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

Frank E. Curtis, Lehigh University

presented at

UPenn Optimization Seminar

April 3, 2025

Outline

Outline

Supervised Learning

Expected/empirical risk minimization:

- feature vector X defined over $\mathcal{X}$
- label Y defined over $\mathcal{Y}$
- (X, Y) defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$

Given a prediction function $p : \mathcal{X} \times \mathbb{R}^d \to \mathcal{Y}$ and loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, solve

$$\min_{w \in \mathbb{R}^d} \int_{\mathcal{X} \times \mathcal{Y}} \ell(p(x, w), y) d\mathbb{P}(x, y) \approx \min_{w \in \mathbb{R}^d} \frac{1}{N} \sum_{i=1}^{N} \ell(p(x_i, w), y_i),$$

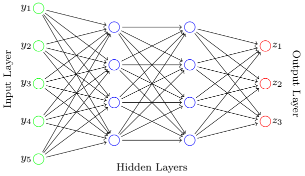
where $\{(x_i, y_i)\}_{i=1}^N$ is a set of sample feature-label pairs.

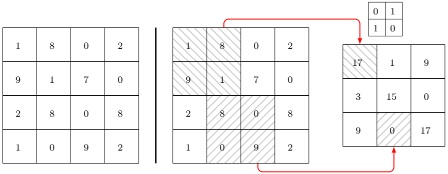
Training faster/better: Choice of data, p, ℓ, and optimization algorithm.

Prediction and loss functions

These are critical, but not my scope. Related to today's talk:

- Simple, classical models $\iff$ enormous, fully connected, overparameterized ones
- The prediction function model/architecture constrains the search
- . . . but there are other ways.





Constrained training/optimization

Constraints can be used to influence training.

- One option is to embed constraints within the prediction function p
- ...e.g., a layer defining p involves solving equations or an optimization problem.
- These remain with every forward pass after the model is trained.

Another option is to impose constraints during training $\Rightarrow$ constrained optimization.

- p constrains the search for a model
- ...additional constraints (data-driven?) refine it further.
- *These constraints can also greatly influence training algorithm behavior!*

Note: This is already done with fine-tuning, e.g., over subspaces, low-rank changes, etc.

Aside: Constrained optimization

Let's simplify notation to focus on the optimization algorithm:

$$\int_{\mathcal{X}\times\mathcal{Y}} \ell(p(x,w),y)\mathrm{d}\mathbb{P}(x,y) =: f(w)$$

Generally, one might consider various paradigms for imposing the constraints:

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

For now, assume constraint values and derivatives can be computed:

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

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

Aside: Penalization

Suppose that $f : \mathbb{R}^d \to \mathbb{R}$, $c_{\mathcal{E}} : \mathbb{R}^d \to \mathbb{R}^{m_{\mathcal{E}}}$, and $c_{\mathcal{I}} : \mathbb{R}^d \to \mathbb{R}^{m_{\mathcal{I}}}$ are locally Lipschitz and consider

$$\min_{w \in \mathbb{R}^d} \ f(w) \ \ \text{s.t.} \ \ c_{\mathcal{E}}(w) = 0 \ \ \text{and} \ \ c_{\mathcal{I}}(w) \leq 0.$$

Two common, essentially equivalent ways of solving such a problem:

- *move* constraints to objective and use an unconstrained method to solve

$$\min_{w \in \mathbb{R}^d} \ f(w) + \lambda v(w) \ \ \text{e.g.} \ \ v(w) = \|c_{\mathcal{E}}(w)\| + \|\max\{c_{\mathcal{I}}(w), 0\}\|$$

- employ a penalty or augmented Lagrangian method

One can refer to this as *penalization*, *regularization*, *soft constraints*, etc.

Aside: Calmness and exact penalization

$$\min_{w \in \mathbb{R}^d} f(w) \quad \text{s.t.} \quad c_{\mathcal{E}}(w) = 0 \;\text{ and }\; c_{\mathcal{I}}(w) \leq 0 \tag{P}$$

