Self-Correcting Variable-Metric Algorithms

Frank E. Curtis, Lehigh University

involving joint work with

Daniel P. Robinson, Johns Hopkins University

Workshop on Nonlinear Optimization Algorithms and Industrial Applications
Fields Institute, Toronto, Ontario, Canada

4 June 2016

Outline

Outline

Unconstrained optimization

Consider unconstrained optimization problems of the form

$$\min_{x \in \mathbb{R}^n} \; f(x).$$

Deterministic, smooth

- gradient $\rightarrow$ Newton methods

Stochastic, smooth

- stochastic gradient $\rightarrow$ batch Newton methods

Deterministic, nonsmooth

- subgradient $\rightarrow$ bundle / gradient sampling methods

Balance between extremes

For deterministic, smooth optimization, a nice balance achieved by quasi-Newton:

$$x_{k+1} \leftarrow x_k - \alpha_k W_k g_k,$$

where

- $\alpha_k > 0$ is a stepsize;
- $g_k \leftarrow \nabla f(x_k)$ (or an approximation of it);
- $\{W_k\}$ is updated dynamically.

We all know:

- local rescaling based on iterate/gradient displacements
- only first-order derivatives required
- no linear system solves required
- global convergence guarantees (say, with line search)
- superlinear local convergence rate

How can we carry these ideas to other settings?

Issues for Enhancements

Convex to nonconvex

- Positive definiteness not maintained automatically $(\star)$

Deterministic to stochastic

- $(\star)$ and scaling matrices not independent from gradients $(W_k g_k)$

Smooth to nonsmooth

- Scaling matrices tend to singularity

Issues for Enhancements: Proposed Solutions

Convex to nonconvex

- Positive definiteness not maintained automatically $(\star)$
- Skipping, damping

Deterministic to stochastic

- $(\star)$ and scaling matrices not independent from gradients $(W_k g_k)$
- Skipping, damping, regularization

Smooth to nonsmooth

- Scaling matrices tend to singularity
- (Wolfe) line search, bundles or gradient sampling

Issues for Enhancements: Proposed Solutions: Remaining Issues

Convex to nonconvex

- Positive definiteness not maintained automatically ($\star$)
- Skipping, damping
- poor performance from skipping or under-/over-damping ($\star\star$)

Deterministic to stochastic

- ($\star$) and scaling matrices not independent from gradients ($W_k g_k$)
- Skipping, damping, regularization
- ($\star\star$) and over-regularization (e.g., adding δI to all updates)

Smooth to nonsmooth

- Scaling matrices tend to singularity
- (Wolfe) line search, bundles or gradient sampling
- intertwined $\{x_k\}$, $\{\alpha_k\}$, $\{g_k\}$, and $\{W_k\}$

Overview

Propose two methods for unconstrained optimization

- exploit self-correcting properties of BFGS-type updates
  - Powell (1976); Ritter (1979, 1981); Werner (1978); Byrd, Nocedal (1989)
- properties of Hessians offer useful bounds for inverse Hessians
- forget about superlinear convergence,

$$\lim_{k\to\infty} \frac{\|(H_k - H_*)s_k\|_2}{\|s_k\|_2} = 0 \quad \text{(not relevant here!)}$$

Overview

Propose two methods for unconstrained optimization

- exploit self-correcting properties of BFGS-type updates
  - Powell (1976); Ritter (1979, 1981); Werner (1978); Byrd, Nocedal (1989)
- properties of Hessians offer useful bounds for inverse Hessians
- forget about superlinear convergence,

$$\lim_{k\to\infty} \frac{\|(H_k - H_*)s_k\|_2}{\|s_k\|_2} = 0 \quad \text{(not relevant here!)}$$

Stochastic, nonconvex:

- Proposal: Twist on updates, different than others proposed
- Result: More stable behavior than basic stochastic quasi-Newton

Deterministic, nonsmooth:

- Proposal: Generic algorithmic framework enjoying self-correcting properties
- Result: Improved performance(?), guide for convergence for other methods

Outline

BFGS-type updates

Inverse Hessian and Hessian approximation[1] updating formulas ($s_k^T v_k > 0$):

$$W_{k+1} \leftarrow \left(I - \frac{v_k s_k^T}{s_k^T v_k}\right)^T W_k \left(I - \frac{v_k s_k^T}{s_k^T v_k}\right) + \frac{s_k s_k^T}{s_k^T v_k}$$

$$H_{k+1} \leftarrow \left(I - \frac{s_k s_k^T H_k}{s_k^T H_k s_k}\right)^T H_k \left(I - \frac{s_k s_k^T H_k}{s_k^T H_k s_k}\right) + \frac{v_k v_k^T}{s_k^T v_k}$$

