Quasi-Newton Methods

Javier Peña

Convex Optimization 10-725/36-725

Last time: primal-dual interior-point methods

Consider the problem

$$\begin{aligned} \min_x \quad & f(x) \\ \text{subject to} \quad & Ax = b \\ & h(x) \leq 0 \end{aligned}$$

Assume $f, h_1, \ldots, h_m$ are convex and differentiable. Assume also that strong duality holds.

Central path equations:

$$\begin{aligned} \nabla f(x) + \nabla h(x) u + A^{\mathsf{T}} v &= 0 \\ U h(x) + \tau \mathbf{1} &= 0 \\ Ax - b &= 0 \\ u, -h(x) &> 0. \end{aligned}$$

Primal-dual interior-point algorithm

Let

$$r(x,u,v) := \begin{bmatrix} \nabla f(x) + \nabla h(x)u + A^{\mathsf{T}}v \\ Uh(x) + \tau \mathbf{1} \\ Ax - b \end{bmatrix},$$

Crux of each iteration

$$(x^+, u^+, v^+) := (x,u,v) + \theta(\Delta x, \Delta u, \Delta v)$$

where $(\Delta x, \Delta u, \Delta v)$ is the Newton step:

$$\begin{bmatrix} \nabla^2 f(x) + \sum_i u_i \nabla^2 h_i(x) & \nabla h(x) & A^{\mathsf{T}} \\ U\nabla h(x)^{\mathsf{T}} & H(x) & 0 \\ A & 0 & 0 \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta u \\ \Delta v \end{bmatrix} = -r(x,u,v)$$

Here $U = \mathsf{Diag}(u),\ H(x) = \mathsf{Diag}(h(x))$.

Outline

Today:

- Motivation for quasi-Newton methods
- Most popular updates: SR1, DFP, BFGS, Broyden class
- Superlinear convergence
- Limited memory BFGS
- Stochastic quasi-Newton methods

Gradient descent and Newton revisited

Consider the unconstrained, smooth optimization problem

$$\min_x \ f(x)$$

where f is twice differentiable, and $\mathrm{dom}(f) = \mathbb{R}^n$.

Gradient descent method

$$x^+ = x - t\nabla f(x)$$

Newton's method

$$x^+ = x - t\nabla^2 f(x)^{-1}\nabla f(x)$$

Quasi-Newton methods

Two main steps in Newton's method:

- Compute Hessian $\nabla^2 f(x)$
- Solve the system of equations

$$\nabla^2 f(x)p = -\nabla f(x).$$

Each of these two steps could be expensive.

Quasi-Newton method

Use instead

$$x^+ = x + tp$$

where

$$Bp = -\nabla f(x)$$

for some approximation B of $\nabla^2 f(x)$.

Want B easy to compute and $Bp = g$ easy to solve.

A bit of history

In the mid 1950s W. Davidon was a physicist at Argonne National Lab. He was using a coordinate descent method to solve a long optimization calculation that kept crashing the computer before finishing.

Davidon figured out a way to accelerate the computation — the first quasi-Newton method ever. Although Davidon's contribution was a major breakthrough in optimization, his original paper was rejected. In 1991, after more than 30 years, his paper was published in the first issue of the SIAM Journal on Optimization.

In addition to his remarkable work in optimization, Davidon was a peace activist. Part of his story is nicely described in *The Burglary* a book published shortly after he passed away in 2013.

Secant equation

We would like B^k to approximate $\nabla^2 f(x^k)$, that is

$$\nabla f(x^k + s) \approx \nabla f(x^k) + B^k s.$$

Once $x^{k+1} = x^k + s^k$ is computed, we would like a new B^{k+1}.

Idea: since B^k already contains some information, make some suitable update.

Reasonable requirement for B^{k+1}

$$\nabla f(x^{k+1}) = \nabla f(x^k) + B^{k+1} s^k$$

or equivalently

$$B^{k+1} s^k = \nabla f(x^{k+1}) - \nabla f(x^k).$$

Secant equation

The latter condition is called the secant equation and written as

$$B^{k+1}s^k = y^k \quad \text{or simply } B^+ s = y$$

where $s^k = x^{k+1} - x^k$ and $y^k = \nabla f(x^{k+1}) - \nabla f(x^k)$.