Definition : Calmness

Problem (P) is calm at $w \in \mathbb{R}^d$ with respect to $\|\cdot\|$ if and only if there exist $(\epsilon, \delta) \in (0, \infty) \times (0, \infty)$ such that, for all $(\overline{w}, s) \in \mathbb{R}^d \times \mathbb{R}^d_{\geq 0}$ with $\|\overline{w} - w\| \leq \epsilon$, $\|s\| \leq \epsilon$, $-s \leq c_{\mathcal{E}}(\overline{w}) \leq s$, and $c_{\mathcal{I}}(\overline{w}) \leq s$, one has

$$f(\overline{w}) + \delta\|s\| \geq f(w).$$

Theorem : Exact penalization

Suppose $w_* \in \mathbb{R}^d$ is a local minimizer of (P), $v : \mathbb{R}^d \to \mathbb{R}$ is defined by $\|c_{\mathcal{E}}(w)\| + \|\max\{c_{\mathcal{I}}(w), 0\}\|$, and (P) is calm at w_* with respect to $\|\cdot\|$. Then, for some $\lambda_* \in (0, \infty)$, the point w_* is a local minimizer of

$$f + \lambda v \quad \textit{for all} \quad \lambda \in [\lambda_*, \infty).$$

Motivation

It is a mistake to overemphasize the relevance of this theory for practical use.

- Exact penalization only applies for minimizers
- . . . and requires a parameter that cannot be known in advance.
- In practice, subject to a computational budget, a minimizer is not reached
- . . . and the use of stochastic algorithms makes the theory even less relevant.

Penalization/regularization/soft-constraints can cause *slow* progress far from a minimizer.

Overall, our aim in this talk is to convince you that:

- It is worthwhile to explore the use of constrained optimization for informed learning.
- Penalization is not often the best route; there are other/better algorithms to consider.

Outline

Motivation

Stochastic SQP

Extensions

Conclusion

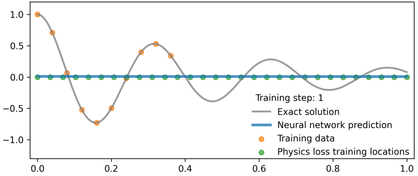
Equality-constrained example

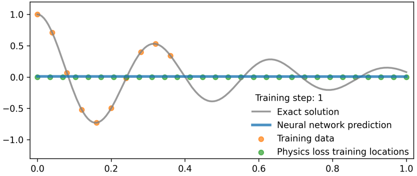
Consider the problem to learn the solution of a parametric partial differential equation (PDE):

- $\mathcal{P}(\phi, u) = 0$, where ϕ are parameters and u solves the PDE with respect to ϕ
- $\mathcal{G}(\phi, y, w)$ predicts u, where y encodes PDE domain and w are trainable parameters
- $\{(\phi_i, y_i, u_i)\}_{i\in\mathcal{S}_1}$ and $\{(\phi_i, y_i)\}_{i\in\mathcal{S}_2}$ are datasets

Our training problem involves (at least) two possible terms:

$$\frac{1}{|\mathcal{S}_1|}\sum_{i\in\mathcal{S}_1}\|u_i - \mathcal{G}(\phi_i, y_i, w)\|^p \qquad \text{and/or} \qquad \frac{1}{|\mathcal{S}_2|}\sum_{i\in\mathcal{S}_2}\|\mathcal{P}(\phi_i, \mathcal{G}(\phi_i, y_i, w))\|^q$$





Problem from https://benmoseley.blog/blog/, $m\,\frac{d^2u(t)}{dt^2} + \mu\,\frac{du(t)}{t} + ku(t) = 0$

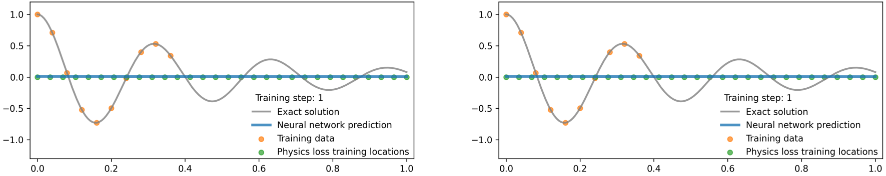
Equality-constrained example

Consider the problem to learn the solution of a parametric partial differential equation (PDE):

- $\mathcal{P}(\phi, u) = 0$, where ϕ are parameters and u solves the PDE with respect to ϕ
- $\mathcal{G}(\phi, y, w)$ predicts u, where y encodes PDE domain and w are trainable parameters
- $\{(\phi_i, y_i, u_i)\}_{i \in \mathcal{S}_1}$ and $\{(\phi_i, y_i)\}_{i \in \mathcal{S}_2}$ are datasets

Our training problem involves (at least) two possible terms:

$$\frac{1}{|\mathcal{S}_1|} \sum_{i \in \mathcal{S}_1} \|u_i - \mathcal{G}(\phi_i, y_i, w)\|^p \qquad \text{and/or} \qquad \mathcal{P}(\phi_i, \mathcal{G}(\phi_i, y_i, w)) = 0$$