- These satisfy secant-type equations

$$W_{k+1} v_k = s_k \quad \text{and} \quad H_{k+1} s_k = v_k,$$

  but these are not very relevant for this talk.
- Choosing $v_k \leftarrow y_k := g_{k+1} - g_k$ yields standard BFGS, but we consider

$$v_k \leftarrow \beta_k s_k + (1 - \beta_k)\alpha_k y_k \quad \text{for some} \quad \beta_k \in [0, 1].$$

  This scheme is important to preserve self-correcting properties.

[1] "Hessian" and "inverse Hessian" used loosely in nonsmooth settings

Geometric properties of Hessian update: Burke, Lewis, Overton (2007)

Consider the matrices (which only depend on s_k and H_k, not g_k!)

$$P_k := \frac{s_k s_k^T H_k}{s_k^T H_k s_k} \quad \text{and} \quad Q_k := I - P_k.$$

Both H_k-orthogonal projection matrices (i.e., idempotent and H_k-self-adjoint).

- P_k yields H_k-orthogonal projection onto $\text{span}(s_k)$.
- Q_k yields H_k-orthogonal projection onto $\text{span}(s_k)^{\perp_{H_k}}$.

Geometric properties of Hessian update: Burke, Lewis, Overton (2007)

Consider the matrices (which only depend on s_k and H_k, not g_k!)

$$P_k := \frac{s_k s_k^T H_k}{s_k^T H_k s_k} \quad \text{and} \quad Q_k := I - P_k.$$

Both H_k-orthogonal projection matrices (i.e., idempotent and H_k-self-adjoint).

- P_k yields H_k-orthogonal projection onto $\text{span}(s_k)$.
- Q_k yields H_k-orthogonal projection onto $\text{span}(s_k)^{\perp_{H_k}}$.

Returning to the Hessian update:

$$H_{k+1} \leftarrow \underbrace{\left(I - \frac{s_k s_k^T H_k}{s_k^T H_k s_k}\right)^T H_k \left(I - \frac{s_k s_k^T H_k}{s_k^T H_k s_k}\right)}_{\text{rank } n-1} + \underbrace{\frac{v_k v_k^T}{s_k^T v_k}}_{\text{rank } 1}$$

- Curvature projected out along $\text{span}(s_k)$
- Curvature corrected by $\frac{v_k v_k^T}{s_k^T v_k} = \left(\frac{v_k v_k^T}{\|v_k\|_2^2}\right)\left(\frac{\|v_k\|_2^2}{v_k^T W_{k+1} v_k}\right)$ (inverse Rayleigh).

Self-correcting properties of Hessian update

Since curvature is constantly projected out, what happens after many updates?

Self-correcting properties of Hessian update

Since curvature is constantly projected out, what happens after many updates?

Theorem 1 (Byrd, Nocedal (1989))

Suppose that, for all k, there exists $\{\eta, \theta\} \subset \mathbb{R}_{++}$ such that

$$\eta \leq \frac{s_k^T v_k}{\|s_k\|_2^2} \quad \text{and} \quad \frac{\|v_k\|_2^2}{s_k^T v_k} \leq \theta. \tag{KEY}$$

Then, for any $p \in (0,1)$, there exist constants $\{\iota, \kappa, \lambda\} \subset \mathbb{R}_{++}$ such that, for any $K \geq 2$, the following relations hold for at least $\lceil pK \rceil$ values of $k \in \{1, \dots, K\}$:

$$\iota \leq \frac{s_k^T H_k s_k}{\|s_k\|_2 \|H_k s_k\|_2} \quad \text{and} \quad \kappa \leq \frac{\|H_k s_k\|_2}{\|s_k\|_2} \leq \lambda.$$

Proof technique.

Building on work of Powell (1976), involves bounding growth of

$$\gamma(H_k) = \text{tr}(H_k) - \ln(\det(H_k)).$$

Self-correcting properties of inverse Hessian update

Rather than focus on superlinear convergence results, we care about the following.

Corollary 2

Suppose the conditions of Theorem 1 hold. Then, for any $p \in (0,1)$, there exist constants $\{\mu, \nu\} \subset \mathbb{R}_{++}$ such that, for any $K \geq 2$, the following relations hold for at least $\lceil pK \rceil$ values of $k \in \{1, \dots, K\}$:

$$\mu \|g_k\|_2^2 \leq g_k^T W_k g_k \quad \textit{and} \quad \|W_k g_k\|_2^2 \leq \nu \|g_k\|_2^2$$

Proof sketch.

