Quasi-Newton Methods

Last modified: 2026-10-31 15:22:31 (PDT)

Quasi-Newton methods

Source and license

This section is adapted from the quasi-Newton lecture of Zhou [1], used under the MIT License. The license text is:

> Copyright (c) 2025 ucla-biostat-257
>
> Permission is hereby granted, free of charge, to any person obtaining a copy of this software and associated documentation files (the "Software"), to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, subject to the following conditions:
>
> The above copyright notice and this permission notice shall be included in all copies or substantial portions of the Software.
>
> THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.

From the source, this section takes the order of ideas: the cost of Newton's method, the secant condition, and the BFGS and limited-memory BFGS updates. The source states the updates as solutions of nearest-matrix problems without proof; here the BFGS update

is written out. The Davidon-Fletcher-Powell and rank-1 updates, the nearest-matrix derivation, the history, and the Julia code are not part of this section.

Definition 0.1 (Secant condition)**.** Let $f : \mathbb{R}^p \to \mathbb{R}$ have first partial derivatives, and let $\underset{\sim}{x}, \underset{\sim}{x}^+ \in \mathbb{R}^p$. Write $\underset{\sim}{s} \overset{\text{def}}{=} \underset{\sim}{x}^+ - \underset{\sim}{x}$ for the step and $y \overset{\text{def}}{=} \frac{\partial}{\partial \underset{\sim}{x}} f(\underset{\sim}{x}^+) - \frac{\partial}{\partial \underset{\sim}{x}} f(\underset{\sim}{x})$ for the change in the gradient. A symmetric matrix $\mathbf{B}$ satisfies the **secant condition** [2, Ch. 6] for $(\underset{\sim}{s}, \underset{\sim}{y})$ when

$$\mathbf{B}\underset{\sim}{s} = \underset{\sim}{y}.$$

Example 0.1 (The Hessian of a quadratic function satisfies the secant condition)**.** Let $f(\underset{\sim}{x}) = \frac{1}{2}\underset{\sim}{x}^\top \mathbf{Q}\underset{\sim}{x} - \underset{\sim}{b} \cdot \underset{\sim}{x}$ with $\mathbf{Q}$ symmetric. Then $\frac{\partial}{\partial \underset{\sim}{x}} f(\underset{\sim}{x}) = \mathbf{Q}\underset{\sim}{x} - \underset{\sim}{b}$, so for any $\underset{\sim}{x}$ and $\underset{\sim}{x}^+$

$$\begin{aligned} \underset{\sim}{y} &= (\mathbf{Q}\underset{\sim}{x}^+ - \underset{\sim}{b}) - (\mathbf{Q}\underset{\sim}{x} - \underset{\sim}{b}) && \text{(gradient at each point)} \\ &= \mathbf{Q}(\underset{\sim}{x}^+ - \underset{\sim}{x}) && \text{(collect terms)} \\ &= \mathbf{Q}\underset{\sim}{s}. && \text{(definition of } \underset{\sim}{s}) \end{aligned}$$

So the Hessian $\mathbf{Q}$ of the quadratic function satisfies the secant condition for every step. For a general f, $\underset{\sim}{y} \approx \mathbf{H}_f \underset{\sim}{s}$ holds for a short step, with $\mathbf{H}_f$ the Hessian of f near the step, which is why a matrix that satisfies the secant condition is a reasonable stand-in for the Hessian.

Definition 0.2 (Quasi-Newton method)**.** Let $f : \mathbb{R}^p \to \mathbb{R}$ have first partial derivatives, let $\underset{\sim}{x}^{(0)}$ be a starting point, and let $\mathbf{H}^{(0)}$ be a positive definite matrix. A **quasi-Newton method** [2, Ch. 6] produces, for $t = 0, 1, 2, \ldots$,

$$\underset{\sim}{x}^{(t+1)} \overset{\text{def}}{=} \underset{\sim}{x}^{(t)} - \alpha^{(t)} \mathbf{H}^{(t)} \frac{\partial}{\partial \underset{\sim}{x}} f(\underset{\sim}{x}^{(t)}),$$

where $\alpha^{(t)} > 0$ is a step length chosen by a line search, and $\mathbf{H}^{(t+1)}$ is a symmetric positive definite matrix whose inverse satisfies the secant condition, that is, $\mathbf{H}^{(t+1)} \underset{\sim}{y}^{(t)} = \underset{\sim}{s}^{(t)}$, with $\underset{\sim}{s}^{(t)} = \underset{\sim}{x}^{(t+1)} - \underset{\sim}{x}^{(t)}$ and $\underset{\sim}{y}^{(t)} = \frac{\partial}{\partial \underset{\sim}{x}} f(\underset{\sim}{x}^{(t+1)}) - \frac{\partial}{\partial \underset{\sim}{x}} f(\underset{\sim}{x}^{(t)})$.