Problem from https://benmoseley.blog/blog/, $m\frac{d^2u(t)}{dt^2} + \mu\frac{du(t)}{t} + ku(t) = 0$

Inequality-constrained example

Suppose that one wants the covariance between a feature and the prediction to be limited by ϵ:

$$\min_{w \in \mathbb{R}^d} \frac{1}{|\mathcal{S}_1|} \sum_{(x_i,y_i)\in\mathcal{S}_1} \ell(p(x_i,w),y_i) \quad \text{s.t.} \quad -\epsilon \leq \frac{1}{|\mathcal{S}_2|} \sum_{(x_i,y_i)\in\mathcal{S}_2} (a_i - \overline{a})p(x_i,w) \leq \epsilon$$

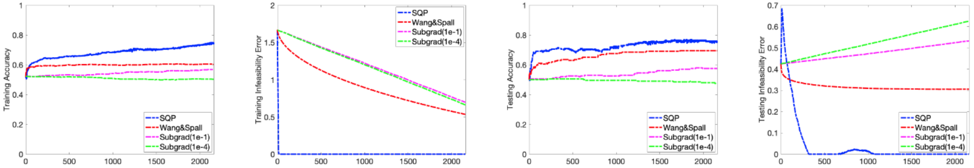


FIG. 5.5. *CPU time versus training accuracy, training infeasibility error, testing accuracy, and testing infeasibility error for a representative run of SQP, Wang & Spall, subgradient* (10^{-1}), *and subgradient* (10^{-4}) *with the* **German** *data set.*

Stochastic gradient method

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

Algorithm SG : Stochastic gradient method

1: choose an initial point $w_1 \in \mathbb{R}^n$ and step sizes $\{\alpha_k\} > 0$
2: **for** $k \in \{1, 2, \dots\} =: \mathbb{N}$ **do**
3: set $w_{k+1} \leftarrow w_k - \alpha_k g_k$, where $g_k \approx \nabla f(w_k)$
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 $\{(w_k, g_k)\}$ as a realization of $\{(W_k, G_k)\}$, where for all $k \in \mathbb{N}$

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

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

Convergence of SG

Let $\mathbb{E}[\cdot]$ = 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

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

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

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

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

SG theory

Taking total expectation, one arrives at

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

Theorem

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

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

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

$$\textit{(further steps)} \quad \Longrightarrow \nabla f(W_k) \to \infty \textit{ almost surely.}$$

Sequential quadratic optimization (SQP)

Consider

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

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

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

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

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

Stochastic SQP

Algorithm guided by merit function with adaptive parameter τ defined by

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

Algorithm : Stochastic SQP

choose $w_1 \in \mathbb{R}^n$, $\tau_0 \in (0, \infty)$, $\{\beta_k\} \in (0, 1]^{\mathbb{N}}$

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

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

compute step size: set

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

then $w_{k+1} \leftarrow w_k + \alpha_k d_k$

end for

Convergence theory in *deterministic setting*

Assumption

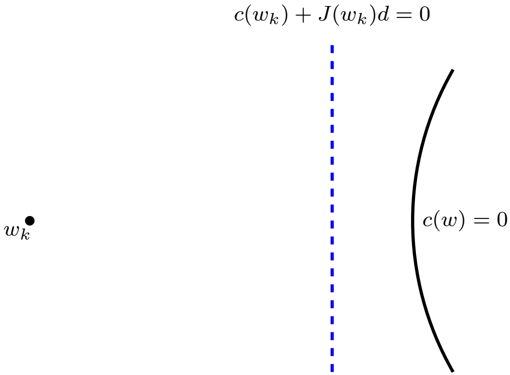
- *f, c, ∇f, and J bounded and Lipschitz*
- *singular values of J bounded away from zero*
- $u^T H_k u \geq \zeta \|u\|_2^2$ *for all* $u \in \text{Null}(J_k)$ *for all* $k \in \mathbb{N}$

