BFGS and SR1

Review BFGS.

Let $\rho_k = (y_k^T s_k)^{-1}$, $s_k = x_{k+1} - s_k$, $y_k = \nabla f_{k+1} - \nabla f_k$

And solve

$$H_{k+1} = \arg\min_{J} \; \| J - H_k \|_F$$

The new H should be as similar to the old H as possible (keep good information!)

Frobenious norm

$$\text{s.t.} \quad J = J^T$$

Stay symmetric

$$J y_k = s_k$$

Satisfy the secant equation

The results:

$$H_{k+1} = \underbrace{(I - \rho_k s_k y_k^T) H_k (I - \rho_k y_k s_k^T)}_{\text{rank 1 update}} + \underbrace{\underbrace{\rho_k}_{\text{scalar}} \underbrace{s_k s_k^T}_{\text{matrix}}}_{\text{rank 1 matrix}}$$

- We want H to be positive definite so that the Newton step $p_k = -H_{k+1} \nabla f_k$ results in $x_{k+1} = x_k + p_k$ the minimizer of the quadratic model $m(p) = f_k + \nabla f_k^T p + \frac{1}{2} p^T \nabla^2 f_k \, p$. If H is not pos. def. then the Newton step may or may not be a descent direction of f.

- H_{k+1} will be positive definite
  (a) H_k is positive definite, and
  (b) $y_k^T s_k > 0$.

  Condition (a) can be met with choice of H_0.
  Condition (b) can be met by using a linesearch with strong Wolfe conditions.

Consider Poorly Conditioned Hessian Matrices.

If $y_k^T s_k$ is very small (or if not implementing strong Wolfe conditions) or negative then the Newton step becomes unreliable. We can find a "nearby" matrix which is more strongly positive definite.

Suppose H has eigenvalues $\lambda_1 \le \lambda_2 \le \cdots \le \lambda_n$ with corresponding eigenvectors $u_1, u_2, \ldots, u_n$. That is, $Hu_i = \lambda_i u_i$.

The condition number of symmetric matrix H is defined as $c = \frac{|\lambda_n|}{|\lambda_1|}$.

If $\lambda_1 < 0$ then H is indefinite.

If $|\lambda_1| \ll |\lambda_n|$ then H is poorly conditioned ($c \gg 1$) or nearly indefinite.

Both of these situations can result in numerical instabilities and/or poor convergence in BFGS.

Consider the "nearby" matrix $H + \mu I$. Notice:

$$(H+\mu I)u_i = Hu_i + \mu I u_i = \lambda_i u_i + \mu u_i = (\lambda_i + \mu) u_i$$

So, u_i is also an eigenvector of $H + \mu I$, but with eigenvalue $\lambda_i + \mu$.

Thus, by adding a multiple of I to H we create a new symmetric matrix with similar eigenstructure. In particular,

$$c_\mu = \frac{|\lambda_n + \mu|}{|\lambda_1 + \mu|}$$

so that we can adjust the condition number for robust matrix computations.

Why should we trust a modified model function?

Ultimately, we want to know if the model provides a descent direction.

$$m(p) = f_k + \nabla f_k^T p + \tfrac{1}{2} p^T B_k p$$

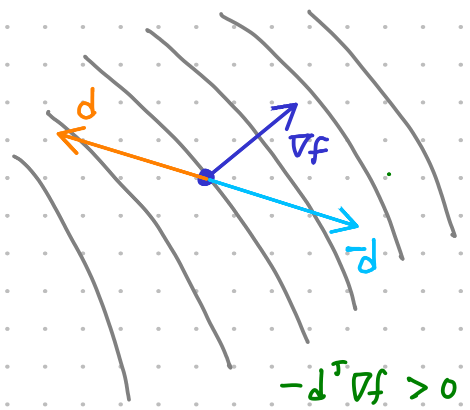
$$m_\mu(p) = f_k + \nabla f_k^T p + \tfrac{1}{2} p^T (B_k + \mu I) p$$

① The gradient descent direction is the same for $m(p)$ and $m_\mu(p)$ $\quad (-\nabla f_k)$

