Self-Correcting Variable Metric Algorithms

Frank E. Curtis, Lehigh University

U.S.-Mexico Workshop on Optimization and its Applications
Mérida, Yucatán, Mexico

8 January 2016

Outline

Motivation

Geometric and Self-Correcting Properties of BFGS-type Updating

Stochastic, Nonconvex Optimization

Deterministic, Convex, Nonsmooth Optimization

Summary

Outline

Motivation

Geometric and Self-Correcting Properties of BFGS-type Updating

Stochastic, Nonconvex Optimization

Deterministic, Convex, Nonsmooth Optimization

Summary

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?

Enhancements

Convex to nonconvex

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

Deterministic to stochastic

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

Smooth to nonsmooth

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

Enhancements

Convex to nonconvex

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

Deterministic to stochastic

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

Smooth to nonsmooth

- Scaling matrices necessarily(?) 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; 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 I have seen
- Result: More stable behavior than basic stochastic quasi-Newton

Deterministic, convex, nonsmooth:

- Proposal: Acceptance/rejection mechanism (no constrained QPs to solve)
- Result: Improved behavior over line search approach

Deterministic, nonconvex, nonsmooth (not this talk): see Curtis, Que (2015)

Outline

Motivation

Geometric and Self-Correcting Properties of BFGS-type Updating

Stochastic, Nonconvex Optimization

Deterministic, Convex, Nonsmooth Optimization

Summary

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 update, but I choose

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

  This inverse damping is important to preserve self-correcting properties.

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

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 are 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

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 are 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:

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

- Curvature projected out along 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

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

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

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

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

Geometric and Self-Correcting Properties of BFGS-type Updating

Stochastic, Nonconvex Optimization

Deterministic, Convex, 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 $\{x_k\}$,

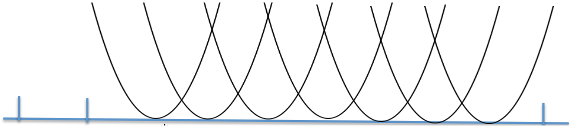
- f is continously differentiable and bounded below and
- ∇f is Lipscthiz continuous with constant $L > 0$,

but

- neither f nor ∇f can be computed exactly.

What has been done?

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



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

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 $\beta_k \leftarrow \min\{\beta \in [0,1] : v_k \leftarrow \beta s_k + (1-\beta)\alpha_k y_k$ satisfies (KEY)$\}$.
  Set $v_k \leftarrow \beta_k s_k + (1-\beta_k)\alpha_k y_k$.
  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

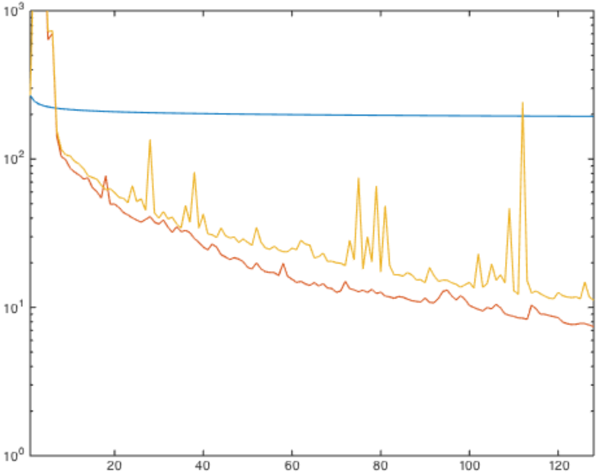
$$\mu \|g_k\|_2^2 \leq g_k^T W_k g_k \quad \text{and} \quad \|W_k g_k\|_2^2 \leq \nu \|g_k\|_2^2.$$

