Stochastic-Gradient-Based Algorithms for Solving Nonlinearly Constrained Optimization Problems

Frank E. Curtis, Lehigh University

presented at

SIAM Conference on the Mathematics of Data Science

Atlanta, Georgia

October 22, 2024

Outline

Outline

Constrained continuous optimization

Consider the setting of solving constrained continuous optimization problems of the form

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

when at any $x \in \mathbb{R}^n$ one has that

- $c_{\mathcal{E}}(x)$ and $c_{\mathcal{I}}(x)$ can be computed exactly
- $\nabla c_{\mathcal{E}}(x)$ and $\nabla c_{\mathcal{I}}(x)$ can be computed exactly
- $f(x)$ and $\nabla f(x)$ cannot be computed exactly—only have (unbiased) estimates

Supervised learning

Aim: Determine a prediction function $p(\cdot, x)$ in a family $\mathcal{P}$ by finding the optimal x for

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

where $\{(a_j, b_j)\}_{j=1}^{n_0}$ is a set of known input-output pairs.

Supervised learning, informed with *soft* constraints

To incorporate some prior knowledge (e.g., physical laws), we may consider

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

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

Supervised learning, informed with *hard* constraints

Alternatively, or in addition, we may include some hard constraints

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

$$\text{s.t. } \varphi(p(\tilde{a}_j,x),\dots,\tilde{b}_j) = 0 \text{ (or } \leq 0) \;\; \text{for some } \; i \in \{1,\dots,n_c\}$$

which has a significant effect on performance if (and only if!) certain algorithms are employed

Expected-loss training problems

For the sake of generality/generalizability, the expected-loss objective function can be written as

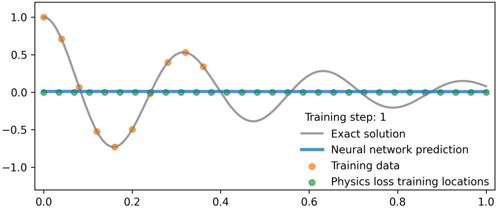
$$\int_{\mathcal{A}\times\mathcal{B}} \ell(p(a,x),b)\mathrm{d}\mathbb{P}(a,b) \equiv \mathbb{E}_{\omega}[F(x,\omega)] =: f(x)$$

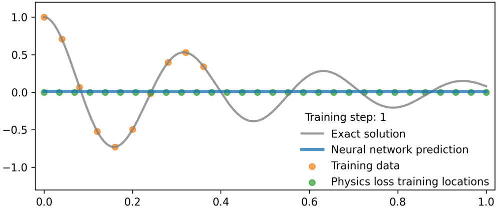
The constraints, on the other hand, can be expressed as

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

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

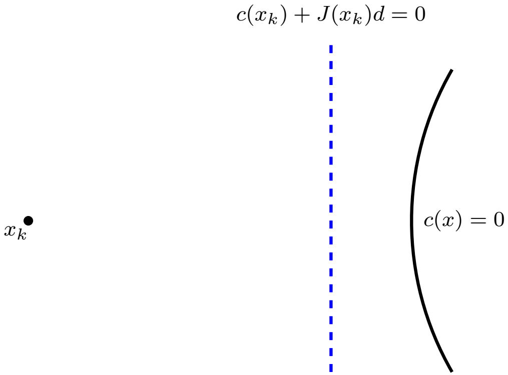
Predicting movement of a spring



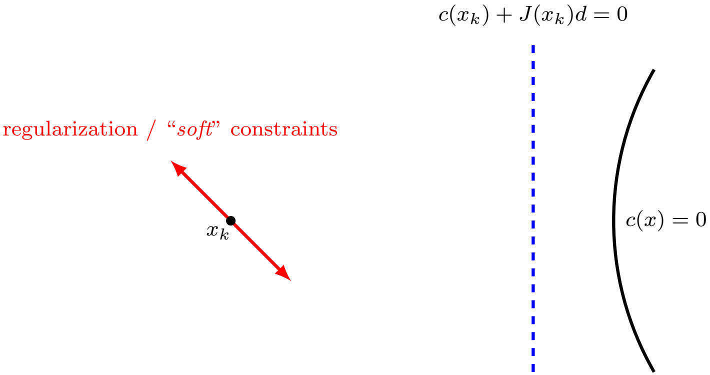


Problem from https://benmoseley.blog/blog/

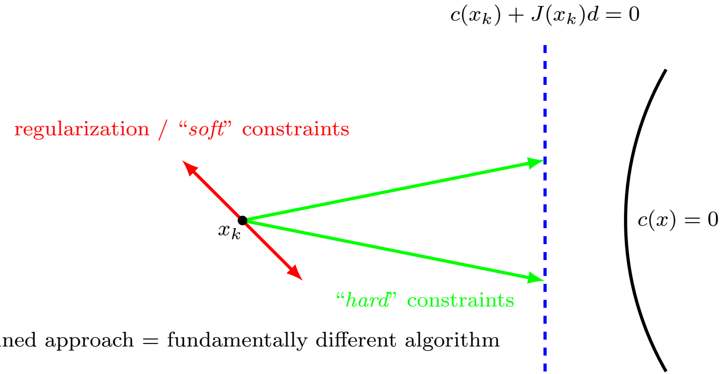
SQP illustration: Why does it work?