② The newton step for $m_\mu(p)$ is $\quad d = -(B_k + \mu I)^{-1} \nabla f_k$

This is a descent direction for $f(x)$ and $m(p)$ if $\quad -d^T \nabla f_k > 0$

Suppose μ such that $\lambda_1 + \mu > 0$. Then



$$
\begin{aligned}
-d^T \nabla f_k &= \left((B_k + \mu I)^{-1} \nabla f_k \right)^T \nabla f_k \\
&= \nabla f_k^T (B_k + \mu I)^{-1} \nabla f_k \\
&\geq \nabla f_k^T (\lambda_n + \mu)^{-1} \nabla f_k \\
&= (\lambda_n + \mu)^{-1} \| \nabla f_k \|^2 \\
&> 0
\end{aligned}
$$

$(\lambda_n + \mu)^{-1}$ is the minimum eigenvalue

We can trust the model $m_\mu(p)$ because its Newton step is a descent direction of our function f at $p=0$.

The SR1 method (symmetric rank 1)

This quasi Newton update is simpler than BFGS. We will derive the update and then seek to understand its properties. It will be useful in the right contexts.

We seek update $\underbrace{B_{k+1} = B_k + \sigma v v^T}_{\text{ansatz}}$ for $\sigma = \pm 1$

and satisfying the secant equation $y_k = B_{k+1} s_k$.

Apply the ansatz to the secant equation:

$$
\begin{aligned}
y_k &= (B_k + \sigma v v^T) s_k \\
&= B_k s_k + \sigma v v^T s_k \qquad (v^T s_k \text{ is a scalar}) \\
&= B_k s_k + \sigma (v^T s_k) v
\end{aligned}
$$

$$\Rightarrow \quad y_k - B_k s_k = (\sigma v^T s_k) v$$

so v is some scalar multiple of $y_k - B_k s_k$. Let $v = a(y_k - B_k s_k)$, then we have

$$
\begin{aligned}
y_k - B_k s_k &= \sigma a^2 (y_k - B_k s_k)^T s_k (y_k - B_k s_k) \\
y_k - B_k s_k &= \left[\sigma a^2 (y_k - B_k s_k)^T s_k\right] (y_k - B_k s_k)
\end{aligned}
$$

So, $\sigma a^2 (y_k - B_k s_k)^T s_k = 1$. Allowing for $\sigma = \pm 1$, we have the solution

$$\sigma = \pm 1 \qquad a = \pm \left| (y_k - B_k s_k)^T s_k \right|^{-1/2}$$

with associated update (H_{k+1} is similarly determined)

$$B_{k+1} = B_k + \frac{w w^T}{s_k^T w}, \quad w = y_k - B_k s_k$$

$$H_{k+1} = H_k + \frac{z z^T}{y_k^T z}, \quad z = s_k - H_k y_k$$

rank 1 updates and easy to compute!

But ...

- The constant "a" may not exist. This occurs when either

  (a) $y_k = B_k s_k$

  Here, the updates $y_k = \nabla f_{k+1} - \nabla f_k$ and $s_k = x_{k+1} - x_k$ already satisfy the secant equation. That would indicate that no update is necessary because B and H already capture all known curvature information. $B_{k+1} = B_k$.

  or (b) $(y_k - B_k s_k)^T s_k = 0$

  Here, we have an unfortunate step s_k which is orthogonal to any corrections induced by s_k. That is, no rank 1 update exists that satisfies the secant equation.

- Also, σ may be -1. In this case $s_k^T y_k < s_k^T B_k s_k$.

  This means that the curvature is low compared to the prediction of B. In other words, the function appears concave in direction s_k.

  $\Rightarrow$ the resulting B_{k+1} may not be positive definite.

Though ...

- In practice, the SR1 method is known to more quickly generate good approximate Hessian matrices compared to rank 2 update methods.

Convergence Theorems for BFGS

Theorem Let B_0 be any symmetric positive definite initial matrix, $f(x)$ be twice continuously differentiable on $\mathbb{R}^n$, and let x_0 be any initial point for which $L = \{x \in \mathbb{R}^n \mid f(x) \leq f(x_0)\}$ is convex. Furthermore, suppose there exist positive scalars m and M such that $m\|z\|^2 \leq z^T \nabla^2 f(x) z \leq M\|z\|^2$ for all $z \in \mathbb{R}^n$ and $x \in L$. Then the sequence $\{x_k\}$ generated by the BFGS algorithm (with $\epsilon = 0$) converges to the minimizer x^* of f.