Example 0.2 (One quasi-Newton step on a quadratic function)**.** Take $f(\underset{\sim}{x}) = \frac{1}{2}\underset{\sim}{x}^\top \mathbf{Q}\underset{\sim}{x} - \underset{\sim}{b} \cdot \underset{\sim}{x}$ with $\mathbf{Q} = \begin{bmatrix} 4 & 1 \\ 1 & 3 \end{bmatrix}$ and $\underset{\sim}{b} = (1, 2)^\top$ (Example 0.1), $\underset{\sim}{x}^{(0)} = \underset{\sim}{0}$ and $\mathbf{H}^{(0)} = \mathbf{I}_2$. For a quadratic function, the step length that minimizes f along a direction $\underset{\sim}{d}$ is $\alpha = -\frac{\partial}{\partial \underset{\sim}{x}} f(\underset{\sim}{x}) \cdot \underset{\sim}{d} / \underset{\sim}{d}^\top \mathbf{Q}\underset{\sim}{d}$.

```r
Q <- matrix(c(4, 1, 1, 3), nrow = 2)
b <- c(1, 2)
f <- function(x) 0.5 * sum(x * (Q %*% x)) - sum(b * x)
grad <- function(x) drop(Q %*% x - b)
x0 <- c(0, 0)
H0 <- diag(2)
d <- -drop(H0 %*% grad(x0))
step <- -sum(grad(x0) * d) / drop(t(d) %*% Q %*% d)
x1 <- x0 + step * d
s <- x1 - x0
y <- grad(x1) - grad(x0)
```

The step length is 0.25, which moves the iterate to $\underset{\sim}{x}^{(1)} = (0.25, 0.5)$ and lowers f from 0 to -0.625. The difference of the gradients, $\underset{\sim}{y}^{(0)} = (1.5, 1.75)$, is $\mathbf{Q}\underset{\sim}{s}^{(0)}$ (the secant condition for a quadratic function).

i Remark (Why a quasi-Newton method is used)

Remark 0.1 (Why a quasi-Newton method is used). Because $\mathbf{H}^{(t)}$ is positive definite, $-\mathbf{H}^{(t)}\frac{\partial}{\partial \underset{\sim}{x}} f(\underset{\sim}{x}^{(t)})$ is a descent direction, so, when the gradient is not $\underset{\sim}{0}$, a line search finds a step that decreases f. Newton's method needs the Hessian and the solution of a system of equations at every iteration, which costs about p^3 operations for a dense $p \times p$ Hessian, and requires the formulas for its entries. A quasi-Newton method needs only the gradient and an update of $\mathbf{H}^{(t)}$ costing about p^2 operations. Near a minimizer the method speeds up superlinearly: faster than gradient descent, slower than Newton's method [2, Ch. 6]. Replacing $\mathbf{H}^{(t)}$ by $\mathbf{I}_p$ in the step formula gives gradient descent.

Definition 0.3 (BFGS update (Broyden-Fletcher-Goldfarb-Shanno))**.** Let $\mathbf{H}$, an approximation of the inverse Hessian, be a $p \times p$ positive definite matrix, and let $\underset{\sim}{s}, \underset{\sim}{y} \in \mathbb{R}^p$ satisfy $\underset{\sim}{y} \cdot \underset{\sim}{s} > 0$. With $\rho \stackrel{\text{def}}{=} 1/\underset{\sim}{y} \cdot \underset{\sim}{s}$, the **BFGS update** [2, Ch. 6] of $\mathbf{H}$ is

$$\mathbf{H}^{+} \stackrel{\text{def}}{=} (\mathbf{I}_p - \rho \underset{\sim}{s}\underset{\sim}{y}^{\top})\mathbf{H}(\mathbf{I}_p - \rho \underset{\sim}{y}\underset{\sim}{s}^{\top}) + \rho \underset{\sim}{s}\underset{\sim}{s}^{\top}.$$