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

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

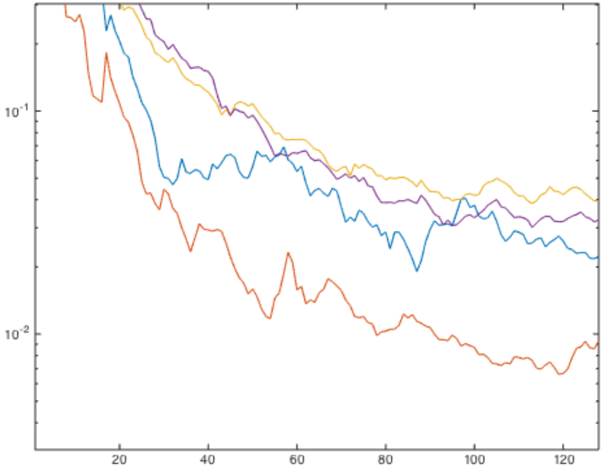
Recompute g_{k+1}, $\hat{g}_{k+1}$, and W_{k+1} until these hold.

Numerical Experiments



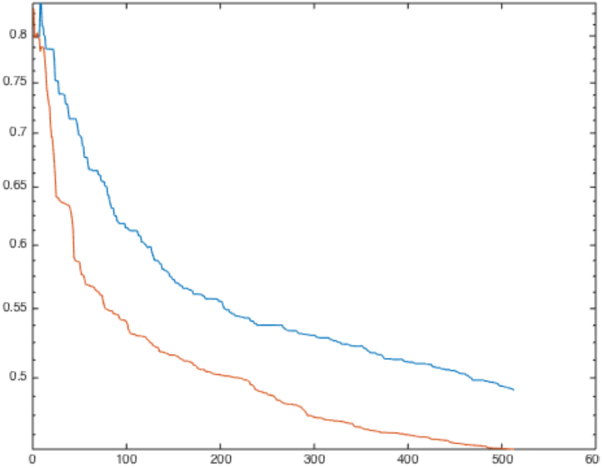
quadratic with noisy gradients

Numerical Experiments



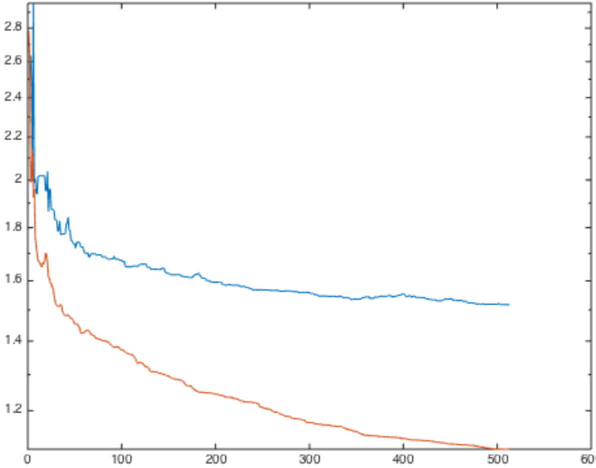
sum of quadratics, stochastic gradients

Numerical Experiments



UCI (breast cancer), stochastic gradients

Numerical Experiments



UCI (mnist), stochastic gradients

Outline

Motivation

Geometric and Self-Correcting Properties of BFGS-type Updating

Stochastic, Nonconvex Optimization

Deterministic, Convex, Nonsmooth Optimization

Summary

Convex, nonsmooth optimization

Consider unconstrained optimization problems of the form

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

where

- f is convex and nonsmooth.

The following algorithm

- maintains self-correcting property in "outer" algorithm while
- "inner" iterations lift curvature until sufficient decrease holds.

Algorithm VM-AR : Variable-Metric Algorithm with Acceptance/Rejection

Choose $x_1 \in \mathbb{R}^n$.
Set $g_1 \in \partial f(x_1)$.
Choose a symmetric positive definite $W_1 \in \mathbb{R}^{n \times n}$.
for $k = 1, 2, \ldots$ **do**
  Set s_k (next page).
  Set $x_{k+1} \leftarrow x_k + s_k$.
  Set $g_{k+1} \in \partial f(x_{k+1})$.
  Set $y_k \leftarrow g_{k+1} - g_k$.
  Set $\beta_k \leftarrow \min\{\beta \in [0,1] : v_k \leftarrow \beta s_k + (1-\beta) y_k \text{ satisfies (KEY)}\}$.
  Set $v_k \leftarrow \beta_k s_k + (1-\beta_k) y_k$.
  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

