Stochastic-Gradient-based Algorithms for Nonconvex Constrained Optimization and Learning

Frank E. Curtis, Lehigh University

presented at

AlgoPerf Workshop

Meta / ML Commons

February 12, 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 the 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

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

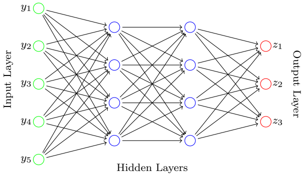
where $\{(x_i, y_i)\}_{i=1}^{N}$ is a set of sample feature-label pairs.

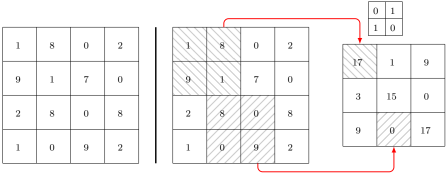
Training faster/better: Choice of p, ℓ, and optimization algorithm.

Prediction and loss functions

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

- Simple, classical models $\Longleftrightarrow$ 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: In some sense 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

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

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 $\|\overline{w} - w\| \leq \epsilon$, $\|s\| \leq \epsilon$, $-s \leq c_{\mathcal{E}}(w) \leq s$, 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 the appropriate route; there are other/better algorithms to consider.

Outline

- Motivation
- Stochastic SQP
- Extensions
- Conclusion

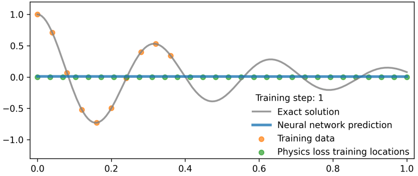
Equality-constrained example

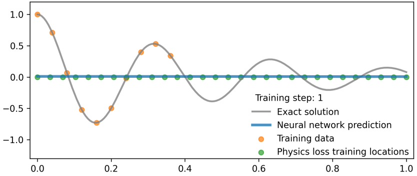
Consider the problem to learn the solution of a parametric partial differential equation (PDE):

- $\mathcal{P}(\phi, u) = 0$, 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/, $m \frac{d^2 u(t)}{dt^2} + \mu \frac{du(t)}{t} + k u(t) = 0$

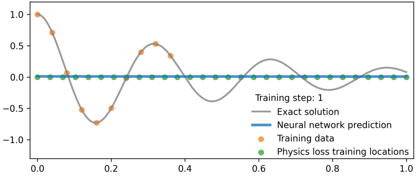
Equality-constrained example

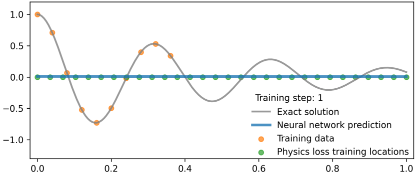
Consider the problem to learn the solution of a parametric partial differential equation (PDE):

- $\mathcal{P}(\phi, u) = 0$, 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:

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

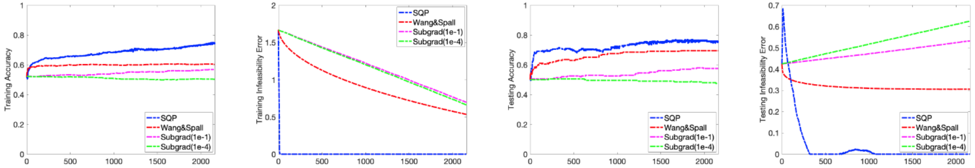
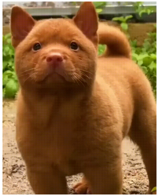


FIG. 5.5. *CPU time versus training accuracy, training infeasibility error, testing accuracy, and testing infeasibility error for a representative run of SQP, Wang & Spall, subgradient* (10^{-1})*, and subgradient* (10^{-4}) *with the* **German** *data set.*

Other examples

Ideas (tested and untested):

- $\frac{dp}{da}(x_i, w) \leq 0 \equiv$ change in predicted value w/ change in input
- $\ell(p(x_i, w), y_i) < \ell(p(x_j, w), y_j) \equiv$ difference in loss
- $\frac{d\ell}{da}(p(x_i, w), y_i) \leq 0 \equiv$ change in loss w/ change in input



Stochastic SQP (equality constraints only, $c(w) = 0$)

