Stochastic Algorithms for Continuous Optimization with Nonlinear Constraints

Frank E. Curtis, Lehigh University

involving joint work with

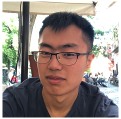
Albert S. Berahas (U. of Michigan), **Xin Jiang** (Lehigh), **Vyacheslav Kungurtsev** (Czech TU), **Suyun Liu** (Amazon), **Michael O'Neill** (UNC Chapel Hill), **Daniel P. Robinson** (Lehigh), **Qi Wang** (Lehigh), **Baoyu Zhou** (Chicago Booth)

presented at

European Conference on Computational Optimization

September 25, 2023

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," to appear in *Mathematics of Operations Research*.
- 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," to appear in *Optimization Letters*.
- 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.

Outline

Motivation

Stochastic SQP

Extensions

Conclusion

Outline

Motivation

Stochastic SQP

Extensions

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 continuously differentiable

- 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.)

Learning: Prediction function

Our aim is to determine a prediction function $p \in \mathcal{P}$, where $\mathcal{P}$ is some family of functions, such that

$$p(a_j)$$

yields an accurate prediction corresponding to any given input feature vector a_j.

Learning: Prediction function, parameterized

For practicality, let us say that the family is parameterized by some vector x such that

$$p(a_j, x)$$

yields an accurate prediction corresponding to any given input feature vector a_j.

Learning: Supervised

In the context of supervised learning, we have known input-output pairs $\{(a_j, b_j)\}_{j=1}^{n_o}$, then

$$\min_{x \in \mathbb{R}^n} \frac{1}{n_o} \sum_{j=1}^{n_o} \ell(p(a_j, x), b_j)$$

becomes our empirical-loss training problem to determine the optimal parameter vector x.

Learning: Supervised and regularized

If, in addition, we aim to impose some structure on the solution x, then we may consider

$$\min_{x \in \mathbb{R}^n} \; \frac{1}{n_o} \sum_{j=1}^{n_o} \ell(p(a_j, x), b_j) + r(x)$$

where r is a *regularization* function.

Learning: Supervised and regularized

If, in addition, we aim to impose some structure on the solution x, then we may consider

$$\min_{x \in \mathbb{R}^n} \ \frac{1}{n_o} \sum_{j=1}^{n_o} \ell(p(a_j, x), b_j) + r(x)$$

where r is a *regularization* function. But is this the right approach for *informed* learning?

Learning: Supervised and informed with *soft* constraints

Added to the loss (e.g., mean-squared error or other data-fitting term), we might consider

$$\min_{x\in\mathbb{R}^n}\ \frac{1}{n_o}\sum_{j=1}^{n_o}\ell(p(a_j,x),b_j)+\frac{1}{n_c}\sum_{j=1}^{n_c}\phi(p(\tilde{a}_j,x),\dots,\tilde{b}_j)$$

where $\{(\tilde{a}_j,\tilde{b}_j)\}_{j=1}^{n_c}$ are some known input-output pairs and ϕ encodes known information.

Learning: Supervised and informed through layer design

Another viable approach is to embed information through the prediction function itself such that

$$\min_{x \in \mathbb{R}^n} \frac{1}{n_o} \sum_{j=1}^{n_o} \ell(\hat{p}(a_j, x), b_j)$$

ensures that information is enforced with every forward pass. (Expense?)

Learning: Supervised and informed with *hard* constraints

Back to the "original" family for p, how about imposing hard constraints during training, as in

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \ & \frac{1}{n_o} \sum_{j=1}^{n_o} \ell(p(a_j, x), b_j) \\ \text{s.t. } & \varphi(p(\tilde{a}_j, x), \dots, \tilde{b}_j) = 0 \text{ (or } \leq 0) \ \text{ for all } \ i \in \{1, \dots, n_c\} \end{aligned}$$

such that we restrict attention to functions that are informed implicitly?