* $f(x)$ is twice continuously differentiable because we need to enforce some properties of the hessian.

* $L = \{x \in \mathbb{R}^n \mid f(x) \leq f(x_0)\}$ is convex. The sublevel set of value $f(x_0)$ is convex. Because of * we also have a global minimizer in L. This is a strong assumption, but we need it.

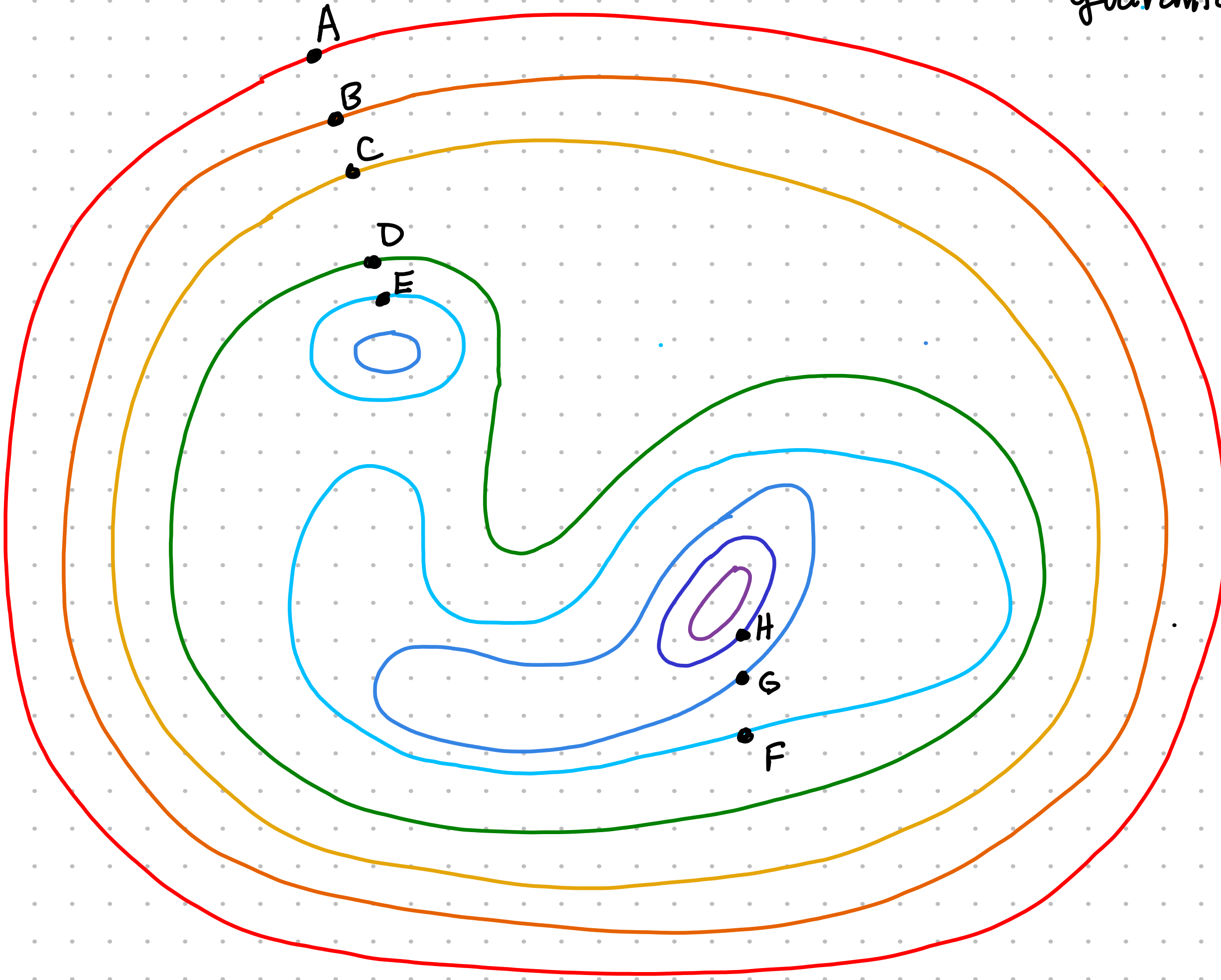
* $m\|z\|^2 \leq z^T \nabla^2 f(x) z \leq M\|z\|^2$. This condition tells us that the hessian is positive definite in L and the eigenvalues are bounded away from zero. (strongly convex).

