Stochastic-Gradient-based Algorithms for Solving Nonconvex Constrained Optimization Problems

Frank E. Curtis, Lehigh University

presented at

Numerical Analysis Seminar

University of Maryland, College Park

October 15, 2024

Outline

Outline

Motivation

Stochastic Algorithms for Nonconvex Optimization

Extensions and Experimental Results

Conclusion

Appendix

Learning: Prediction function

Aim: Determine a prediction function p from a family $\mathcal{P}$ such that

$$p(a_j)$$

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

Learning: Prediction function, parameterized

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 *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* x.

Learning: Supervised and regularized

If 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 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. Is this good for *informed* learning?

Learning: Supervised and informed through model design

One approach is to embed information in the prediction function itself, so

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

ensures that information is enforced with every forward pass. (Is this enough and/or efficient?)

Learning: Supervised and informed with *soft* constraints

Added to the loss (e.g., mean-squared error), 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 known input-output pairs and ϕ encodes information.

Learning: Supervised and informed with *hard* constraints

Alternatively, how about *hard* constraints during training, as in

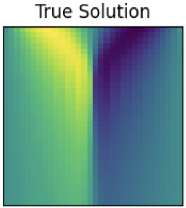
$$\min_{x \in \mathbb{R}^n} \ \frac{1}{n_o} \sum_{j=1}^{n_o} \ell(p(a_j, x), b_j)$$
$$\text{s.t.} \ \ \varphi(p(\tilde{a}_j, x), \dots, \tilde{b}_j) = 0 \ (\text{or} \leq 0) \ \ \text{for all} \ \ i \in \{1, \dots, n_c\}$$

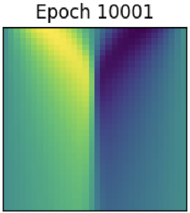
such that we restrict attention to functions that are informed implicitly?

Motivation and challenges: Stochastic algorithms for constrained optimization

Motivated by informed learning when model design + regularization is insufficient

- physics-informed machine learning
- fair (supervised) machine learning
- ... but algorithms are general-purpose, e.g., also for simulation optimization





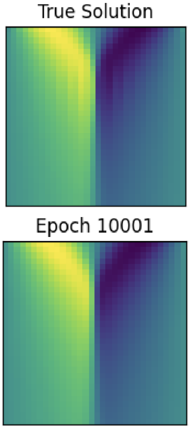
Motivation and challenges: Stochastic algorithms for constrained optimization

Motivated by informed learning when model design + regularization is insufficient

- physics-informed machine learning
- fair (supervised) machine learning
- . . . but algorithms are general-purpose, e.g., also for simulation optimization

Same challenges and questions as for unconstrained:

- convergence/complexity guarantees (adaptive algorithms)
- computational complexity
- stability guarantees
- generalization properties



Motivation and challenges: Stochastic algorithms for constrained optimization

Motivated by informed learning when model design + regularization is insufficient

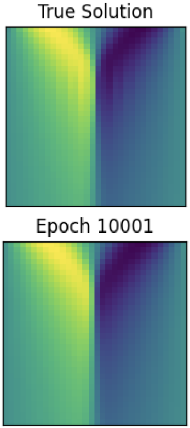
- physics-informed machine learning
- fair (supervised) machine learning
- . . . but algorithms are general-purpose, e.g., also for simulation optimization

Same challenges and questions as for unconstrained:

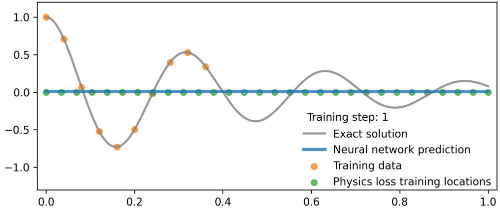
- convergence/complexity guarantees (adaptive algorithms)
- computational complexity
- stability guarantees
- generalization properties

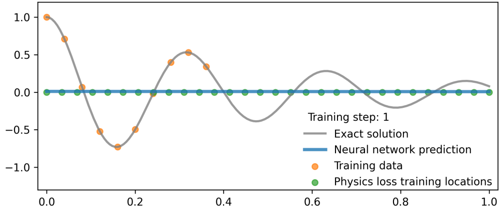
New challenges for *handling constraints as constraints*:

- (i.e., avoid penalty methods, augmented Lagrangian, etc.)
- balancing the objective and constraints
- degeneracy and infeasibility



Predicting movement of a spring





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

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

In this talk, constraints values and derivatives can be computed:

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

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

Outline

Stochastic gradient method

Consider $\min_{x \in \mathbb{R}^n} f(x)$, where $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$ is Lipschitz continuous with constant L.