Expected-loss training problems

For the sake of generality/generalizability, the expected-loss objective function is

$$\int_{\mathcal{A}\times\mathcal{B}} \ell(p(a,x),b)\mathrm{d}\mathbb{P}(a,b) \equiv \mathbb{E}_\omega[F(x,\omega)] =: f(x).$$

One might consider various paradigms for imposing the constraints:

- expectation constraints
- (distributionally) robust constraints
- probabilistic (i.e., chance) constraints

For our recent work, we consider constraints whose values and derivatives can be computed:

$$c_{\mathcal{E}}(x) = 0 \quad \text{and} \quad c_{\mathcal{I}}(x) \leq 0$$

e.g., as in imposing a fixed set of constraints corresponding to a fixed set of sample data.

Physics-informed learning (e.g., PINNs)

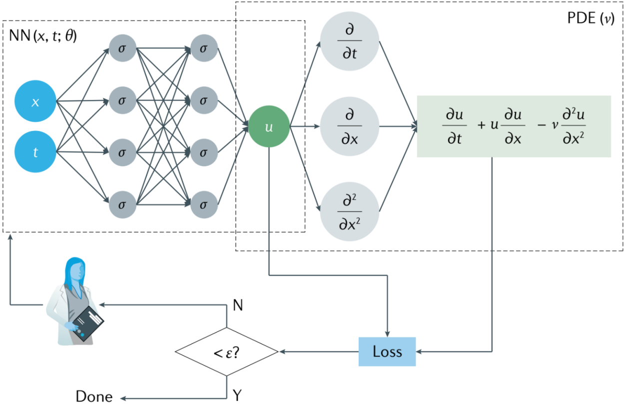


Photo: Karniadakis et al.

Fair learning

Let

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

Given a prediction function p and loss ℓ, the expected-loss minimization problem is

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

However, the resulting 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, etc.)
- For example, in binary classification, disparate impact may be expressed as the constraint

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

Constrained optimization (stochastic algorithms)

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

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

- Classical applications under uncertainty, constrained DNN training, etc.
- Besides cases involving a deterministic equivalent...
- ... very few algorithms so far (mostly penalty methods)

Outline

Motivation

Stochastic SQP

Extensions

Conclusion

Equality-constrained setting (to start)

Consider the *equality-constrained* optimization problem:

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

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:

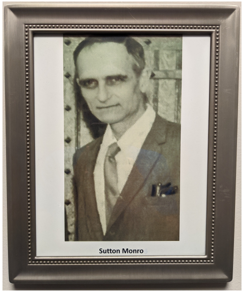
- Feasible methods are not tractable
- ... so no projection methods, Frank-Wolfe, etc.
- "Two-phase" methods are not effective
- ... so should not search for feasibility, then optimize.

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

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

Stochastic gradient method (SG)

Stochastic approximation by Herbert Robbins and Sutton Monro (1951)



Sutton Monro, former Lehigh faculty member

Stochastic gradient (*not* descent)

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

Algorithm SG : Stochastic Gradient

1: choose an initial point $x_1 \in \mathbb{R}^n$ and step sizes $\{\alpha_k\} > 0$
2: **for** $k \in \{1, 2, \dots\}$ **do**
3: set $x_{k+1} \leftarrow x_k - \alpha_k g_k$, where $\mathbb{E}[G_k|\mathcal{F}_k] = \nabla f(X_k)$ and $\mathbb{E}[\|G_k - \nabla f(X_k)\|_2^2|\mathcal{F}_k] \leq M$
4: **end for**

Notation: $\{(x_k, g_k)\}$ is a realization of the stochastic process $\{(X_k, G_k)\}$ with filtration $\{\mathcal{F}_k\}$

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

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

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

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* $k \in \mathbb{N}$*:*

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

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

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

SG illustration

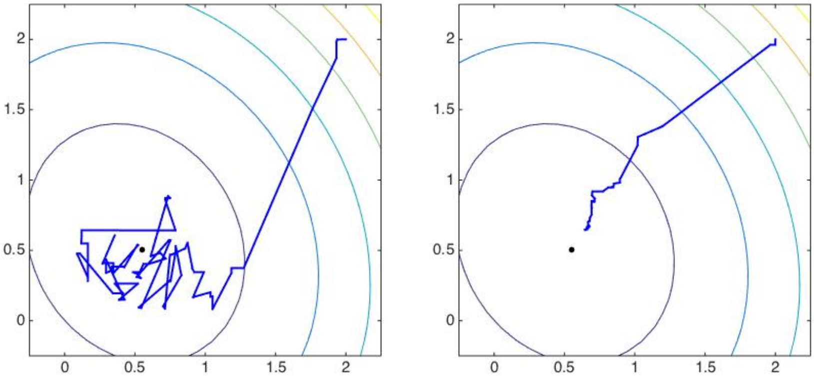


Figure: SG with fixed step size (left) vs. diminishing step sizes (right)