Theorem

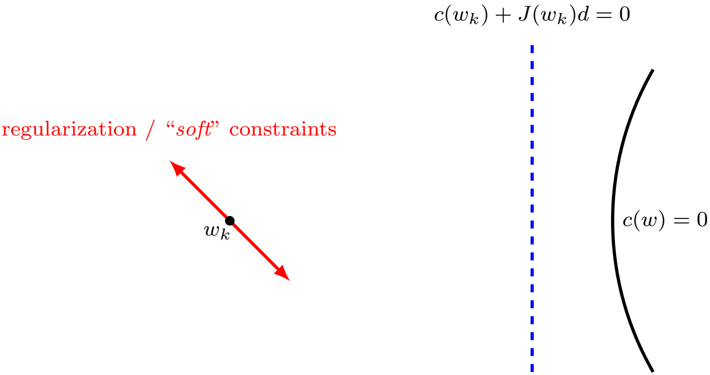
- $\{\alpha_k\} \geq \alpha_{\min}$ *for some* $\alpha_{\min} > 0$
- $\{\tau_k\} \geq \tau_{\min}$ *for some* $\tau_{\min} > 0$
- $\Delta q(w_k, \tau_k, \nabla f(w_k), d_k) \to 0$ *implies optimality error vanishes, specifically,*

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

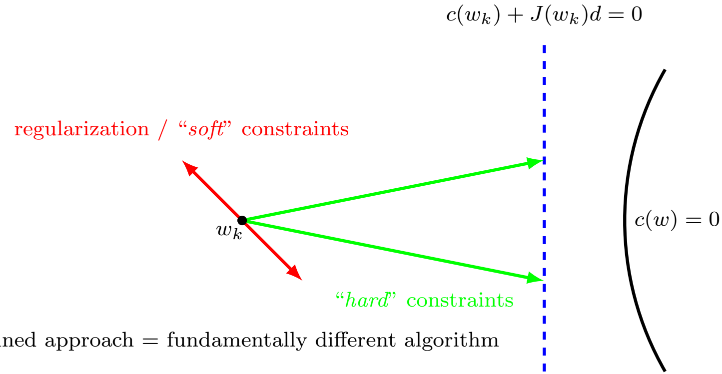
SQP illustration



SQP illustration



SQP illustration

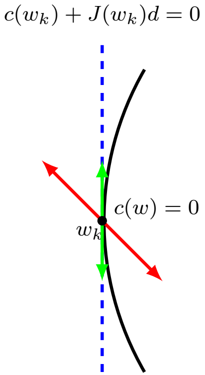


constrained approach = fundamentally different algorithm

SQP illustration

regularization / "*soft*" constraints

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



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

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

Lemma

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

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

Good merit parameter behavior

Theorem 6

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

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

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

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

Good merit parameter behavior

Theorem 6

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

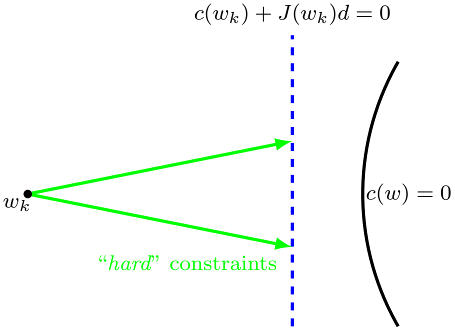
Then, conditioned on $\mathcal{E}$,

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

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

Key observation

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



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

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

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

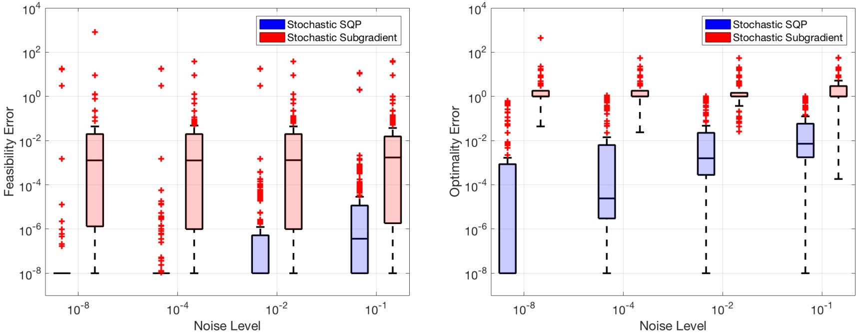


Figure: Box plots for feasibility errors (left) and optimality errors (right).