In addition to the secant equation, we would like

(i) B^+ symmetric

(ii) B^+ "close" to B

(iii) B positive definite $\Rightarrow$ B^+ positive definite

Symmetric rank-one update (SR1)

Try an update of the form

$$B^+ = B + auu^\mathsf{T}$$

Observe

$$B^+ s = y \Rightarrow (au^\mathsf{T} s)u = y - Bs$$

The latter can hold only if u is a multiple of $y - Bs$.

It follows that the only symmetric rank-one update that satisfies the secant equation is

$$B^+ = B + \frac{(y - Bs)(y - Bs)^\mathsf{T}}{(y - Bs)^\mathsf{T} s}.$$

Sherman-Morrison-Woodbury formula

A low-rank update on a matrix corresponds to a low rank update on its inverse.

Theorem (Sherman-Morrison-Woodbury formula)

Assume $A \in \mathbb{R}^{n\times n}$, and $U, V \in \mathbb{R}^{n\times d}$. Then $A + UV^{\mathsf{T}}$ is nonsingular if and only if $I + V^{\mathsf{T}}A^{-1}U$ is nonsingular. In that case

$$(A + UV^{\mathsf{T}})^{-1} = A^{-1} - A^{-1}U(I + V^{\mathsf{T}}A^{-1}U)^{-1}A^{-1}$$

Thus for the SR1 update the inverse H of B is also easily updated:

$$H^{+} = H + \frac{(s - Hy)(s - Hy)^{\mathsf{T}}}{(s - Hy)^{\mathsf{T}}y}.$$

SR1 is simple but has two shortcomings: it may fail if $(y - Bs)^{\mathsf{T}}s \approx 0$ and it does not preserve positive definiteness.

Davidon-Fletcher-Powell (DFP) update

Try a rank-two update

$$H^+ = H + auu^\mathsf{T} + bvv^\mathsf{T}.$$

The secant equation yields

$$s - Hy = (au^\mathsf{T}y)u + (bv^\mathsf{T}y)v.$$

Putting $u = s,\ v = Hy$, and solving for a, b we get

$$H^+ = H - \frac{Hyy^\mathsf{T}H}{y^\mathsf{T}Hy} + \frac{ss^\mathsf{T}}{y^\mathsf{T}s}$$

By Sherman-Morrison-Woodbury we get a rank-two update on B

$$\begin{aligned} B^+ &= B + \frac{(y - Bs)y^\mathsf{T}}{y^\mathsf{T}s} + \frac{y(y - Bs)^\mathsf{T}}{y^\mathsf{T}s} - \frac{(y - Bs)^\mathsf{T}s}{(y^\mathsf{T}s)^2}yy^\mathsf{T} \\ &= \left(I - \frac{ys^\mathsf{T}}{y^\mathsf{T}s}\right) B \left(I - \frac{sy^\mathsf{T}}{y^\mathsf{T}s}\right) + \frac{yy^\mathsf{T}}{y^\mathsf{T}s} \end{aligned}$$

DFP update – alternate derivation

Find B^+ closest to B in some norm so that B^+ satisfies the secant equation and is symmetric:

$$\begin{aligned} \min_{B^+} \quad & \|B^+ - B\|_? \\ \text{subject to} \quad & B^+ = (B^+)^\mathsf{T} \\ & B^+ s = y \end{aligned}$$

What norm to use?

Curvature condition

Observe: B^+ positive definite and $B^+s = y$ imply

$$y^\mathsf{T} s = s^\mathsf{T} B^+ s > 0.$$

The inequality $y^\mathsf{T} s > 0$ is called the curvature condition.

Fact: if $y, s \in \mathbb{R}^n$ and $y^\mathsf{T} s > 0$ then there exists M symmetric and positive definite such that $Ms = y$.

DFP update again

Solve

$$\begin{aligned} \min_{B^+} \quad & \|W^{-1}(B^+ - B)W^{-\mathsf{T}}\|_F \\ \text{subject to} \quad & B^+ = (B^+)^\mathsf{T} \\ & B^+ s = y \end{aligned}$$

where $W \in \mathbb{R}^{n \times n}$ is nonsingular and such that $WW^\mathsf{T} s = y$.