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 positive definite over 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

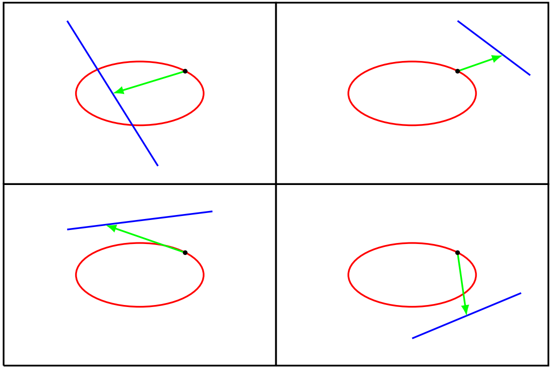


Figure: Illustrations of SQP subproblem solutions

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

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

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

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

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

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

- $\{\alpha_k\} \geq \alpha_{\min}$ *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 optimality error vanishes, specifically,*

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

Toward stochastic SQP

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

  step sizes determined by Lipschitz constants

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

SQP with adaptive step sizes

Algorithm SQP w/o line search

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

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

  compute step size: set

$$\widehat{\alpha}_k \leftarrow \frac{2(1-\eta)\Delta q(x_k, \tau_k, \nabla f(x_k), d_k)}{(\tau_k L + \Gamma)\|d_k\|_2^2} \quad \text{and} \quad \widetilde{\alpha}_k \leftarrow \widehat{\alpha}_k - \frac{4\|c_k\|_1}{(\tau_k L + \Gamma)\|d_k\|_2^2}$$

  then

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

  then set $x_{k+1} \leftarrow x_k + \alpha_k d_k$
end for

Convergence theory: Nearly identical as for SQP w/ line search.

Stochastic SQP with adaptive step sizes

Algorithm : Stochastic SQP

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

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

5: compute step size: set

$$\widehat{\alpha}_k \leftarrow \frac{\beta_k \Delta q(x_k, \tau_k, g_k, d_k)}{(\tau_k L + \Gamma)\|d_k\|_2^2} \quad \text{and} \quad \widetilde{\alpha}_k \leftarrow \widehat{\alpha}_k - \frac{4\|c_k\|_1}{(\tau_k L + \Gamma)\|d_k\|_2^2}$$

6: then

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

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

Assume $\{g_k\}$ is a realization of $\{G_k\}$ with $\mathbb{E}[G_k|\mathcal{F}_k] = \nabla f(X_k)$ and $\mathbb{E}[\|G_k - \nabla f(X_k)\|_2^2|\mathcal{F}_k] \leq M$

Fundamental lemma

Recall in the unconstrained setting that

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

Lemma

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

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

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 $\mathcal{E}$*, for large* k*, one finds*

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

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

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 $\mathcal{E}$, for large k, one finds

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

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

Poor merit parameter behavior

$\{\mathcal{T}_k\} \searrow 0$:

- cannot occur if $\|G_k - \nabla f(X_k)\|_2$ is bounded uniformly
- occurs with small probability if distribution of G_k has "small tails"

$\{\mathcal{T}_k\}$ remains too large:

- under a modest assumption, occurs with probability zero

Numerical results: (Matlab) 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 penalty parameter