Follows simply after algebraic manipulations from the result of Theorem 1, using the facts that $s_k = -\alpha_k W_k g_k$ and $W_k = H_k^{-1}$ for all k.

Outline

Motivation

Self-Correcting Properties of BFGS-type Updating

Stochastic, Nonconvex Optimization

Deterministic, Nonsmooth Optimization

Summary

Stochastic, nonconvex optimization

Consider unconstrained optimization problems of the form

$$\min_{x \in \mathbb{R}^n} \; f(x),$$

where, in an open set containing $\{x_k\}$,

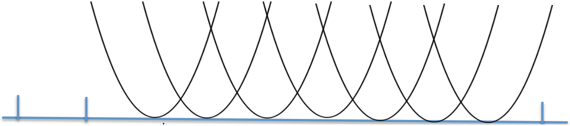
- f is continously differentiable and bounded below and
- ∇f is Lipschitz continuous with constant $L > 0$,

but

- neither f nor ∇f can be computed exactly.

What has been done?

$$H_{k+1} s_k = y_k$$



$$y_k \leftarrow \nabla f(x_{k+1}, \xi_k) - \nabla f(x_k, \xi_k) \quad \text{or} \quad y_k \leftarrow \left(\sum_{\xi_{k+1} \in \Xi_{k+1}} \nabla^2 f(x_{k+1}, \xi_{k+1}) \right) s_k$$

false consistency?

Algorithm VM-DS : Variable-Metric Algorithm with Diminishing Stepsizes

Choose $x_1 \in \mathbb{R}^n$.
Set $g_1 \approx \nabla f(x_1)$.
Choose a symmetric positive definite $W_1 \in \mathbb{R}^{n \times n}$.
Choose a positive scalar sequence $\{\alpha_k\}$ such that

$$\sum_{k=1}^{\infty} \alpha_k = \infty \quad \text{and} \quad \sum_{k=1}^{\infty} \alpha_k^2 < \infty.$$

for $k = 1, 2, \ldots$ **do**
    Set $s_k \leftarrow -\alpha_k W_k g_k$.
    Set $x_{k+1} \leftarrow x_k + s_k$.
    Set $g_{k+1} \approx \nabla f(x_{k+1})$.
    Set $y_k \leftarrow g_{k+1} - g_k$.
    Set $\beta_k \leftarrow \min\{\beta \in [0,1] : v(\beta) := \beta s_k + (1-\beta)\alpha_k y_k$ satisfies (KEY)$\}$.
    Set $v_k \leftarrow v(\beta_k)$.
    Set

$$W_{k+1} \leftarrow \left(I - \frac{v_k s_k^T}{s_k^T v_k}\right)^T W_k \left(I - \frac{v_k s_k^T}{s_k^T v_k}\right) + \frac{s_k s_k^T}{s_k^T v_k}.$$

end for

Global convergence theorem

Theorem 3 (Bottou, Curtis, Nocedal (2016))

Suppose that, for all k, there exists a scalar constant $\rho > 0$ *such that*

$$-\nabla f(x_k)^T \mathbb{E}_{\xi_k}[W_k g_k] \leq -\rho \|\nabla f(x_k)\|_2^2,$$

and there exist scalars $\sigma > 0$ *and* $\tau > 0$ *such that*

$$\mathbb{E}_{\xi_k}[\|W_k g_k\|_2^2] \leq \sigma + \tau \|\nabla f(x_k)\|_2^2.$$

Then, $\{\mathbb{E}[f(x_k)]\}$ *converges to a finite limit and*

$$\liminf_{k \to \infty} \mathbb{E}[\nabla f(x_k)] = 0.$$

Proof technique.

Follows from the critical inequality

$$\mathbb{E}_{\xi_k}[f(x_{k+1})] - f(x_k) \leq -\alpha_k \nabla f(x_k)^T \mathbb{E}_{\xi_k}[W_k g_k] + \alpha_k^2 L \mathbb{E}_{\xi_k}[\|W_k g_k\|_2^2].$$

Reality

The conditions in this theorem cannot be verified in practice.

- They require knowing $\nabla f(x_k)$.
- They require knowing $\mathbb{E}_{\xi_k}[W_k g_k]$ and $\mathbb{E}_{\xi_k}[\|W_k g_k\|_2^2]$
- ... but W_k and g_k are not independent!
- That said, Corollary 2 ensures that they hold with $g_k = \nabla f(x_k)$; recall

$$\mu\|g_k\|_2^2 \leq g_k^T W_k g_k \quad \text{and} \quad \|W_k g_k\|_2^2 \leq \nu\|g_k\|_2^2.$$

Reality

The conditions in this theorem cannot be verified in practice.