Broyden-Fletcher-Goldfarb-Shanno (BFGS) update

Same ideas as the DFP update but with roles of B and H exchanged.

Secant equation

$$H^+ y = s \Leftrightarrow B^+ s = y$$

Closeness to H:

$$\begin{aligned} \min_{H^+} \quad & \|W^{-1}(H^+ - H)W^{-\mathsf{T}}\|_F \\ \text{subject to} \quad & H^+ = (H^+)^{\mathsf{T}} \\ & H^+ y = s \end{aligned}$$

where $W \in \mathbb{R}^{n \times n}$ is nonsingular and $WW^{\mathsf{T}} y = s$.

BFGS update

Swapping H and B and y and s in the DFP update we get

$$B^+ = B - \frac{Bss^\mathsf{T}B}{s^\mathsf{T}Bs} + \frac{yy^\mathsf{T}}{y^\mathsf{T}s}$$

and

$$\begin{aligned} H^+ &= H + \frac{(s - Hy)s^\mathsf{T}}{y^\mathsf{T}s} + \frac{s(s - Hy)^\mathsf{T}}{y^\mathsf{T}s} - \frac{(s - Hy)^\mathsf{T}y}{(y^\mathsf{T}s)^2}ss^\mathsf{T} \\ &= \left(I - \frac{sy^\mathsf{T}}{y^\mathsf{T}s}\right) H \left(I - \frac{ys^\mathsf{T}}{y^\mathsf{T}s}\right) + \frac{ss^\mathsf{T}}{y^\mathsf{T}s} \end{aligned}$$

Both DFP and BFGS preserve positive definiteness: if B is positive definite and $y^\mathsf{T}s > 0$ then B^+ is positive definite.

BFGS is more popular than DFP. It also has a "self-correcting" property.

The Broyden class

SR1, DFP, and BFGS are some of many possible quasi-Newton updates. The *Broyden class* of updates is defined by

$$B^+ = (1-\phi)B^+_{\text{BFGS}} + \phi B^+_{\text{DFP}}, \quad \text{for } \phi \in \mathbb{R}.$$

By putting $v := \frac{y}{y^\mathsf{T} s} - \frac{Bs}{s^\mathsf{T} Bs}$ we can rewrite the above as

$$B^+ = B - \frac{Bss^\mathsf{T} B}{s^\mathsf{T} Bs} + \frac{yy^\mathsf{T}}{y^\mathsf{T} s} + \phi(s^\mathsf{T} Bs)vv^\mathsf{T}.$$

Observe

- BFGS and DFP correspond to $\phi = 0$ and $\phi = 1$ respectively.
- SR1 corresponds to $\phi = \frac{y^\mathsf{T} s}{y^\mathsf{T} s - s^\mathsf{T} Bs}$

Superlinear convergence

Back to our main goal:

$$\min_x \ f(x)$$

where f is twice differentiable, and $\mathrm{dom}(f) = \mathbb{R}^n$.

Quasi-Newton method

- Pick initial x^0 and B^0
- For $k = 0, 1, \ldots$
  - Solve $B^k p^k = -\nabla f(x^k)$
  - Pick t_k and let $x^{k+1} = x^k + t_k p^k$
  - update B^k to B^{k+1}

  end for

Superlinear convergence

Under suitable assumptions on the objective function $f(x)$ and on the step size t_k we get superlinear convergence:

$$\lim_{k\to\infty} \frac{\|x^{k+1} - x^*\|}{\|x^k - x^*\|} = 0.$$

Step length (Wolfe conditions)

Assume t is chosen (via backtracking) so that

$$f(x + tp) \leq f(x) + \alpha_1 t \nabla f(x)^\mathsf{T} p$$

and

$$\nabla f(x + tp)^\mathsf{T} p \geq \alpha_2 |\nabla f(x)^\mathsf{T} p|$$

for $0 < \alpha_1 < \alpha_2 < 1$.

Superlinear convergence

The crux of superlinear convergence is the following technical result.