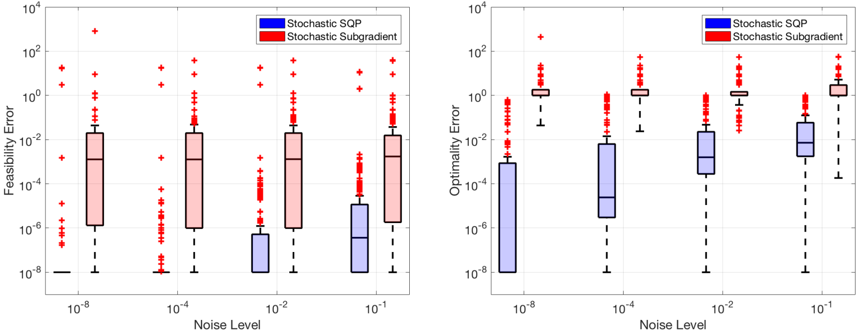


Figure: Box plots for feasibility errors (left) and optimality errors (right).

Outline

Motivation

Stochastic SQP

Extensions

Conclusion

Summary

Since our original work, we have considered various extensions.

- stronger convergence guarantees (convergence in probability $\rightarrow$ almost-sure convergence)
- convergence of Lagrange multiplier estimates
- relaxed constraint qualifications
- worst-case complexity guarantees
- generally constrained problems (with inequality constraints as well)
- interior-point methods
- iterative linear system solvers and inexactness

Almost-sure convergence of merit function value

Convergence of the algorithm is driven by the exact merit function

$$\phi_\tau(X) = \tau f(X) + \|c(X)\|$$

Reductions in a local model of ϕ_τ can be tied to a stationarity measure

$$\Delta q_\tau(X, \nabla f(X), H, D^{\text{true}}) \qquad \sim \qquad \|\nabla f(X) + \nabla c(X) Y\|^2 + \|c(X)\|$$

Lemma

Suppose $\mathbb{E}[G_k | \mathcal{F}_k] = \nabla f(X_k)$ *and* $\mathbb{E}[\|G_k - \nabla f(X_k) | \mathcal{F}_k\|^2] \leq M$*. Then, by a classical theorem of Robbins and Siegmund (1971), one finds that, almost surely,*

$$\lim_{k \to \infty} \{\phi_\tau(X_k)\} \text{ exists and is finite and}$$

$$\liminf_{k \to \infty} \Delta q_\tau(X_k, \nabla f(X_k), H_k, D_k^{\text{true}}) = 0$$

Almost-sure convergence of the primal iterates

If $\{X_k\}$ stays within a neighborhood of x_* almost surely, where x_* is a stationary point at which a generalization of the Polyak–Łojasiewicz condition holds, then almost-sure convergence follows:

Theorem

Suppose that there exists $x_* \in \mathcal{X}$ *with* $c(x_*) = 0$, $\mu \in \mathbb{R}_{>1}$, *and* $\epsilon \in \mathbb{R}_{>0}$ *such that for all*

$$x \in \mathcal{X}_{\epsilon,x_*} := \{x \in \mathcal{X} : \|x - x_*\|_2 \leq \epsilon\}$$

one finds that

$$\phi_\tau(x) - \phi_\tau(x_*) \begin{cases} = 0 & \text{if } x = x_* \\ \in (0, \mu(\tau\|Z(x)^T\nabla f(x)\|_2^2 + \|c(x)\|_2)] & \text{otherwise,} \end{cases}$$

where for all $x \in \mathcal{X}_{\epsilon,x_*}$ *one defines* $Z(x) \in \mathbb{R}^{n\times(n-m)}$ *as some orthonormal matrix whose columns form a basis for the null space of* $\nabla c(x)^T$*. Then, if* $\limsup_{k\to\infty}\{\|X_k - x_*\|_2\} \leq \epsilon$ *almost surely, it follows that*

$$\{\phi_\tau(X_k)\} \xrightarrow{a.s.} \phi_\tau(x_*), \quad \{X_k\} \xrightarrow{a.s.} x_*, \quad \text{and} \quad \left\{\begin{bmatrix}\nabla f(X_k) + \nabla c(X_k)Y_k^{\text{true}} \\ c(X_k)\end{bmatrix}\right\} \xrightarrow{a.s.} 0.$$

