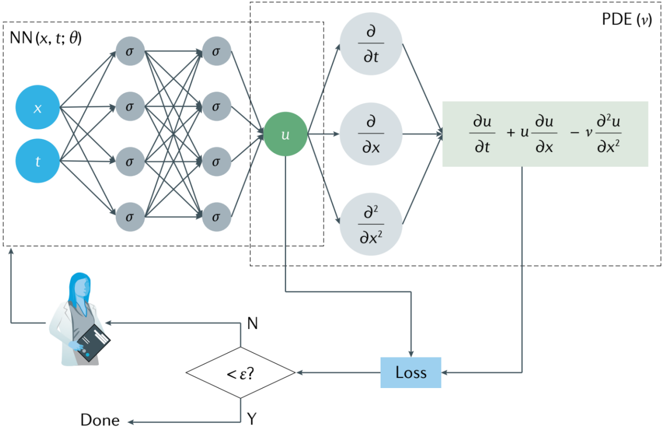


Photo: Karniadakis et al.

Motivation #3: Fair learning

Let

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

and consider

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

This loss might not be fair between subgroups in the population.

- Various criteria related to fairness (e.g., demographic parity, equalized odds, equalized opportunity) leading to various measures (e.g., accuracy equality, disparate impact, measures conditioned on outcome, measures conditioned on prediction)
- For example, in binary classification, disparate impact asks for the following *constraints* to hold:

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

Regularized optimization

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

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

where $r : \mathbb{R}^n \times \mathbb{R}$ is often convex and potentially nonsmooth, but this can be computationally expensive (due to need to tune hyperparameters), especially to achieve *exact* satisfaction

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

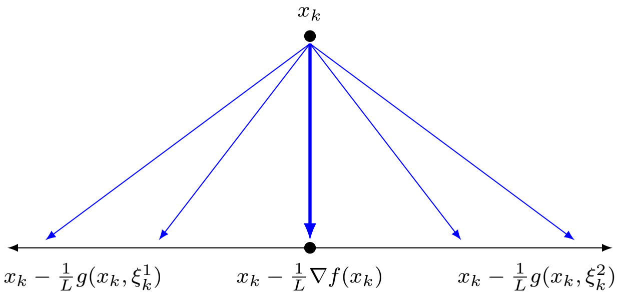
$$\begin{aligned} \min_{x \in \mathbb{R}^n} \ & f(x), \quad \text{where} \quad f(x) = \mathbb{E}_\omega[F(x,\omega)] \\ \text{s.t.} \ & c_\mathcal{E}(x) = 0 \\ & c_\mathcal{I}(x) \leq 0 \end{aligned}$$

Outline

Stochastic gradient (*not* descent) method

Consider the unconstrained setting where $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$ is Lipschitz continuous with constant L, namely,

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



Adaptive stochastic gradient method

This method can be made *adaptive* in various ways

- step-size selection
- scaling matrix
- error in gradient estimator

That said, in the *fully stochastic regime*, the convergence behavior boils down to the same thing.

Stochastic gradient method

Algorithm SG : Stochastic Gradient

```
1: choose an initial point x_1 ∈ ℝ^n and step sizes {α_k} ⊂ ℝ_{>0}
2: for all k ∈ ℕ do
3:     set x_{k+1} ← x_k − α_k g_k, where g_k ≈ ∇f(x_k)
4: end for
```

Formally, $\{(x_k, g_k)\}$ is a realization of the stochastic process $\{(X_k, G_k)\}$, where

- $\mathcal{F}_1 = \sigma(x_1)$ and, for $k \geq 2$, $\mathcal{F}_k$ is the σ-algebra generated by $\{G_1, \ldots, G_{k-1}\}$
- (for simplicity) $\mathbb{E}[G_k|\mathcal{F}_k] = \nabla f(X_k)$ and $\mathbb{E}[\|G_k - \nabla f(X_k)\|_2^2|\mathcal{F}_k] \leq M$

The algorithm achieves *eventual descent in expectation* with appropriate step-size selection:

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

SG theory

Theorem SG

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

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

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

and $\{\nabla f(X_k)\} \to 0$ *almost surely*

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 positive definite over $\text{Null}(J)$, two viewpoints:

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

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

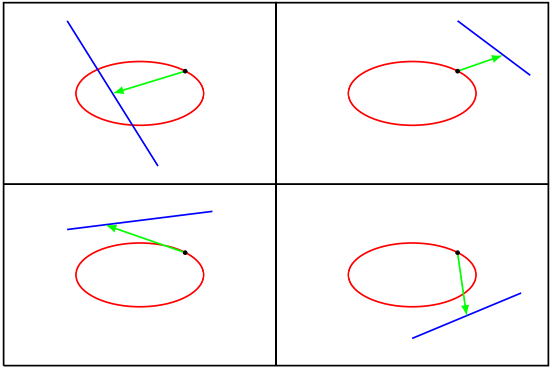


Figure: Illustrations of SQP subproblem solutions

SQP with backtracking line search

Algorithm guided by merit function with adaptive parameter τ defined by

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

Algorithm : SQP w/ line search

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

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

    update merit parameter: set τ_k to ensure

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

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

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

end for

SQP with prescribed step-size rule

Algorithm : SQP w/ prescribed step-size rule

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

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

  update merit parameter: set τ_k to ensure

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

  compute step size: set $x_{k+1} \leftarrow x_k + \alpha_k d_k$ where, for sufficiently small $\beta_k \in \mathbb{R}_{>0}$,

$$\alpha_k \leftarrow \frac{2(1-\eta)\beta_k \tau_k}{\tau_k L + \Gamma}$$

end for

Convergence theory

Assumption

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

Theorem

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

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

Stochastic SQP with adaptive step sizes

Algorithm : Stochastic SQP

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

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

3: compute step: solve

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

4: update merit parameter: set τ_k to ensure

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

5: compute adaptive step-size bound: set $\tilde{\alpha}_k$ as the largest value of $\alpha \in \mathbb{R}_{\geq 0}$ such that

$$0 \geq \varphi_k(\alpha) = (\eta - 1)\alpha\beta_k \Delta l(x_k, \tau_k, g_k, d_k) + \|c_k + \alpha \nabla c(x_k)^T d_k\|_2 - \|c_k\|_2 + \alpha(\|c_k\|_2 - \|c_k + \nabla c(x_k)^T d_k\|_2) + \tfrac{1}{2}(\tau_k L + \Gamma)\alpha^2 \|d_k\|_2^2$$

6: compute step size: set $x_{k+1} \leftarrow x_k + \alpha_k d_k$ where, for sufficiently small $\beta_k \in \mathbb{R}_{>0}$,

$$\alpha_k \in [\alpha_{k,\min}, \alpha_{k,\max}], \text{ with } \alpha_{k,\min} \leftarrow \frac{2(1-\eta)\beta_k \tau_k}{\tau_k L + \Gamma}$$

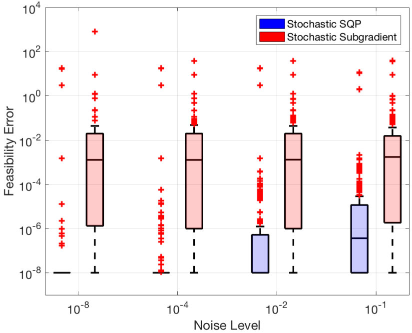
$$\alpha_{k,\max} \leftarrow \min\{\tilde{\alpha}_k, \alpha_{k,\min} + \theta\beta_k^2\}$$

7: **end for**

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

CUTE problems with noise added to gradients with different noise levels

- StochasticSQP vs. stochastic subgradient method



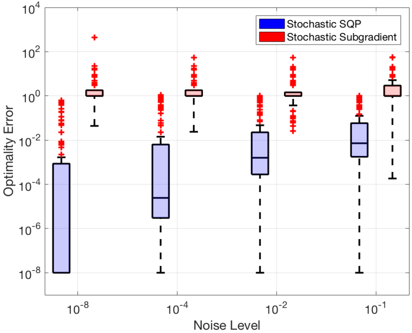


Figure: Box plots for feasibility errors (left) and optimality errors (right).

Fundamental lemma

Recall in the unconstrained setting that

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

Lemma

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

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

Good merit parameter behavior

Lemma

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

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

Theorem

Conditioned on $\mathcal{E}$, one finds

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

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

Good merit parameter behavior

Lemma

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

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

Theorem

Conditioned on $\mathcal{E}$, one finds

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

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

Lagrange multiplier convergence

How about convergence of the Lagrange multiplier sequence?

- The prior theorem considers the *true* multplier that we do not compute.
- The *last* multiplier is always subject to error.

If the primal iterates do not converge, then is there hope of anything?

We (upcoming paper with Xin Jiang and Qi Wang) have conditions under which

- the stationarity measure and primal iterates converge almost surely (like for SG), and
- correspondingly, an *averaged* multiplier sequence converges almost surely.

A consequence of the martingale central limit theorem.