Theorem (Dennis-Moré)

Assume f is twice differentiable, $x^k \to x^*$ such that $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*)$ is positive definite. If the search direction p^k satisfies

$$\lim_{k\to\infty} \frac{\|\nabla f(x^k) - \nabla^2 f(x^k)p^k\|}{\|p^k\|} = 0 \tag{1}$$

then there exists k_0 such that

(i) *The step length $t_k = 1$ satisfies Wolfe conditions for $k \geq k_0$.*

(ii) *If $t_k = 1$ for $k \geq k_0$ then $x^k \to x^*$ superlinearly.*

Under suitable assumptions, DFP and BFGS updates ensure (1) holds for $p^k = -H^k \nabla f(x^k)$ and we get superlinear convergence.

Limited memory BFGS (LBFGS)

For large problems, exact quasi-Newton updates becomes too costly.

An alternative is to maintain a compact approximation of the matrices: save only a few $n \times 1$ vectors and compute the matrix implicitly.

The BFGS method computes the search direction

$$p = -H \nabla f(x)$$

where H is updated via

$$H^+ = \left(I - \frac{s y^\mathsf{T}}{y^\mathsf{T} s} \right) H \left(I - \frac{y s^\mathsf{T}}{y^\mathsf{T} s} \right) + \frac{s s^\mathsf{T}}{y^\mathsf{T} s}$$

LBFGS

Instead of computing and storing H, compute an implicit modified version of H by maintaining the last m pairs (y, s).

Observe

$$H^+ g = p + (\alpha - \beta)s$$

where

$$\alpha = \frac{s^\mathsf{T} g}{y^\mathsf{T} s},\ q = g - \alpha y,\ p = Hq,\ \beta = \frac{y^\mathsf{T} p}{y^\mathsf{T} s}$$

Hence Hg can be computed via two loops of length k if H is obtained after k BFGS updates.

For $k \geq m$ LBFGS limits the length of these loops to m.

LBFGS

LBFGS computation of $p^k = -H^k \nabla f(x^k)$

1. $q := -\nabla f(x^k)$
2. for $i = k-1, \ldots, \min(k-m, 0)$
   $\alpha_i := \frac{(s^i)^\mathsf{T} q}{(y^i)^\mathsf{T} s^i}$
   $q := q - \alpha y^i$
   end for
3. $p := H^{0,k} q$
4. for $i = \min(k-m, 0), \ldots, k-1$
   $\beta := \frac{(y^i)^\mathsf{T} p}{(y^i)^\mathsf{T} s^i}$
   $p := p + (\alpha_i - \beta) s^i$
5. return p

In step 3 $H^{0,k}$ is the "initial" H. Popular choice

$$H^{0,k} := \frac{(y^{k-1})^\mathsf{T} s^{k-1}}{(y^{k-1})^\mathsf{T} y^{k-1}} I$$

Stochastic quasi-Newton methods

Consider the problem

$$\min_x \ \mathbb{E}(f(x;\xi))$$

where $f(x,\xi)$ depends on a random input ξ.

Stochastic gradient descent (SGD)

Use a draw of ξ to compute a random gradient of the objective function

$$x^{k+1} = x^k - t_k \nabla f(x^k, \xi_k)$$

Stochastic quasi-Newton template

Extend previous ideas

$$x^{k+1} = x^k - t_k H^k \nabla f(x^k, \xi_k)$$

Stochastic quasi-Newton methods

Some challenges:

- Theoretical limitations: stochastic iteration cannot have faster than sublinear rate of convergence.
- Additional cost: a major advantage of SGD (and similar algorithms) is their low cost per iteration. Could the additional overhead of a quasi-Newton method be compensated?
- Conditioning of scaling matrices: updates on H depend on consecutive gradient estimates. The noise in the random gradient estimates can be a hindrance in the updates.

Online BFGS

The most straightforward adaptation of quasi-Newton methods is to use BFGS (or LBFGS) with

$$s^k = x^{k+1} - x^k, \quad y^k = \nabla f(x^{k+1}, \xi_k) - \nabla f(x^k, \xi_k)$$

Key: use the same ξ_k in the two above stochastic gradients.

Maintain H^k via BFGS or LBFGS updates.

This approach, referred to as online BFGS, is due to Schraudolph-Yu-Günter.