* $\{x_k\} \to x^*$. We get convergence to the global minimizer of f on $\mathbb{R}^n$ (not just on L)

The main theorem is surprisingly weak in that the assumptions are not met for most interesting optimization problems. However, we can say that once x_k satisfies all of the conditions, then convergence is sure. Also consider:

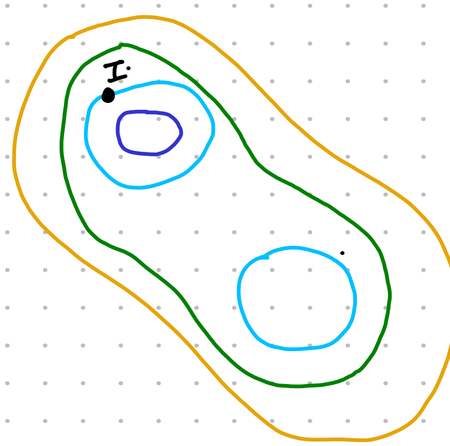
Suppose x^* is a strict local minimizer of twice continuously differentiable function $f(x)$. Then there exists a neighborhood $N(x^*, r)$ over which $\nabla^2 f(x) > 0$. (Note: N is convex)

Consider the smooth function shown below as a contour plot. Assume that $L = \{ x \in \mathbb{R}^n \mid f(x) \leq f(A) \}$ is convex. For which x_0 is convergence to x^* guaranteed?



x_0	$\to x^*$?	L convex?	$\nabla^2 f(x) \succ 0$
A	maybe	yes	no
B	maybe	yes	no
C	maybe	yes	no
D	maybe	no	no
E	maybe	yes	no
F	maybe	no	no
G	maybe	no	no
H	yes.	yes	yes
I	maybe	no	yes

separate example



Theorem Let B_0 be any symmetric positive definite initial matrix, $f(x)$ be twice continuously differentiable on $D \subseteq \mathbb{R}^n$ and let $x_0 \in D$. Furthermore, suppose there exist positive scalar M such that $|z^T \nabla^2 f(x) z| \leq M \|z\|^2$ for all $z \in \mathbb{R}^n$ and $x \in L$. Finally, suppose $X = \{x \in \mathbb{R}^n \mid \nabla f(x) = 0, \nabla^2 f(x) > 0\} \neq \phi$. Then the sequence $\{x_k\}$ generated by the BFGS algorithm (with $\epsilon = 0$) converges superlinearly to some $x^* \in X$.

Local convergence result is very strong.

Paraphrased:

"If at least one strict local minimizer of f exists then BFGS will find one."

A sequence $\{x_k\}$ is said to converge to x^* if $\lim_{k \to \infty} \|x_k - x^*\| = 0$.

$\{x_k\}$ converges linearly to x^* if there exists $\tilde{k}$ and $c \in [0,1)$ such that

$$\|x_{k+1} - x^*\| \leq c \|x_k - x^*\| \text{ for all } k > \tilde{k}.$$

$\{x_k\}$ converges quadratically to x^* if there exists $\tilde{k}$ and $c \in [0,1)$ such that

$$\|x_{k+1} - x^*\| \leq c \|x_k - x^*\|^2 \text{ for all } k > \tilde{k}.$$

$\{x_k\}$ converges superlinearly to x^* if there exists $\tilde{k}$ and $\{c_k\} \to 0$ such that

$$\|x_{k+1} - x^*\| \leq c_k \|x_k - x^*\| \text{ for all } k > \tilde{k}.$$