Example 0.3 (A BFGS update from the quasi-Newton step)**.** Continue Example 0.2 with $\mathbf{H} = \mathbf{I}_2$, $\underset{\sim}{s} = \underset{\sim}{s}^{(0)}$, and $\underset{\sim}{y} = \underset{\sim}{y}^{(0)}$.

```
rho <- 1 / sum(y * s)
left <- diag(2) - rho * s %*% t(y)
H1 <- left %*% H0 %*% t(left) + rho * s %*% t(s)
H1
#>          [,1]   [,2]
#> [1,]  0.6625 -0.425
#> [2,] -0.4250  0.650
drop(H1 %*% y)
#> [1] 0.25 0.50
```

Here $y \cdot s = 1.25 > 0$, so $\rho = 0.8$. The updated matrix satisfies $\mathbf{H}^+ y = (0.25, 0.5)$, which is s itself, and its smallest eigenvalue is 0.231204, so $\mathbf{H}^+$ is positive definite.

i Remark (The curvature condition $y \cdot s > 0$)

Remark 0.2 (The curvature condition $y \cdot s > 0$). If the strict inequality of Theorem 4 in Convexity in Several Variables holds for distinct points, then $y \cdot s > 0$ for $s = x^+ - x \neq 0$ and $y = \frac{\partial}{\partial x} f(x^+) - \frac{\partial}{\partial x} f(x)$. Add the versions at (x^+, x) and at (x, x^+):

$$f(x^+) + f(x) > f(x) + f(x^+) + \frac{\partial}{\partial x} f(x) \cdot s - \frac{\partial}{\partial x} f(x^+) \cdot s,$$

so $0 > -y \cdot s$. For a general f, a line search that enforces a curvature condition on the step length (the Wolfe conditions [2, Ch. 3]) guarantees $y \cdot s > 0$. Such a step can be found along a descent direction when f cannot keep decreasing forever.

Definition 0.4 (Limited-memory BFGS (L-BFGS))**. Limited-memory BFGS** (**L-BFGS**) [2, Ch. 7] is the quasi-Newton method that never forms $\mathbf{H}^{(t)}$. It stores only the m most recent pairs $(s^{(j)}, y^{(j)})$, with m small, such as 5 or 20, and obtains $\mathbf{H}^{(t)} \frac{\partial}{\partial x} f(x^{(t)})$ implicitly, through a few vector operations, as if the BFGS update for those m pairs were applied to a simple starting matrix, such as a multiple of $\mathbf{I}_p$.

Example 0.4 (L-BFGS applies the update without forming the matrix)**.** Continue Example 0.3 with $m = 1$ stored pair $(s^{(0)}, y^{(0)})$ and starting matrix $\mathbf{I}_2$. Below, $\mathbf{H}^+ \frac{\partial}{\partial x} f(x^{(1)})$ is formed from vectors alone, starting from `H0`, then compared with the matrix product.

```
g1 <- grad(x1)
alpha <- rho * sum(s * g1)
q <- g1 - alpha * y
r <- drop(H0 %*% q)
beta <- rho * sum(y * r)
r <- r + s * (alpha - beta)
rbind(from_vectors = r, from_matrix = drop(H1 %*% g1))
#>                 [,1]   [,2]
#> from_vectors 0.4375 -0.375
#> from_matrix  0.4375 -0.375
```

The two rows agree to within 5.6e-17, although the first never forms the matrix. BFGS stores a dense $p \times p$ matrix and spends about p^2 operations per iteration; L-BFGS stores $2m$ vectors with p entries and spends about mp operations per iteration, so it remains usable when p is in the millions.

[1] H. Zhou, "Quasi-Newton methods." Lecture notes for Biostat 257, Statistical Computing, University of California, Los Angeles, 2025. Accessed: Oct. 26, 2026. [Online]. Available: https://ucla-biostat-257.github.io/2025spring/slides/24-quasinewton/quasinewton.html

[2] J. Nocedal and S. J. Wright, *Numerical optimization*, 2nd ed. in Springer series in operations research and financial engineering. Springer, 2006. doi: 10.1007/978-0-387-40065-5.