SQP illustration: Why does it work?



SQP illustration: Why does it work?

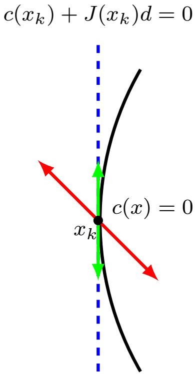


constrained approach = fundamentally different algorithm

SQP illustration: Why does it work?

regularization / "*soft*" constraints

"*hard*" constraints $\Longrightarrow$ step in null space



Outline

Motivation

Almost-Sure Convergence of Stochastic SQP

Numerical Experiments and P-Adam

Conclusion

Constrained stochastic optimization

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

where

- $f(x) = \mathbb{E}_\omega[F(x,\omega)]$
- c is continuously differentiable
- ∇f has Lipschitz constant L
- ∇c has Lipschitz constant Γ
- stationarity conditions:

$$\nabla f(x) + \nabla c(x) y = 0$$
$$c(x) = 0$$

Algorithm : Stochastic SQP

1: choose $x_1 \in \mathbb{R}^n$, $\tau \in \mathbb{R}_{>0}$
2: **for** $k \in \{1, 2, \dots\}$ **do**
3: estimate gradient: $g_k \approx \nabla f(x_k)$
4: compute step: solve

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

5: choose step size: for small $\beta_k \in \mathbb{R}_{>0}$,

$$\alpha_k \leftarrow \frac{\beta_k \tau}{\tau L + \Gamma}$$

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

Convergence in probability to stationarity

Assumption

- *τ is sufficiently small*
- *$\{\beta_k\} = \mathcal{O}(1/k)$ with β_1 sufficiently small*

Theorem (Berahas, Curtis, Robinson, Zhou (2021))

$$\liminf_{k\to\infty} \mathbb{E}\left[\|\nabla f(X_k) + \nabla c(X_k)^T Y_k^{\text{true}}\|^2 + \|c(X_k)\|\right] = 0$$

This shows that over some sequence the expected stationarity measure vanishes, but

- it does not guarantee that $\{X_k\}$ converges in any sense and
- the values $\{Y_k^{\text{true}}\}$ are not realized by the algorithm, so
- it does not guarantee anything about $\{Y_k\}$

Multipliers are important for verifying stationarity, active-set identification, etc.

Toward stronger guarantees

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* $\mathbb{E}[\|G_k - \nabla f(X_k)|\mathcal{F}_k\|^2] \leq \sigma^2$. *Using Robbins and Siegmund (1971) with*

$$P_k := \tfrac{\beta_k \tau}{\tau L + \Gamma} \Delta q_\tau(X_k, \nabla f(X_k), H_k, D_k^{\text{true}}), \quad Q_k := \tfrac{\beta_k^2 \tau^2 \sigma^2}{2\zeta(\tau L + \Gamma)}, \quad \text{and} \quad R_k := \phi_\tau(X_k) - \tau f_{\inf}$$

shows 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 multipliers as a (noisy) mapping of the primal iterates

In a standard manner, it can be shown that

$$Y_k = M_k(H_k(\nabla c(X_k)^\dagger)^T c(X_k) - G_k) \in \mathbb{R}^m,$$

where M_k is a product of a pseudoinverse of the derivative of c at X_k and a projection matrix:

$$M_k = \nabla c(X_k)^\dagger (I - H_k Z_k (Z_k^T H_k Z_k)^{-1} Z_k^T) \in \mathbb{R}^{m\times n}$$

If $\{X_k\} \xrightarrow{a.s.} x_*$, then one would expect

- $\{Y_k^{\text{true}}\} \xrightarrow{a.s.} y_*$ (i.e., as above with $\nabla f(X_k)$ in place of G_k)
- $\{Y_k\}$ noisy with error proportional to error in stochastic gradient estimators

True and average 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

$$\|Y_k - y_*\|_2 \leq \kappa_y \|X_k - x_*\|_2 + r^{-1}\|\nabla f(X_k) - G_k\|_2$$
$$\text{and} \quad \|Y_k^{\text{true}} - y_*\|_2 \leq \kappa_y \|X_k - x_*\|_2,$$

where $\kappa_y := \kappa_H L_c r^{-2} + L r^{-1} + \kappa_{\nabla f} L_{\mathcal{M}}$.

Unfortunately, this means that

- $\{Y_k\}$ *always* has error
- $\{Y_k^{\text{true}}\}$ converges if $\{X_k\}$ does, but these are not realized (requires $\{\nabla f(X_k)\}$)!

