A Self-Correcting Variable-Metric Algorithm for Stochastic Optimization

Frank E. Curtis, Lehigh University

International Conference on Machine Learning (ICML)
New York, NY, USA

21 June 2016
Motivation

Self-Correcting Properties of BFGS-type Updating

Proposed Algorithm

Summary
Outline

Motivation

Self-Correcting Properties of BFGS-type Updating

Proposed Algorithm

Summary
Consider unconstrained optimization problems of the form

$$\min_{w \in \mathbb{R}^d} f(w),$$

where

- $f(w) := \mathbb{E}[F(w, \xi)]$ with expectation w.r.t. distribution of random variable $\xi$;
- $f$ continuously differentiable, bounded below, and potentially nonconvex;
- $\nabla f$ Lipschitz continuous with constant $L > 0$.

**Goal:** Go beyond stochastic gradient (SG) to design improved methods

**Justification?:** with L. Bottou and J. Nocedal (submitted to SIAM Review)
“Optimization Methods for Large-Scale Machine Learning”
http://arxiv.org/abs/1606.04838
Balance between extremes

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

\[ w_{k+1} \leftarrow w_k - \alpha_k M_k g_k, \]

where

- \( \alpha_k > 0 \) is a stepsize;
- \( g_k \leftarrow \nabla f(w_k) \);
- \( \{M_k\} \) is updated dynamically.

Background on quasi-Newton:

- local rescaling of step (overcome ill-conditioning)
- only first-order derivatives required
- no linear system solves required
- global convergence guarantees (say, with line search)
- superlinear local convergence rate

How can the idea be carried over to a stochastic setting?
Previous work: BFGS-type methods

Much focus on the secant equation \((H_{k+1} \sim \text{Hessian approximation})\)

\[
H_{k+1}s_k = y_k \quad \text{where} \quad \begin{cases} 
  s_k := w_{k+1} - w_k \\
  y_k := \nabla f(w_{k+1}) - \nabla f(w_k)
\end{cases}
\]

and an appropriate replacement for the gradient displacement:

\[
y_k \leftarrow \nabla f(w_{k+1}, \xi_k) - \nabla f(w_k, \xi_k)
\]

use same seed
- oLBFGS, Schraudolph et al. (2007)
- SGD-QN, Bordes et al. (2009)
- RES, Mokhtari & Ribeiro (2014)

or

\[
y_k \leftarrow \left( \sum_{i \in S^H_k} \nabla^2 f(w_{k+1}, \xi_{k+1}, i) \right) s_k
\]

use action of step on subsampled Hessian
- SQN, Byrd et al. (2015)

Is this the right focus? Is there a better way (especially for nonconvex \(f\))?
Propose a quasi-Newton method for stochastic (nonconvex) 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
- motivating convergence theory for convex and nonconvex objectives
- dynamic noise reduction strategy
- limited memory variant

Observed stable behavior and overall good performance
Outline

Motivation

Self-Correcting Properties of BFGS-type Updating

Proposed Algorithm

Summary
BFGS-type updates

Inverse Hessian and Hessian approximation updating formulas \((s_k^T v_k > 0)\):

\[
M_{k+1} \leftarrow \left( I - \frac{v_k s_k^T}{s_k^T v_k} \right)^T M_k \left( I - \frac{v_k s_k^T}{s_k^T v_k} \right) + \frac{s_k s_k^T}{s_k^T v_k} 
\]

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

- Satisfy secant-type equations

\[
M_{k+1} v_k = s_k \quad \text{and} \quad H_{k+1} s_k = v_k,
\]

but these are not relevant for this talk.

- Choosing \(v_k \leftarrow y_k := g_{k+1} - g_k\) yields standard BFGS, but in this talk

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

This scheme is important to preserve self-correcting properties.
Geometric properties of Hessian update

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 span($s_k$).
- $Q_k$ yields $H_k$-orthogonal projection onto span($s_k)^{-1}H_k$. 
Geometric properties of Hessian update

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

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

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

- $P_k$ yields $H_k$-orthogonal projection onto span($s_k$).
- $Q_k$ yields $H_k$-orthogonal projection onto span($s_k$)$^\perp H_k$.