With proper tuning it can give some improvement over SGD.

Other stochastic quasi-Newton methods

Byrd-Hansen-Nocedal-Singer propose a stochastic version of LBFGS but with two main changes:

- Perform LBFGS update only every L iterations
- For s and y use respectively

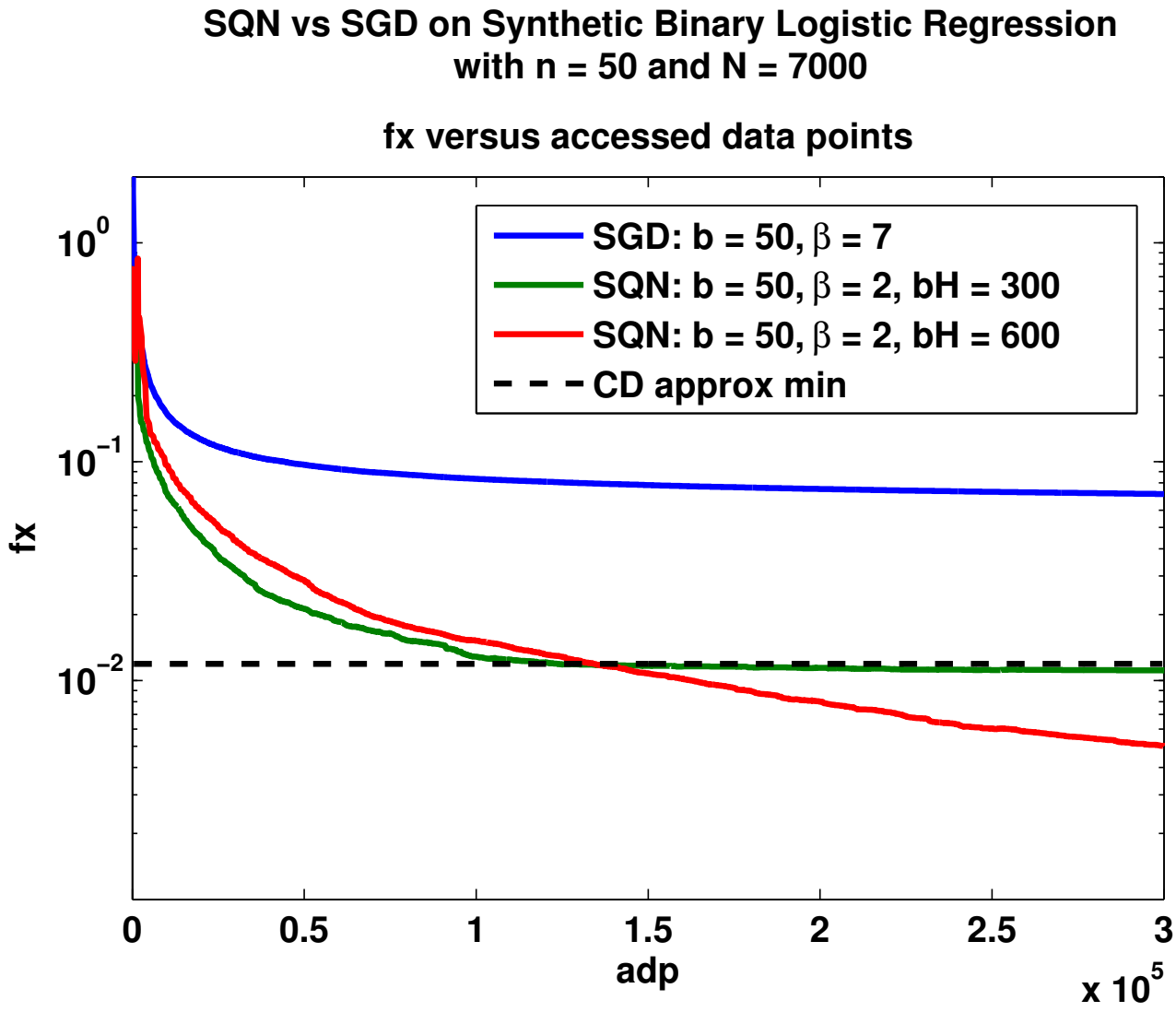
$$s = \bar{x}^t - \bar{x}^{t-1} \quad \text{where} \quad \bar{x}^t = \sum_{i=k-L+1}^{k} x^i$$