Idea: Averaging! Applying the Martingale central limit theorem, one can show that

Theorem

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

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

Outline

Motivation

Almost-Sure Convergence of Stochastic SQP

Numerical Experiments and P-Adam

Conclusion

Projected Adam

Algorithm P-Adam Projection-based Adam

Require: $m_{k-1} \in \mathbb{R}^d$, $v_{k-1} \in \mathbb{R}^d$, $w_k \in \mathbb{R}^d$, $g_k \in \mathbb{R}^d$, $\beta_1 \in (0,1)$, $\beta_2 \in (0,1)$, $\mu \in \mathbb{R}_{>0}$

Compute $\bar{g}_k \leftarrow (I - J(w_k)^T (J(w_k)J(w_k)^T)^{-1} J(w_k)) g_k$

Set $p_k \leftarrow \beta_1 p_{k-1} + (1-\beta_1)\bar{g}_k$

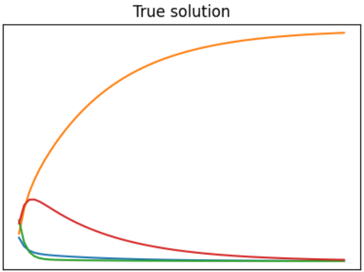
Set $q_k \leftarrow \beta_2 q_{k-1} + (1-\beta_2)(\bar{g}_k \circ \bar{g}_k)$, where $(\bar{g}_k \circ \bar{g}_k)_i = (\bar{g}_k)_i^2$ for all $i \in \{1,\ldots,d\}$

Set $\widehat{p}_k \leftarrow (1/(1-\beta_1^k))p_k$

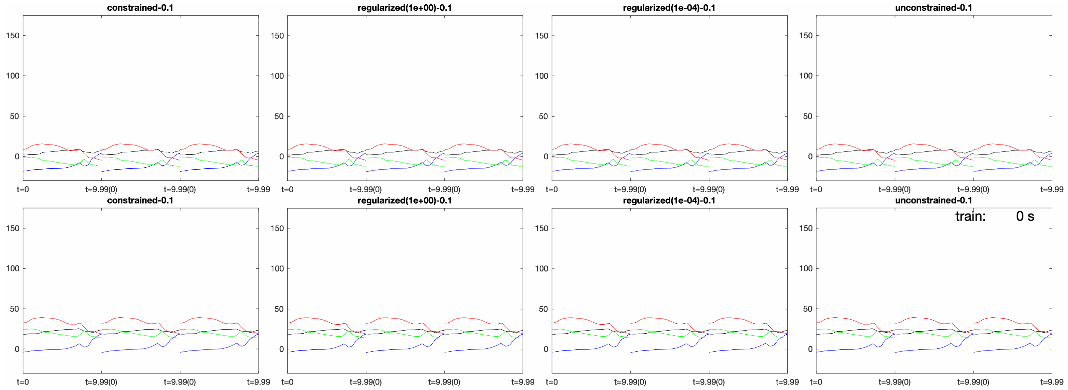
Set $\widehat{q}_k \leftarrow (1/(1-\beta_2^k))q_k$

Compute s_k by solving $\begin{bmatrix} \operatorname{diag}(\sqrt{\widehat{q}_k + \mu}) & J(w_k)^T \\ J(w_k) & 0 \end{bmatrix} \begin{bmatrix} s_k \\ \lambda_k \end{bmatrix} = -\begin{bmatrix} \widehat{p}_k \\ c_k \end{bmatrix}$

Predicting an ODE solution



Mass-balance-informed learning



Outline

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," *Mathematics of Operations Research*, https://doi.org/10.1287/moor.2021.0154, 2023.
- F. E. Curtis, D. P. Robinson, and B. Zhou, "Inexact Sequential Quadratic Optimization for Minimizing a Stochastic Objective Subject to Deterministic Nonlinear Equality Constraints," to appear in *INFORMS Journal on Optimization*, 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," *Mathematical Programming*, https://doi.org/10.1007/s10107-023-01981-1, 2023.
- F. E. Curtis, S. Liu, and D. P. Robinson, "Fair Machine Learning through Constrained Stochastic Optimization and an ϵ-Constraint Method," *Optimization Letters*, https://doi.org/10.1007/s11590-023-02024-6, 2023.
- F. E. Curtis, X. Jiang, and Q. Wang, "Almost-sure convergence of iterates and multipliers in stochastic sequential quadratic optimization," https://arxiv.org/abs/2308.03687.
- 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.

Where do we go from here?

There are many open questions:

- other algorithm variants with same guarantees
- strengthened guarantees (e.g., other growth conditions, convex settings)
- improved worst-case complexity properties
- loosened constraint qualification requirements
- second-order-type methods
- generalization properties
- trade-off analyses (Bottou–Bosquet)
- data-driven constraints

Thank you!

Questions?