Algorithm : Stochastic gradient (w/ diagonal scaling, e.g., ADAM)

choose $w_1 \in \mathbb{R}^d$
for $k \in \{1, 2, \dots\}$ **do**
  set scaling: compute stochastic gradient g_k, choose symmetric positive definite $H_k \in \mathbb{R}^{d \times d}$
  compute step: solve $H_k s_k = -g_k$
  update iterate: set $w_{k+1} \leftarrow w_k + \alpha_k s_k$, where $\alpha_k = \Theta\left(\frac{\beta_k}{L_{\nabla f}}\right)$
end for

Algorithm : Stochastic SQP

choose $w_1 \in \mathbb{R}^d$
for $k \in \{1, 2, \dots\}$ **do**
  set scaling: compute stochastic gradient g_k, choose symmetric positive definite $H_k \in \mathbb{R}^{d \times d}$
  compute step: solve $\begin{bmatrix} H_k & \nabla c(w_k)^T \\ \nabla c(w_k) & 0 \end{bmatrix} \begin{bmatrix} s_k \\ y_k \end{bmatrix} = -\begin{bmatrix} g_k \\ c(w_k) \end{bmatrix}$ (includes $c(w_k) + \nabla c(w_k) s_k = 0$)
  update iterate: set $w_{k+1} \leftarrow w_k + \alpha_k s_k$, where $\alpha_k = \Theta\left(\frac{\beta_k \tau_k}{L_{\nabla f} \tau_k + L_{\nabla c}}\right)$
end for

Fundamental lemma

A fundamental lemma in the analysis of the stochastic gradient method:

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

Lemma

For all $k \in \mathbb{N}$, the change in the merit function ϕ satisfies (before taking expectations)

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

Good merit parameter behavior

For a stochastic gradient method, the fundamental lemma allows one to show that

$$\beta_k = \Theta(1) \implies \mathbb{E}\left[\frac{1}{k}\sum_{j=1}^{k} \|\nabla f(W_j)\|_2^2\right] = \mathcal{O}(\text{constant})$$

$$\beta_k = \Theta\left(\frac{1}{k}\right) \implies \mathbb{E}\left[\frac{1}{\left(\sum_{j=1}^k \beta_j\right)}\sum_{j=1}^{k} \beta_j \|\nabla f(W_j)\|_2^2\right] \to 0 \qquad \left(\text{yields } \liminf_{k\to\infty} \mathbb{E}\left[\|\nabla f(W_j)\|_2^2\right] = 0\right)$$

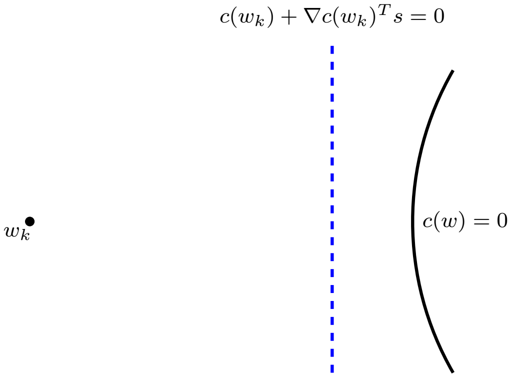
Theorem : Berahas, Curtis, Robinson, Zhou (2021)

Let $\mathcal{E} :=$ event that $\{\mathcal{T}_k\}$ eventually remains constant at $\mathcal{T} \geq \tau_{\min} > 0$. Then, conditioned on $\mathcal{E}$:

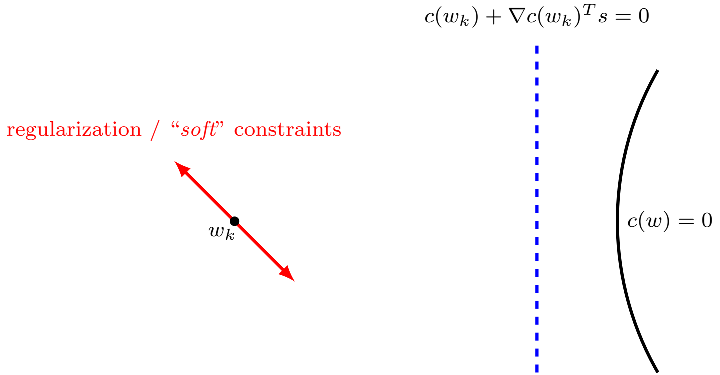
$$\beta_k = \Theta(1) \implies \mathbb{E}\left[\frac{1}{k}\sum_{j=1}^{k} (\|\nabla f(W_j) + \nabla c(W_j) Y_j^{\text{true}}\|_2^2 + \|c(W_j)\|_2)\right] = \mathcal{O}(constant)$$

$$\beta_k = \Theta\left(\frac{1}{k}\right) \implies \mathbb{E}\left[\frac{1}{\left(\sum_{j=1}^k \beta_j\right)}\sum_{j=1}^{k} \beta_j (\|\nabla f(W_j) + \nabla c(W_j) Y_j^{\text{true}}\|_2^2 + \|c(W_j)\|_2)\right] \to 0$$