Main challenges of adaptivity

Adaptivity, such as that for step sizes, is one type of challenge.

- As long as parameter sequences are prescribed, or at least controlled by prescribed sequences, then convergence can be guaranteed, perhaps with some additional steps.
- We have accomplished this as well in the context of an interior-point method (Qi Wang's talk).

Adaptivity of quantities such as the merit parameter is another type of (huge) challenge.

- The function that the algorithm is minimizing is *changing* during the optimization.
- Algorithmic behavior is *not* determined solely by the initial conditions.

I will outline our approach for handling this challenge in the context of proving a worst-case complexity.

Outline

SQP with prescribed step-size rule

First, recall the deterministic algorithm:

Algorithm : SQP w/ prescribed step-size rule

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

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

4: update merit parameter: set τ_k to ensure

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

5: compute step size: set $x_{k+1} \leftarrow x_k + \alpha_k d_k$ where, for sufficiently small $\beta_k \in \mathbb{R}_{>0}$,

$$\alpha_k \leftarrow \frac{2(1-\eta)\beta_k \tau_k}{\tau_k L + \Gamma}$$

6: **end for**

Complexity of deterministic algorithm

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

Lemma 6

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

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

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

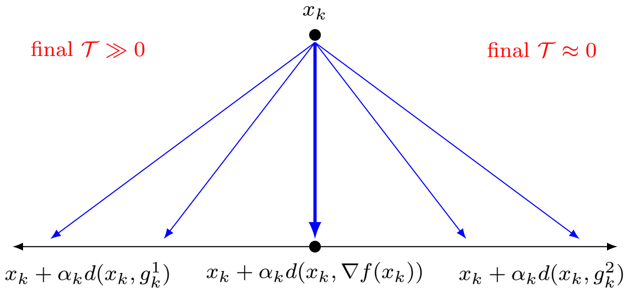
Theorem 7

For any $\epsilon \in (0,1)$, there exists $(\kappa_1, \kappa_2) \in (0,\infty) \times (0,\infty)$ such that $\|\nabla f(x_k) + J_k^T y_k\| \leq \epsilon$ and $\sqrt{\|c_k\|_1} \leq \epsilon$ in a number of iterations no more than

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

Challenge in the stochastic setting

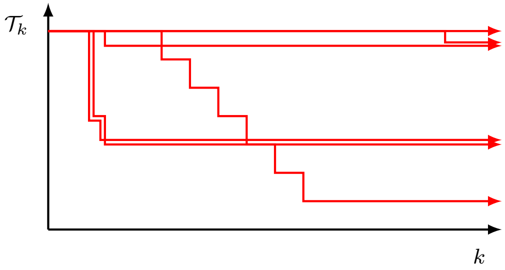
We are minimizing a function that is changing during the optimization.



Challenge in the stochastic setting

In the stochastic setting, minimum $\mathcal{T}$ is not determined by the initial point.

- Even if we assume $\mathcal{T}_k \geq \tau_{\min} > 0$ for all k in any realization, the final $\mathcal{T}$ is not determined.
- This means we cannot cast all reductions in terms of some fixed constant τ.



Our approach

In fact, $\mathcal{T}$ reaching some minimum value is not necessary.

- Important: Diminishing probability of continued imbalance between "true" merit parameter update and "stochastic" merit parameter update.
- In iteration k, the algorithm has obtained the merit parameter value $\mathcal{T}_{k-1}$.
- If the true gradient is computed, then one obtains $\mathcal{T}_k^{\text{trial,true}}$.

Lemma 8

Suppose that the merit parameter is reduced at most $s_{\max}$ times. For any $\delta \in (0,1)$, one finds that

$$\mathbb{P}\left[|\{k : \mathcal{T}_k^{trial,true} < \mathcal{T}_{k-1}\}| \leq \left\lceil \frac{\ell(s_{\max}, \delta)}{p} \right\rceil\right] \geq 1 - \delta,$$

where $p \in (0,1)$ (related to a bounded imbalance assumption we make) and

$$\ell(s_{\max}, \delta) := s_{\max} + \log(1/\delta) + \sqrt{\log(1/\delta)^2 + 2 s_{\max} \log(1/\delta)} > 0.$$

Chernoff bound

How do we get there?