Algorithm VM-AR-step : VM-AR step computation

Set $\overline{W}_{k,1} \leftarrow W_k$.
for $l = 1, 2, \ldots$ **do**
  Set $\bar{s}_{k,l} \leftarrow -\overline{W}_{k,l} g_k$.
  Set $\bar{x}_{k,l} \leftarrow x_k + \bar{s}_{k,l}$.
  if $f(\bar{x}_{k,l}) \leq f(x_k) - \eta g_k^T \overline{W}_{k,l} g_k$ **then**
    break
  else
    Set

$$\overline{M}_{k,l} \leftarrow I - \phi\left(\frac{\bar{s}_{k,l}\bar{s}_{k,l}^T}{\bar{s}_{k,l}^T \bar{s}_{k,l}}\right) \quad \text{then}$$

$$\overline{W}_{k,l+1} \leftarrow \overline{M}_{k,l}^T \overline{W}_{k,l} \overline{M}_{k,l}.$$

  end if
end for
Set $s_k \leftarrow \bar{s}_{k,l}$.

Global convergence theory

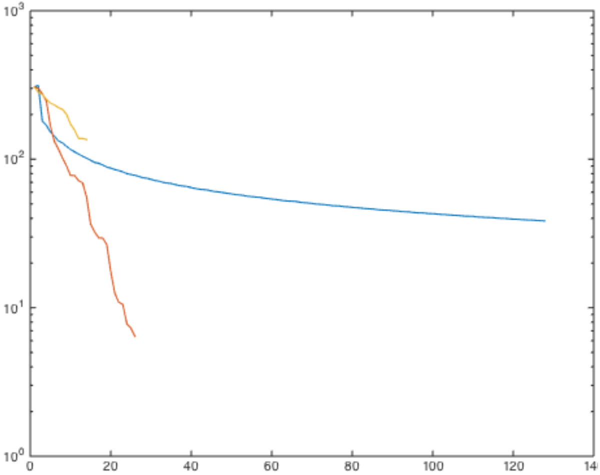
If x_k suboptimal, then Algorithm VM-AR-step terminates finitely:

- finite number of lifts until sufficient decrease is attained.

If f has a minimizer, then x_k converges to a minimizer.

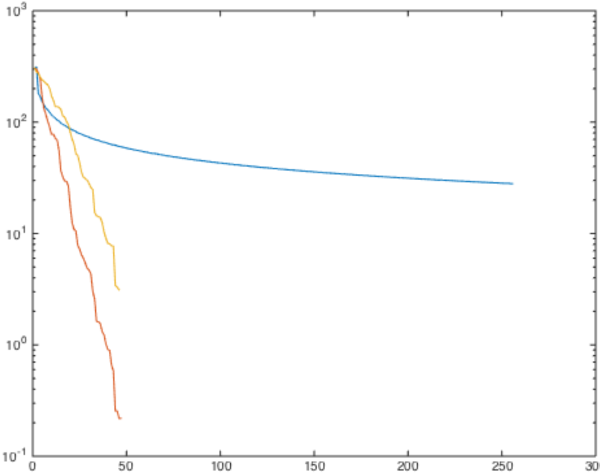
- (Or so I believe! Proof is on-going. Strategy is...)
- Either finite termination or infinite descent steps;
- self-correcting properties of updates yield "nice" initial matrices for all k;
- recall strategy of bundle methods.

Numerical Experiments



quadratic plus polyhedral (function evaluation limit)

Numerical Experiments



quadratic plus polyhedral (more function evaluations)

Outline

Motivation

Geometric and Self-Correcting Properties of BFGS-type Updating

Stochastic, Nonconvex Optimization

Deterministic, Convex, Nonsmooth Optimization

Summary

Contributions

Proposed two methods for unconstrained optimization

- one for stochastic, nonconvex problems
- one for deterministic, convex, nonsmooth problems
- exploit self-correcting properties of BFGS-type updates; Byrd, Nocedal (1989) which builds on work of Powell (1976)

⋆ F. E. Curtis.
Self-Correcting Variable Metric Algorithms.
Working Paper, 2016.