Algorithm SG : Stochastic gradient method

1: choose an initial point $x_1 \in \mathbb{R}^n$ and step sizes $\{\alpha_k\} > 0$
2: **for** $k \in \{1, 2, \dots\} =: \mathbb{N}$ **do**
3: set $x_{k+1} \leftarrow x_k - \alpha_k g_k$, where $g_k \approx \nabla f(x_k)$
4: **end for**

Algorithm[†] behavior is defined by $(\Omega, \mathcal{F}, \mathbb{P})$, where

- $\Omega = \Gamma \times \Gamma \times \Gamma \times \cdots$ (sequence of draws determining stochastic gradients);
- $\mathcal{F}$ is a σ-algebra on Ω, the set of events (i.e., measurable subsets of Ω); and
- $\mathbb{P} : \mathcal{F} \to [0, 1]$ is a probability measure.

View any $\{(x_k, g_k)\}$ as a realization of $\{(X_k, G_k)\}$, where for all $k \in \mathbb{N}$

$$x_k = X_k(\omega) \text{ and } g_k = G_k(\omega) \text{ given } \omega \in \Omega.$$

[†] Robbins and Monro (1951); Sutton Monro = former Lehigh ISE faculty member

Convergence of SG

$\mathbb{E}[\cdot]$ = expectation w.r.t. $\mathbb{P}[\cdot]$. Analyze through associated sub-σ-algebras $\{\mathcal{F}_k\}$.

Assumption

For all $k \in \mathbb{N}$, one has that

- $\mathbb{E}[G_k|\mathcal{F}_k] = \nabla f(X_k)$ *and*
- $\mathbb{E}[\|G_k\|_2^2|\mathcal{F}_k] \leq M + M_{\nabla f}\|\nabla f(X_k)\|_2^2$

By Lipschitz continuity of ∇f and construction of the algorithm, one finds

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

by the assumption and since $f(X_k)$ and $\nabla f(X_k)$ are $\mathcal{F}_k$-measurable.

SG theory

Taking total expectation, one arrives at

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

Theorem

$$\alpha_k = \frac{1}{LM_{\nabla f}} \implies \mathbb{E}\left[\frac{1}{k}\sum_{j=1}^{k}\|\nabla f(X_j)\|_2^2\right] \leq M_k \xrightarrow{k\to\infty} \mathcal{O}\left(\frac{M}{M_{\nabla f}}\right)$$

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

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

$$\textit{(further steps)} \implies \nabla f(X_k) \to 0 \textit{ almost surely.}$$

Constrained optimization

Consider

$$\boxed{\begin{aligned} \min_{x\in\mathbb{R}^n}\ & f(x) \\ \text{s.t. } & c(x) = 0 \end{aligned}}$$

Option: Regularization / soft constraints (penalization), as in

$$\min_{x\in\mathbb{R}^n}\ \tau f(x) + \|c(x)\|_q^p\ (+y^T c(x)),$$

then employ a (stochastic) algorithm for unconstrained optimization.

Constrained optimization

Consider

$$\boxed{\begin{aligned}\min_{x\in\mathbb{R}^n}\ & f(x) \\ \text{s.t. } & c(x) = 0\end{aligned}}$$

Option: Regularization / soft constraints (penalization), as in

$$\min_{x\in\mathbb{R}^n}\ \tau f(x) + \|c(x)\|_q^p\ (+y^T c(x)),$$

then employ a (stochastic) algorithm for unconstrained optimization.

On the positive side, "exact" penalty function theory is well established:

- *can* solve the constrained problem, in theory.