and

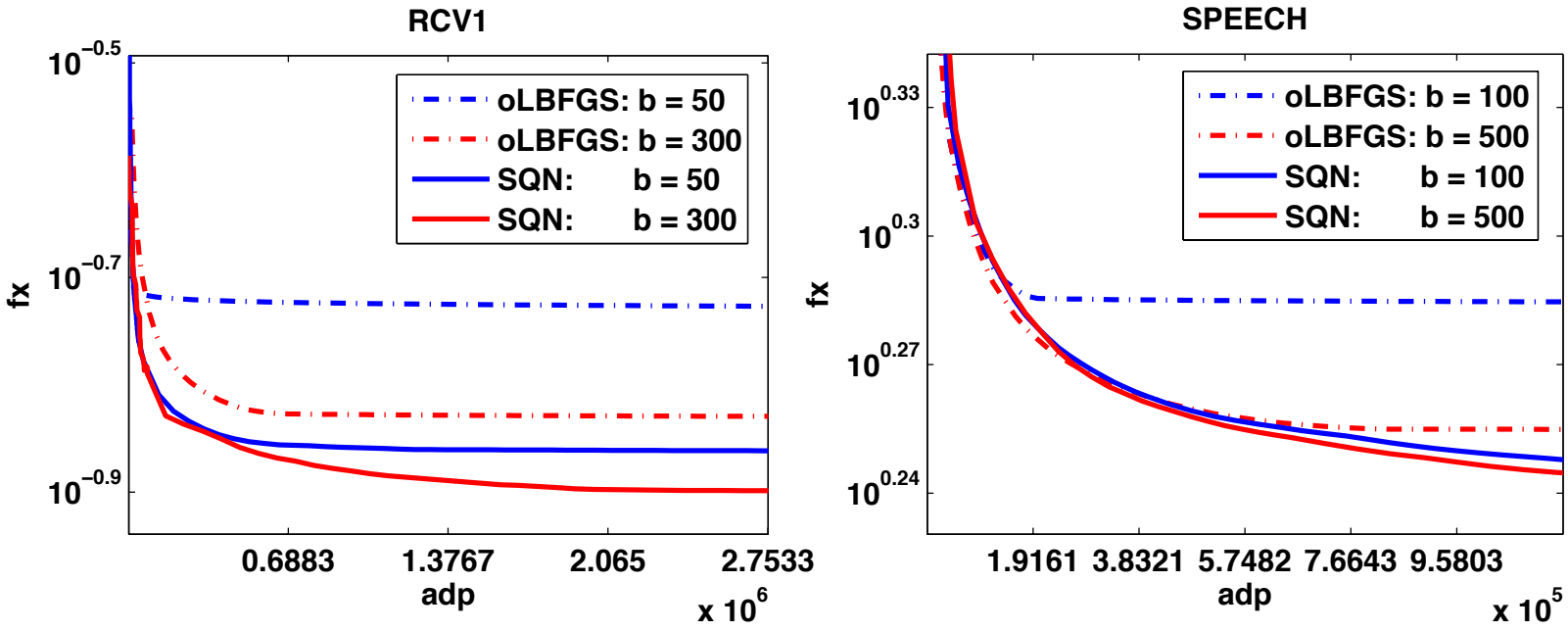
$$y = \widehat{\nabla}^2 F(\bar{x}^t) s$$

where $\widehat{\nabla}^2 F(\bar{x}^t)$ is a Hessian approximation (based on a random subsample).

Some results (Byrd-Hansen-Nocedal-Singer)



More results (Byrd-Hansen-Nocedal-Singer)



References and further reading

- J. Dennis and R. Schnabel (1996), "Numerical methods for unconstrained optimization and nonlinear equations."
- J. Nocedal and S. Wright (2006), "Numerical optimization", Chapters 6 and 7
- N. Schraudolph, J. Yu, S. Günter (2007), "A stochastic quasi-Newton method for online convex optimization."
- L. Bottou, F. Curtis, J. Nocedal (2016), "Optimization methods for large-scale machine learning."