Lagrange multiplier convergence

Theorem

Suppose (x_*, y_*) *is a stationary point. Then, for any* $k \in \mathbb{N}$*, one finds* $\|X_k - x_*\|_2 \leq \epsilon$ *implies*

$$\|Y_k - y_*\|_2 \leq \kappa_y \|X_k - x_*\|_2 + r^{-1}\|\nabla f(X_k) - G_k\|_2$$
$$\textit{and} \;\; \|Y_k^{\text{true}} - y_*\|_2 \leq \kappa_y \|X_k - x_*\|_2 \;\; \textit{for some} \;\; (\kappa, r) \in \mathbb{R}_{>0} \times \mathbb{R}_{>0}.$$

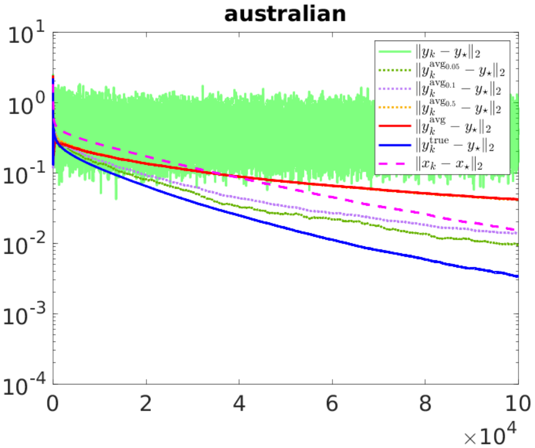
Computed multipliers *always* have error. Consider *averaged* multipliers $\{Y_k^{\text{avg}}\}$ instead.

Theorem

If the iterate sequence converges almost surely to x_**, i.e.,* $\{X_k\} \xrightarrow{a.s.} x_*$*, then*

$$\{Y_k^{\text{true}}\} \xrightarrow{a.s.} y_* \;\; \textit{and} \;\; \{Y_k^{\text{avg}}\} \xrightarrow{a.s.} y_*.$$

Constrained logistic regression: **australian** dataset (LIBSVM)



Relaxing constraint qualifications

Use a step decomposition method, handled infeasible and/or degenerate problems as well.

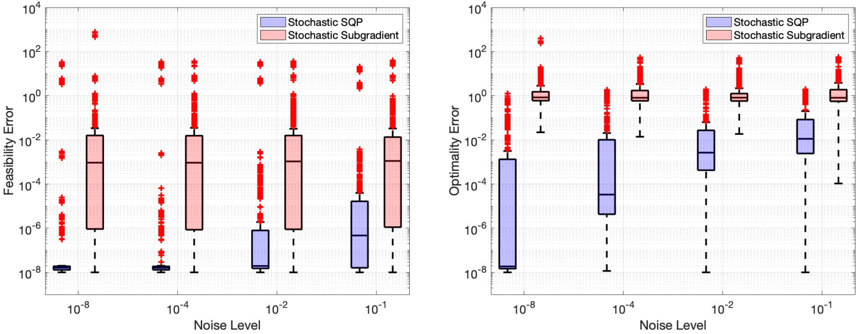


Figure: Box plots for feasibility errors (left) and optimality errors (right).

Complexity of $\mathcal{O}(\epsilon^{-2})$ for deterministic algorithm

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

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

Theorem

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

$$\left(\frac{\tau_0(f_1 - f_{\inf}) + \|c_1\|_1}{\min\{\kappa_1, \kappa_2 \tau_{\min}\}}\right)\epsilon^{-2}.$$

Complexity of $\widetilde{\mathcal{O}}(\epsilon^{-4})$ for stochastic algorithm

Theorem

Suppose the algorithm is run $k_{\max}$ iterations with

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

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

$$\mathbb{E}[\|\nabla f(X_{K_*}) + J(X_{K_*})^T Y_{k_*}^{\text{true}}\|_2^2 + \|c(X_{K_*})\|_1] \leq \frac{\tau_0(f(x_1) - f_{\inf}) + \|c(x_1)\|_1 + M}{\sqrt{k_{\max}+1}} + \frac{(\tau_{-1} - \tau_{\min})(s_{\max}\log(k_{\max}) + \log(1/\delta))}{\sqrt{k_{\max}+1}}$$