Unfortunately, however, such an approach is not ideal:

- appropriate balance (τ and/or y) not known in advance
- $p = 1$ (nonsmooth), $p = 2$ (need $\tau \searrow 0$, ill-conditioning)

Sequential quadratic optimization (SQP)

Consider

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

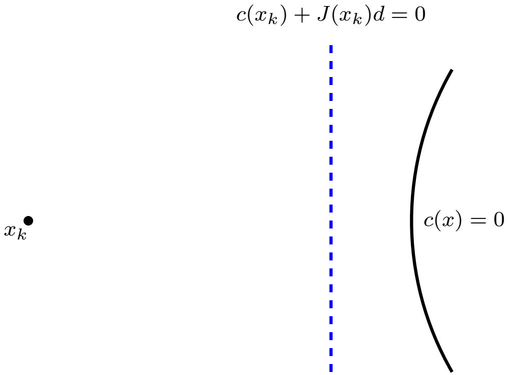
Option: With $J \equiv \nabla c^T$ and H positive definite over Null(J), two viewpoints:

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

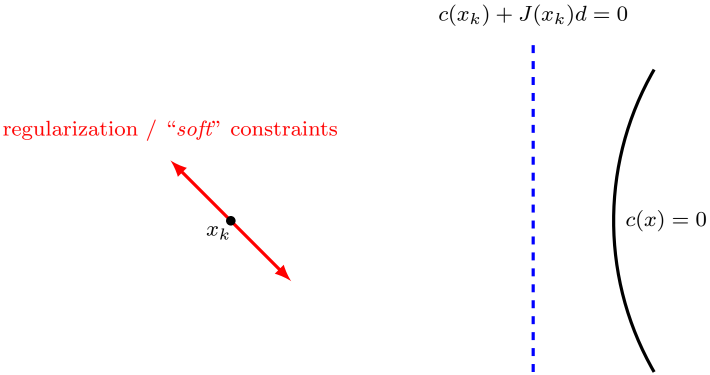
both leading to the same "Newton-SQP system":

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

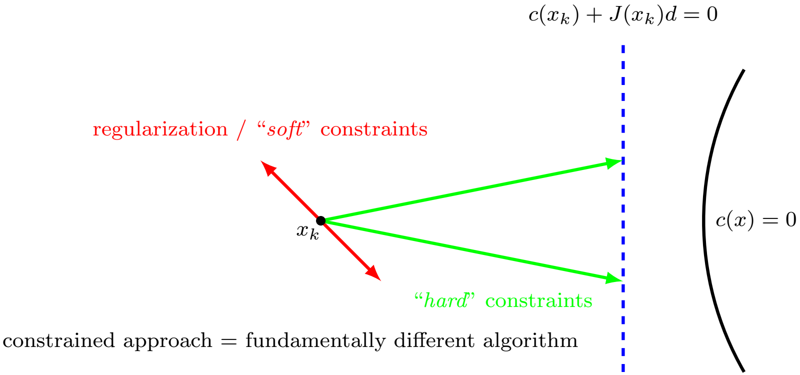
SQP illustration



SQP illustration



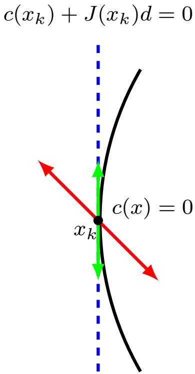
SQP illustration



SQP illustration

regularization / "*soft*" constraints

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



Stochastic SQP

Algorithm guided by merit function with adaptive parameter τ defined by

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

Algorithm : Stochastic SQP

choose $x_1 \in \mathbb{R}^n$, $\tau_0 \in (0,\infty)$, $\{\beta_k\} \in (0,1]^{\mathbb{N}}$
for $k \in \{1,2,\ldots\}$ **do**
  compute step: solve

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

  update merit parameter: set τ_k to ensure

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

  compute step size: set

$$\alpha_k = \Theta\left(\frac{\beta_k \tau_k}{\tau_k L_{\nabla f} + L_J}\right)$$

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

Convergence theory in *deterministic setting*

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,*

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

Stochastic setting: What do we want?

What we want/expect from the algorithm?

Note: We are interested in the stochastic approximation (SA) regime.