Projected Adam

Algorithm P-Adam Projection-based Adam

Require: $\beta_1 \in (0,1)$, $\beta_2 \in (0,1)$, $\mu \in \mathbb{R}_{>0}$

Compute $\bar{g}_k \leftarrow (I - J_k^T (J_k J_k^T)^{-1} J_k) g_k$ (comes "for free" if computing v_k explicitly)

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

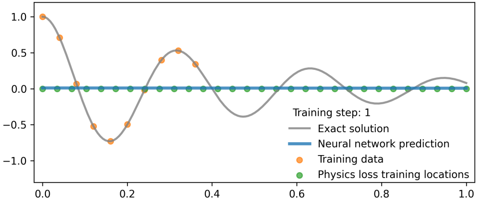
Set $q_k \leftarrow \beta_2 q_{k-1} + (1-\beta_2)(\bar{g}_k \circ \bar{g}_k)$, where $(\bar{g}_k \circ \bar{g}_k)_i = (\bar{g}_k)_i^2$ for all $i \in \{1,\dots,d\}$

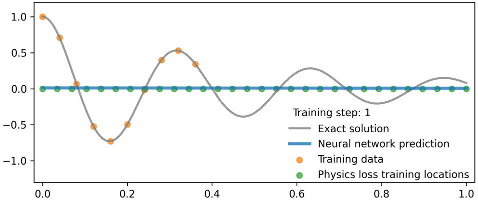
Set $\widehat{p}_k \leftarrow (1/(1-\beta_1^k))p_k$

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

Compute d_k by solving $\begin{bmatrix} \text{diag}(\sqrt{\widehat{q}_k + \mu}) & J_k^T \\ J_k & 0 \end{bmatrix} \begin{bmatrix} d_k \\ y_k \end{bmatrix} = -\begin{bmatrix} \widehat{p}_k \\ c_k \end{bmatrix}$

Accelerated performance with P-Adam





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

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

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

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

Lemma

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

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

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

Almost-sure convergence of the primal iterates

Theorem

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

$$w \in \mathcal{W}_{\epsilon,w_*} := \{w \in \mathcal{W} : \|w - w_*\|_2 \leq \epsilon\}$$

one finds that

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

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

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

Lagrange multiplier convergence

Theorem

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

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

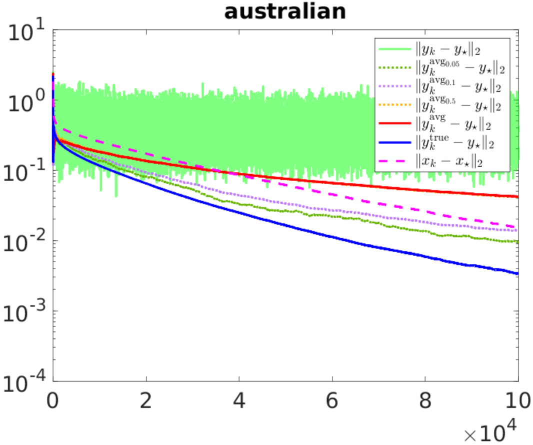
Computed multipliers *always* have error. Consider *averaged* multipliers $\{Y_k^{\text{avg}}\}$:

Theorem

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

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

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



Outline

Summary

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

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

Open questions:

- tradeoff analysis (Bottou and Bousquet)?
- generalization guarantees?
- beyond projected ADAM, etc.?
- Lagrange multiplier estimators for inequality-constrained setting?
- active-set identification?
- expectation/probabilistic constraints?

Constraint engineering

Neural network engineering, feature engineering, and now *constraint engineering*. . .

- The number of constraints m can be controlled:

$$\left.\begin{array}{r} c(p(x_1,w),y_1)=0 \\ c(p(x_2,w),y_2)=0 \\ \vdots \end{array}\right\} \quad vs. \quad \frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}} c(p(x_i,w),y_i)=0.$$

- Selection of constraint data $\{(x_i,y_i)\}_{i\in\mathcal{S}}$ also requires some care.

In all cases, also due to "vanishing gradients" and other possible effects, beware rank-deficient Jacobians:

- Berahas, Curtis, O'Neill, Robinson (2023)

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, "A Stochastic Inexact Sequential Quadratic Optimization Algorithm for Nonlinear Equality-Constrained Optimization," *INFORMS Journal on Optimization*, , 2024.
- 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," *SIAM Journal on Optimization*, 34(4):3592–3622, 2024.
- F. E. Curtis, X. Jiang, and Q. Wang, "Almost-sure convergence of iterates and multipliers in stochastic sequential quadratic optimization," *Journal of Optimization Theory and Applications*, https://rdcu.be/d5OwU, 2024.
- 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," to appear in *SIAM Journal on Optimization*, https://arxiv.org/abs/2304.14907.
- F. E. Curtis, X. Jiang, and Q. Wang, "Single-Loop Deterministic and Stochastic Interior-Point Algorithms for Nonlinearly Constrained Optimization," https://arxiv.org/abs/2408.16186.

Questions?