Theorem

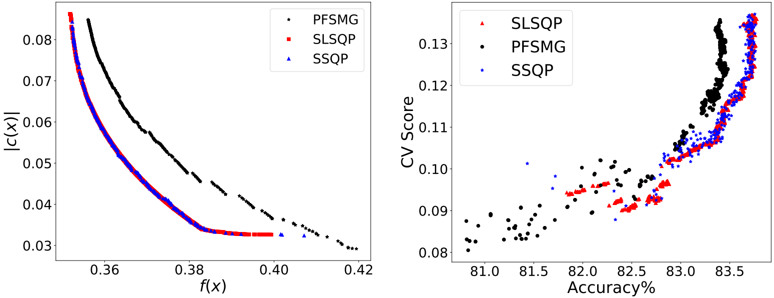
If the stochastic gradient estimates are sub-Gaussian, then w.p. $1-\bar{\delta}$

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

Inequality-constrained optimization

Stochastic SQP for inequality constrained problems

- employed in an ϵ-constraint method for fair machine learning



$$\min_{x\in\mathbb{R}^n} \frac{1}{N^o} \sum_{(v_i,y_i)\in D_o} \ell(x,v_i,y_i) \quad \text{s.t.} \quad -\epsilon \le \frac{1}{N^c} \sum_{(v_i,a_i)\in D_c} (a_i-\overline{a})x^T v_i \le \epsilon$$

Interior-point methods

Stochastic *single-loop* algorithm (prescribed barrier sequence $\{\mu_k\} \searrow 0$) with convergence guarantees.

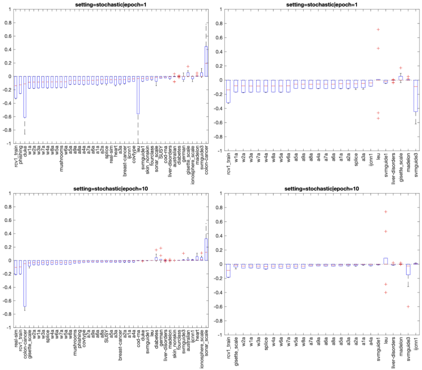


Figure: Deterministic setting (left) and stochastic setting (right)

Iterative methods and inexactness

Inexact subproblem solves

- stochasticity and inexactness(!)

Iterative methods employed to solve

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

termination tests to determine when an inexact solution is sufficient for convergence.

Physics-informed learning

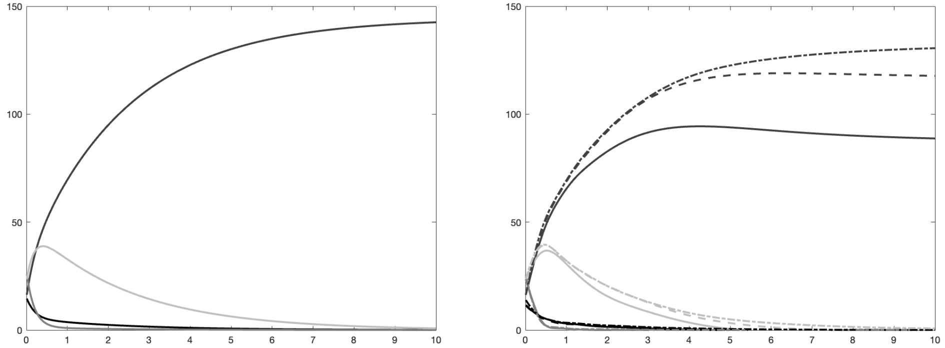


Figure: True solution (left) and predicted solutions (right).

Outline

Motivation

Stochastic SQP

Extensions

Conclusion

Summary

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

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

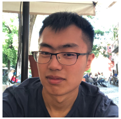
Equality-constraints-only setting:

- convergence in probability with complexity guarantees
- almost-sure convergence of primal iterates and averaged Lagrange multipliers
- relaxed constraint qualifications
- inexact subproblem solves

Generally constrained setting (with inequality constraints as well):

- stochastic SQP
- stochastic interior-point (bounds only so far, but generally constrained in progress)

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," to appear in *Mathematics of Operations Research*.
- 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," to appear in *Optimization Letters*.
- 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.