Ultimately, there are *many* questions to answer:

- convergence guarantees
- complexity guarantees
- tradeoff analysis (Bottou and Bousquet)
- generalization
- large-scale implementations
- beyond first-order (SG) methods

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

Theorem 4

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

Then, conditioned on $\mathcal{E}$,

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

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

Good merit parameter behavior

Theorem 4

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

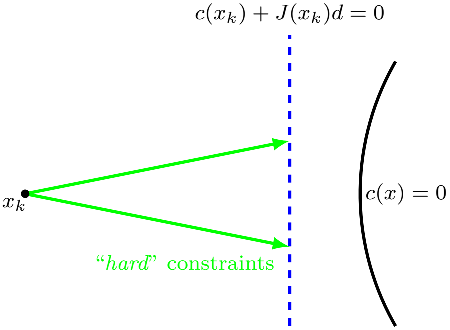
Then, conditioned on $\mathcal{E}$,

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

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

Key observation

Key observation is that $c(X_k)$ and $J(X_k)$ are $\mathcal{F}_k$-measurable.



Therefore, $\mathbb{E}[D_k|\mathcal{F}_k]$ = true step if $\nabla f(X_k)$ were known.

Numerical results: https://github.com/frankecurtis/StochasticSQP

Stochastic SQP (hard constraints) vs. stochastic subgradient (soft constraints)

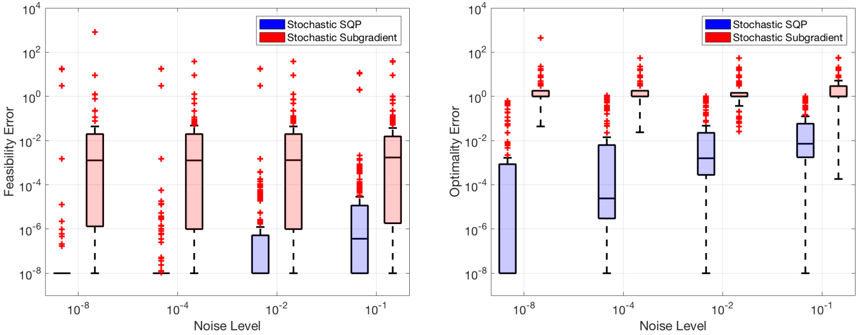


Figure: Box plots for feasibility errors (left) and optimality errors (right).

Outline

Summary

Since our original work, we have considered various extensions.

- stronger convergence guarantees (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
- diagonal scaling methods for saddle-point systems

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

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

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

Theorem

Suppose 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* $J(x)$*. Then, if* $\limsup_{k\to\infty}\{\|X_k - x_*\|_2\} \leq \epsilon$ *almost surely, it follows that*

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

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

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

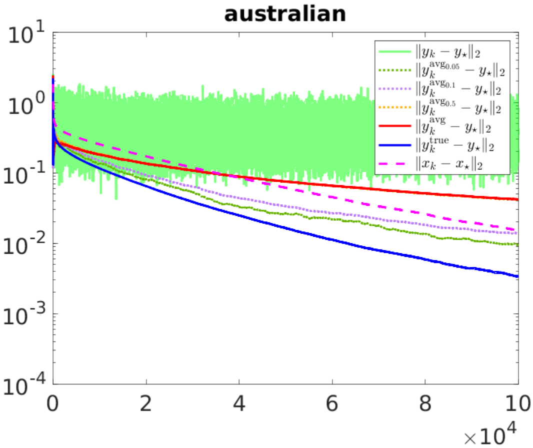
Computed multipliers *always* have error. Consider *averaged* multipliers $\{Y_k^{\text{avg}}\}$:

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)



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*

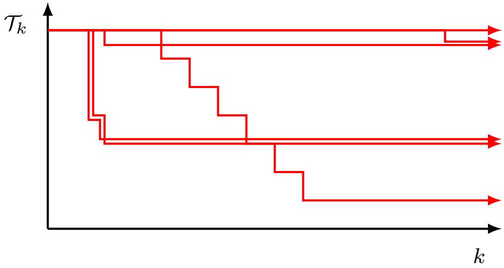
$$\|\nabla f(x_k) + J_k^T y_k\| \leq \epsilon \text{ and } \sqrt{\|c_k\|_1} \leq \epsilon$$

in a number of iterations no more than

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

Challenge in the stochastic setting

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



▸ Details

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

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* $1-\delta$*,*

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

Theorem

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

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

Inequality-constrained: Fair learning

Consider an ϵ-constraint method for fair machine learning:

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

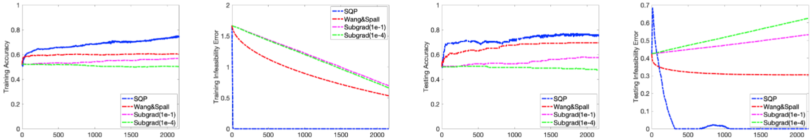


Fig. 5.5. *CPU time versus training accuracy, training infeasibility error, testing accuracy, and testing infeasibility error for a representative run of SQP, Wang & Spall, subgradient* (10^{-1}), *and subgradient* (10^{-4}) *with the* **German** *data set.*

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

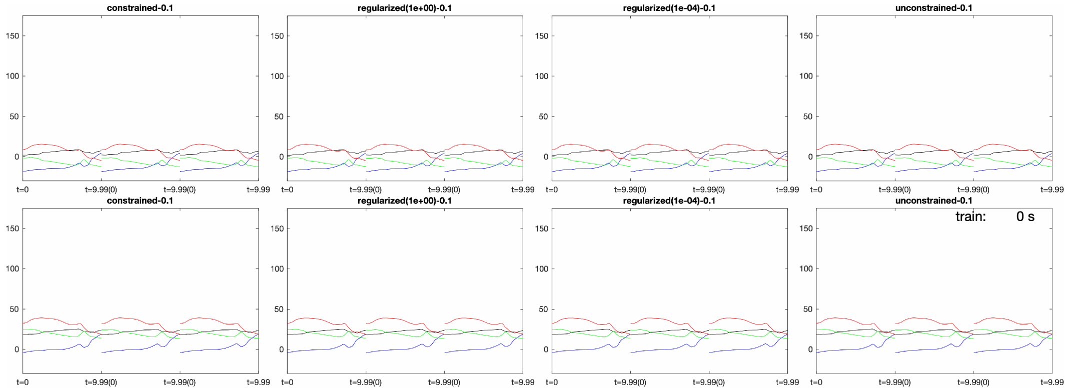
Set $q_k \leftarrow \beta_2 q_{k-1} + (1-\beta_2)(\bar{g}_k \circ \bar{g}_k)$, where $(\bar{g}_k \circ \bar{g}_k)_i = (\bar{g}_k)_i^2$ for all $i \in \{1,\ldots,d\}$

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

Mass-balance



Outline

Summary

Stochastic-gradient/Newton-based algorithms for constrained optimization.

- A lot of work so far, but many open questions.

Open questions:

- stochastic interior-point methods (generally constrained)?
- tradeoff analysis (Bottou and Bousquet)?
- generalization guarantees?
- beyond projected ADAM, etc.?
- Lagrange multiplier estimators?
- active-set identification?
- expectation/probabilistic constraints?

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, 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.

Thank you!

Questions?

Outline

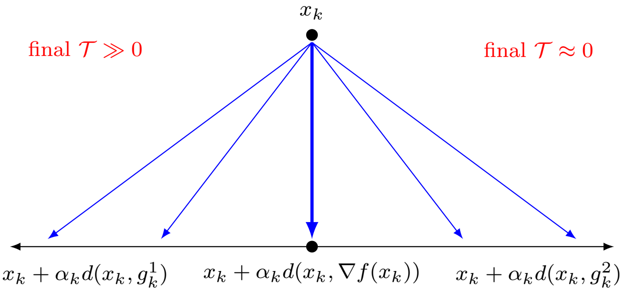
Details

Back

Some details on the tree construction for our complexity analysis...

Challenge in the stochastic setting

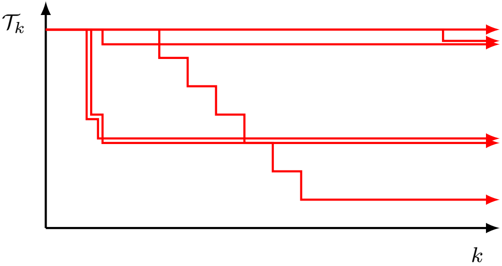
We are minimizing a function that is changing during the optimization.



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 all realizations, 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

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

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

Chernoff bound

How do we get there?

Lemma (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