Returning to the Hessian update:

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

- Curvature projected out along span($s_k$)
- Curvature corrected by $\frac{v_k v_k^T}{s_k^T v_k} = \left( \frac{v_k v_k^T}{\|v_k\|^2} \right) \left( \frac{\|v_k\|^2}{v_k^T M_{k+1} v_k} \right)$ (inverse Rayleigh).
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 exist \( \{\eta, \theta\} \subset \mathbb{R}^{++} \) such that
\[
\eta \leq s_T k v_k \|s_k\|_2^2 \quad \text{and} \quad \|v_k\|_2^2 s_T k v_k \leq \theta.
\]

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 \( \lceil p K \rceil \) values of $k \in \{1, \ldots, K\}$:
\[
\iota \leq s_T k H_k s_k \|s_k\|_2 \|H_k s_k\|_2 \quad \text{and} \quad \kappa \leq \|H_k s_k\|_2 \|s_k\|_2 \leq \lambda.
\]

Proof technique.

Building on work of Powell (1976), etc., involves bounding growth of $\gamma(H_k) = \text{tr}(H_k) - \ln(\det(H_k))$. 

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

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

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 $k \in \{1, \ldots, K\}$:

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

**Proof technique.**

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

$$
\gamma(H_k) = \text{tr}(H_k) - \ln(\text{det}(H_k)).
$$
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 $k \in \{1, \ldots, K\}$:

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

Proof sketch.

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

Motivation

Self-Correcting Properties of BFGS-type Updating

Proposed Algorithm

Summary
**Algorithm SC**: Self-Correcting BFGS Algorithm

1: Choose $w_1 \in \mathbb{R}^d$.
2: Set $g_1 \approx \nabla f(w_1)$.
3: Choose a symmetric positive definite $M_1 \in \mathbb{R}^{d \times d}$.
4: Choose a positive scalar sequence $\{\alpha_k\}$.
5: **for** $k = 1, 2, \ldots$ **do**
6: Set $s_k \leftarrow -\alpha_k M_k g_k$.
7: Set $w_{k+1} \leftarrow w_k + s_k$.
8: Set $g_{k+1} \approx \nabla f(w_{k+1})$.
9: Set $y_k \leftarrow g_{k+1} - g_k$.
10: Set $\beta_k \leftarrow \min\{\beta \in [0, 1] : v(\beta) := \beta s_k + (1 - \beta)\alpha_k y_k \text{ satisfies (KEY)}\}$.
11: Set $v_k \leftarrow v(\beta_k)$.
12: Set

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

13: **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

\[
-\nabla f(w_k)^T \mathbb{E}_{\xi_k} [M_k g_k] \leq -\rho \|\nabla f(w_k)\|^2,
\]

and there exist scalars \( \sigma > 0 \) and \( \tau > 0 \) such that

\[
\mathbb{E}_{\xi_k} [\|M_k g_k\|^2] \leq \sigma + \tau \|\nabla f(w_k)\|^2.
\]

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

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

**Proof technique.**

Follows from the critical inequality

\[
\mathbb{E}_{\xi_k} [f(w_{k+1})] - f(w_k) \leq -\alpha_k \nabla f(w_k)^T \mathbb{E}_{\xi_k} [M_k g_k] + \alpha_k^2 L \mathbb{E}_{\xi_k} [\|M_k g_k\|^2].
\]

Also stronger results for strongly convex \( f \); see paper.
The conditions in this theorem cannot be verified in practice.

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

$$\mu \|g_k\|^2 \leq g_k^T M_k g_k \quad \text{and} \quad \|M_k g_k\|^2 \leq \nu \|g_k\|^2.$$
The conditions in this theorem cannot be verified in practice.

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

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

Stabilized variant (SC-s): Loop over (stochastic) gradient computation until

$$\rho \|\hat{g}_{k+1}\|^2 \leq \hat{g}_{k+1}^T M_{k+1} g_{k+1}$$

and

$$\|M_{k+1} g_{k+1}\|^2 \leq \sigma + \tau \|\hat{g}_{k+1}\|^2.$$

Recompute $g_{k+1}$, $\hat{g}_{k+1}$, and $M_{k+1}$ until these hold.
Numerical Experiments: a1a

logistic regression, data a1a, diminishing stepsizes
Numerical Experiments: rcv1

SC-L and SC-L-s: limited memory variants of SC and SC-s, respectively:

logistic regression, data rcv1, diminishing stepsizes
Numerical Experiments: \textit{mnist}

deep neural network, data \textit{mnist}, diminishing stepsizes
Outline

Motivation

Self-Correcting Properties of BFGS-type Updating

Proposed Algorithm

Summary
Proposed a quasi-Newton method for stochastic (nonconvex) optimization

▶ exploited self-correcting properties of BFGS-type updates
▶ properties of Hessians offer useful bounds for inverse Hessians
▶ motivating convergence theory for convex and nonconvex objectives
▶ dynamic noise reduction strategy
▶ limited memory variant

Observed stable behavior and overall good performance

★ F. E. Curtis.
A Self-Correcting Variable-Metric Algorithm for Stochastic Optimization.