- They require knowing $\nabla f(x_k)$.
- They require knowing $\mathbb{E}_{\xi_k}[W_k g_k]$ and $\mathbb{E}_{\xi_k}[\|W_k g_k\|_2^2]$
- ...but W_k and g_k are not independent!
- That said, Corollary 2 ensures that they hold with $g_k = \nabla f(x_k)$; recall

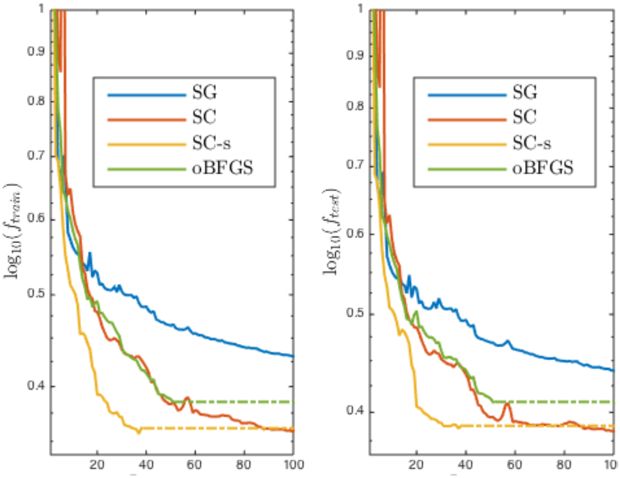
$$\mu \|g_k\|_2^2 \leq g_k^T W_k g_k \quad \text{and} \quad \|W_k g_k\|_2^2 \leq \nu \|g_k\|_2^2.$$

End of iteration k, loop over (stochastic) gradient computation until

$$\rho \|\hat{g}_{k+1}\|_2^2 \leq \hat{g}_{k+1}^T W_{k+1} g_{k+1}$$
$$\text{and} \quad \|W_{k+1} g_{k+1}\|_2^2 \leq \sigma + \tau \|\hat{g}_{k+1}\|_2^2.$$

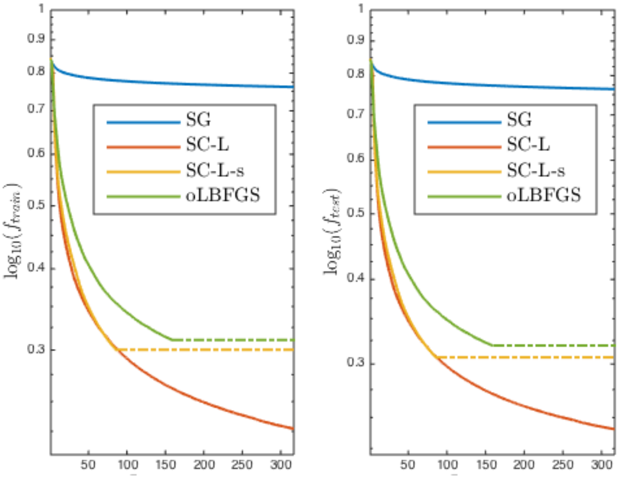
Recompute g_{k+1}, $\hat{g}_{k+1}$, and W_{k+1} until these hold.

Numerical Experiments: a1a



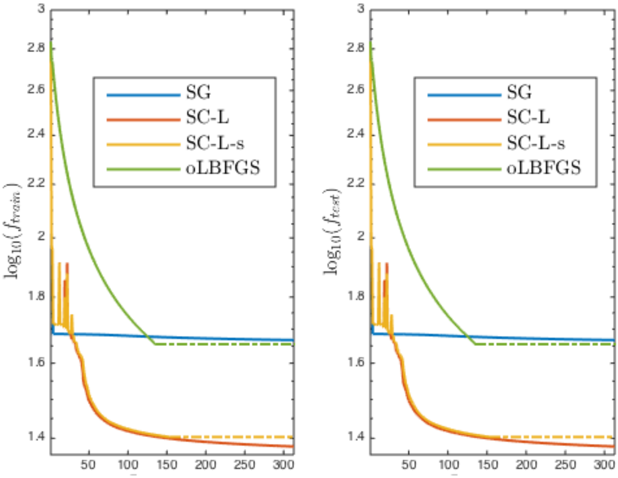
logistic regression, data `a1a`, diminishing stepsizes

Numerical Experiments: rcv1



logistic regression, data `rcv1`, diminishing stepsizes

Numerical Experiments: mnist



deep neural network, data `mnist`, diminishing stepsizes

Outline

Nonsmooth optimization

Consider unconstrained optimization problems of the form

$$\min_{x \in \mathbb{R}^n} \; f(x),$$

where

- f is locally Lipschitz in $\mathbb{R}^n$ and
- differentiable in an open, dense subset of $\mathbb{R}^n$,

but

- nonsmooth.