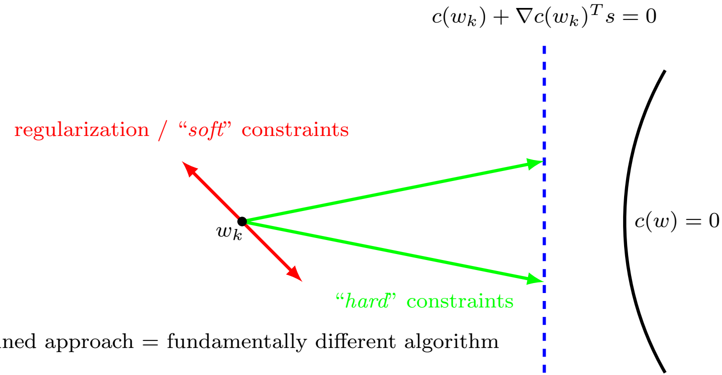
SQP illustration



SQP illustration



SQP illustration

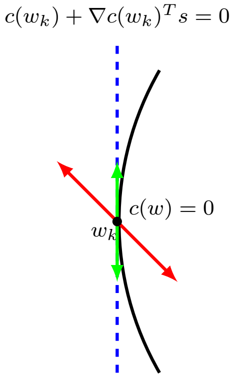


constrained approach = fundamentally different algorithm

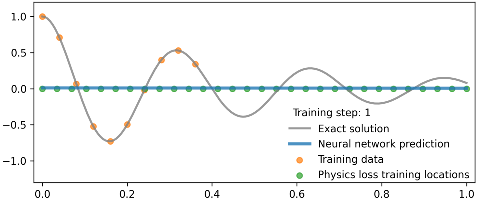
SQP illustration

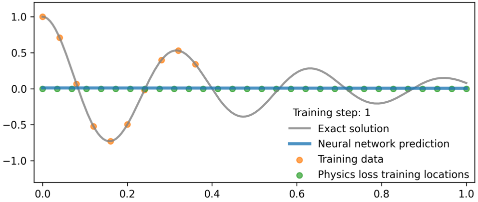
regularization / "*soft*" constraints

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



Accelerated performance





Computational costs

Solve a system with $\begin{bmatrix} H_k & \nabla c(w_k)^T \\ \nabla c(w_k) & 0 \end{bmatrix} \in \mathbb{R}^{(d+m)\times(d+m)}$?!

Direct solves

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

Important notes:

- The number of constraints m can be very small (more on this later).
- H_k can also have nice structure. Let's say (block) *diagonal*.
- $s_k = v_k + u_k$ is the only part needed (usually), where
- ... v_k is the step to the linearized constraints and
- ... u_k is the unique H_k-orthogonal projection of $g_k + H_k v_k$ onto Null(J_k)

$$v_k = -J_k^T \underbrace{(J_k J_k^T)^{-1}}_{m \times m} c_k \quad \text{and} \quad u_k = -(I - \underbrace{H_k^{-1}}_{diag} J_k^T \underbrace{(J_k H_k^{-1} J_k^T)}_{m \times m}{}^{-1} J_k) \underbrace{H_k^{-1}}_{diag} (g_k + H_k v_k)$$

Total cost: $\mathcal{O}(m^2 d + m^3)$

Iterative solves

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

Large sparse indefinite system:

- Iterative linear system solvers based on Lanczos process, building Krylov subspaces
- MINRES, SYMMLQ, preconditioning techniques, etc.
- Eigenvalues cluster nicely, few iterations needed
- Allow inexact solutions! Curtis, Robinson, Zhou (2024)

Constraint preconditioning, factorization reuse

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

Suppose one has a factorization of $\begin{bmatrix} H & J^T \\ J & 0 \end{bmatrix}$, where $H \approx H_k$ and $J \approx J_k$.