Lemma 9 (Chernoff bound, multiplicative form)

Let $\{Y_0, \ldots, Y_k\}$ *be independent Bernoulli random variables. Then, for any* $s_{\max} \in \mathbb{N}$ *and* $\delta \in (0,1)$,

$$\sum_{j=0}^{k} \mathbb{P}[Y_j = 1] \geq \ell(s_{\max}, \delta) \implies \mathbb{P}\left[\sum_{j=0}^{k} Y_j \leq s_{\max}\right] \leq \delta.$$

We construct a tree whose nodes are signatures of possible runs of the algorithm.

- A realization $\{g_0, \ldots, g_k\}$ belongs to a node if and only if a certain number of decreases of $\mathcal{T}$ have occurred and the probability of decrease in the current iteration is in a given closed/open interval.
- Bad leaves are those when the probability of decrease has accumulated beyond a threshold, yet the merit parameter has not been decreased sufficiently often.
- Along the way, we apply a Chernoff bound on a carefully constructed set of (independent Bernoulli) random variables to bound probabilities associated with bad leaves.

Node definition

Let $[k] := \{0, 1, \ldots, k\}$ and define

- $p_{[k]}$ = probabilities of merit parameter decreases
- $w_{[k]}$ = counter of merit parameter decreases

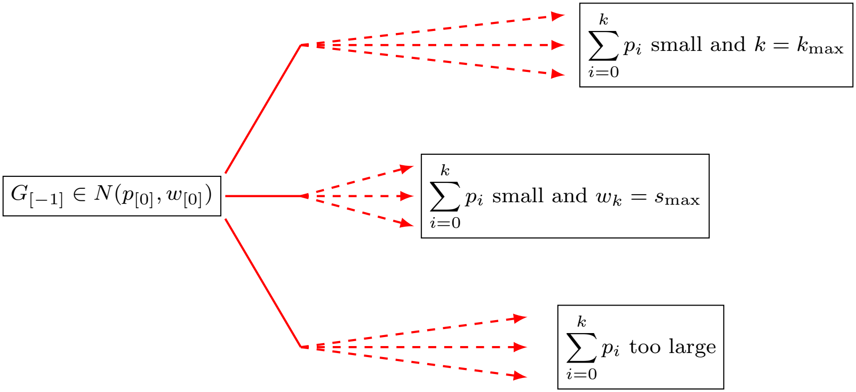
Then, define nodes of the tree according to

$$\boxed{G_{[k-1]} \in N(p_{[k]}, w_{[k]})}$$

if and only if

$$\boxed{\begin{aligned} G_{[k-2]} &\in N(p_{[k-1]}, w_{[k-1]}) \\ \mathbb{P}[\mathcal{T}_k < \mathcal{T}_{k-1} | \mathcal{F}_k] &\in \iota(p_k) \\ \sum_{i=1}^{k-1} \mathbb{1}[\mathcal{T}_i < \mathcal{T}_{i-1}] &= w_k \end{aligned}}$$

Visualization



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

Theorem 10

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 $1-\delta$,

$$\mathbb{E}[\|\nabla f(X_{k_*}) + J(X_{k_*})^T Y_{k_*}\|_2^2 + \|c(X_{k_*})\|_1] \leq \frac{\tau_{-1}(f_0 - f_{\inf}) + \|c_0\|_1 + M}{\sqrt{k_{\max}+1}} + \frac{(\tau_{-1} - \tau_{\min})(s_{\max}\log(k_{\max}) + \log(1/\delta))}{\sqrt{k_{\max}+1}}$$

Theorem 11

If the stochastic gradient estimates are sub-Gaussian, then with probabiliy $1-\bar{\delta}$

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

Outline

Summary

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

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

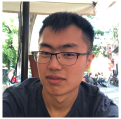
In terms of the design of *adaptive* stochastic algorithms, solving constrained problems presents

- new opportunities
- additional challenges

We have a framework for analyzing stochastic algorithms with adaptive algorithmic parameters

- used to analyze the worst-case complexity of a stochastic SQP algorithm
- results showing that the complexity is on par with the unconstrained setting

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," to appear in *Mathematical Programming*.
- F. E. Curtis, S. Liu, and D. P. Robinson, "Fair Machine Learning through Constrained Stochastic Optimization and an ϵ-Constraint Method," (under third round of review).
- F. E. Curtis, D. P. Robinson, and B. Zhou, "Sequential Quadratic Optimization for Stochastic Optimization with Deterministic Nonlinear Inequality and Equality Constraints," https://arxiv.org/abs/2302.14790.
- F. E. Curtis, V. Kungurtsev, D. P. Robinson, and Q. Wang, "A Stochastic-Gradient-based Interior-Point Algorithm for Solving Smooth Bound-Constrained Optimization Problems," https://arxiv.org/abs/2304.14907.