What has been done?

Many have observed improved performance with quasi-Newton schemes.

"Unadulterated" BFGS

- Lemaráchal (1982)
- Lewis, Overton (2012)

BFGS (with restricted updates)

- Haarala, Miettinen, Mäkelä (2004)
- Curtis, Que (2015)

To our knowledge, none have tried to exploit self-correcting properties.

Subproblems in nonsmooth optimization algorithms

With sets of points, scalars, and (sub)gradients

$$\{x_{k,j}\}_{j=1}^m, \quad \{f_{k,j}\}_{j=1}^m, \quad \{g_{k,j}\}_{j=1}^m,$$

nonsmooth optimization methods involve the primal subproblem

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \ & \left(\max_{j \in \{1,\dots,m\}} \{f_{k,j} + g_{k,j}^T (x - x_{k,j})\} + \tfrac{1}{2}(x - x_k)^T H_k (x - x_k) \right) \\ \text{s.t. } & \|x - x_k\| \leq \delta_k, \end{aligned} \tag{P}$$

but, with $G_k \leftarrow [g_{k,1} \ \cdots \ g_{k,m}]$, it is typically more efficient to solve the dual

$$\begin{aligned} \sup_{(\omega,\gamma) \in \mathbb{R}_+^m \times \mathbb{R}^n} \ & -\tfrac{1}{2}(G_k\omega + \gamma)^T W_k (G_k\omega + \gamma) + b_k^T \omega - \delta_k \|\gamma\|_* \\ \text{s.t. } & \mathbb{1}_m^T \omega = 1. \end{aligned} \tag{D}$$

The primal solution can then be recovered by

$$x_k^* \leftarrow x_k - W_k \underbrace{(G_k \omega_k + \gamma_k)}_{\tilde{g}_k}.$$

Algorithm Self-Correcting BFGS for Nonsmooth Optimization

Choose $x_1 \in \mathbb{R}^n$.
Choose a symmetric positive definite $W_1 \in \mathbb{R}^{n \times n}$.
Choose $\alpha \in (0, 1)$
for $k = 1, 2, \dots$ **do**
    Solve (P)–(D) such that setting

$$\begin{aligned} G_k &\leftarrow \begin{bmatrix} g_{k,1} & \cdots & g_{k,m} \end{bmatrix}, \\ s_k &\leftarrow -W_k(G_k\omega_k + \gamma_k), \\ \text{and} \quad x_{k+1} &\leftarrow x_k + s_k \end{aligned}$$

    yields

$$f(x_{k+1}) \leq f(x_k) - \tfrac{1}{2}\alpha(G_k\omega_k + \gamma_k)^T W_k (G_k\omega_k + \gamma_k).$$

    Choose $y_k \in \mathbb{R}^n$.
    Set $\beta_k \leftarrow \min\{\beta \in [0,1] : v(\beta) := \beta s_k + (1-\beta) y_k$ satisfies (KEY)$\}$.
    Set $v_k \leftarrow v(\beta_k)$.
    Set

$$W_{k+1} \leftarrow \left(I - \frac{v_k s_k^T}{s_k^T v_k}\right)^T W_k \left(I - \frac{v_k s_k^T}{s_k^T v_k}\right) + \frac{s_k s_k^T}{s_k^T v_k}.$$

end for

Instances of the framework

Cutting plane / bundle methods

- Points added incrementally until sufficient decrease obtained
- Finite number of additions until accepted step

Gradient sampling methods

- Points added randomly, incrementally until sufficient decrease obtained
- Sufficient number of iterations with "good" steps

We believe that either could use line search or trust region ideas

Outline

Contributions

Proposed two methods for unconstrained optimization

- one for stochastic, nonconvex problems
- one for deterministic, nonsmooth problems
- exploit self-correcting properties of BFGS-type updates

⋆ F. E. Curtis.
A Self-Correcting Variable-Metric Algorithm for Stochastic Optimization.
In *Proceedings of the 33rd International Conference on Machine Learning*, New York, NY, USA, 2016. JMLR.