- Effective as a preconditioner for an iterative linear system solver ("constraint preconditioner")
- ... Keller, Gould, Wathen (2000)
- Can also simply reuse factorization over multiple steps ("lagged Newton")
- ... Shamanskii (1967); Brown, Brune (2013)
- Similarly, could reuse factorizations for *reduced-space* approach mentioned earlier

Diagonal scaling matrix

What choice for H_k in the constraint setting?

- Typical scaling (e.g., Adam) uses only information from $\{g_k\}$
- Anything different with constraints?

Yes! **Idea**: Avoid accounting for components of $\{g_k\}$ *off* of constraints.

- The normal step $v_k = -J_k^T(J_k J_k^T)^{-1} c_k$ is unaffected by H_k.
- However, the tangential step (in $\mathrm{Null}(J_k)$) is affected:

$$
\begin{aligned}
u_k &= -(I - H_k^{-1} J_k^T (J_k H_k^{-1} J_k^T)^{-1} J_k) H_k^{-1} (g_k + H_k v_k) \\
&= -Z_k (Z_k^T H_k Z_k)^{-1} Z_k^T (g_k + H_k v_k)
\end{aligned}
$$

Idea: To build H_k, project out component of g_k that lies in $\mathrm{Range}(J_k^T)$.

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

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

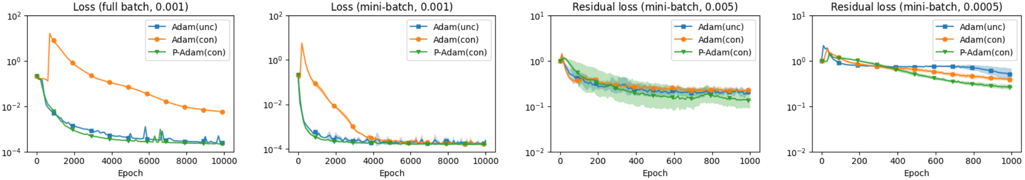
Set $q_k \leftarrow \beta_2 q_{k-1} + (1-\beta_2)(\bar{g}_k \circ \bar{g}_k)$, where $(\bar{g}_k \circ \bar{g}_k)_i = (\bar{g}_k)_i^2$ for all $i \in \{1,\dots,d\}$

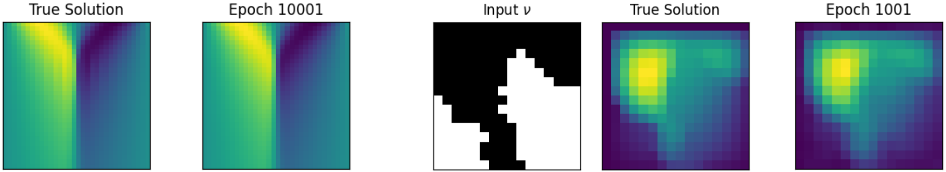
Set $\widehat{p}_k \leftarrow (1/(1-\beta_1^k))p_k$

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

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

Burgers Equation and Darcy Flow





Outline

Motivation

Stochastic SQP

Extensions

Conclusion

Summary

Since our original work, we have considered various extensions.

- iterative linear system solvers and inexactness
- diagonal scaling methods for saddle-point systems
- 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

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 $\|X_k - x_*\|_2 \le \epsilon$ implies

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

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

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

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* $1-\delta$,

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

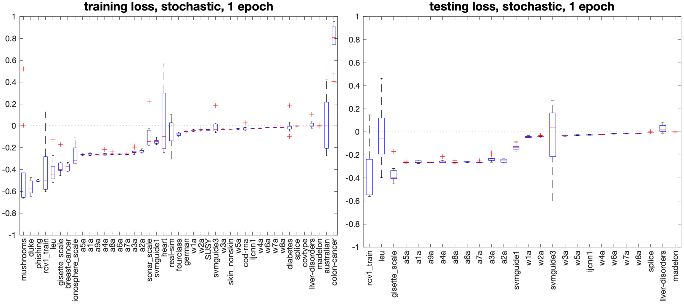
Theorem

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

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

Stochastic-gradient-based interior-point method

Single-loop interior-point (SLIP) method: barrier parameter $\{\mu_k\}$ vanishes by prescribed rate.



Relative performance of SLIP and PSGM, stochastic setting (10 runs each), training neural network models (with one hidden layer) with cross-entropy loss; among 43 training datasets, 26 have testing datasets.

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?
- 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:

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

- 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?