Numerical Optimization: Understanding L-BFGS

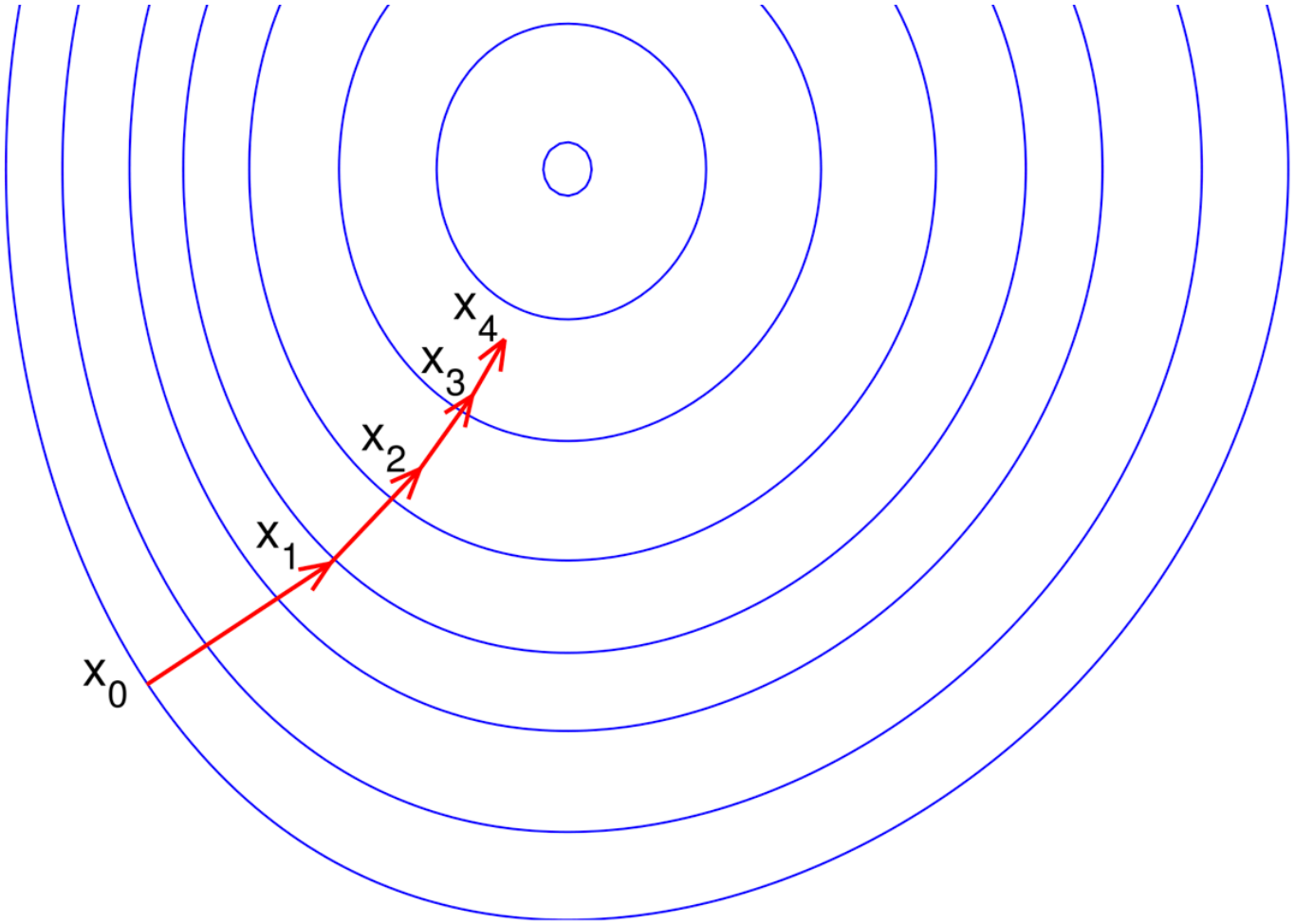
Numerical optimization is at the core of much of machine learning. Once you've defined your model and have a dataset ready, estimating the parameters of your model typically boils down to minimizing some multivariate function $\mathrm{f(x)}$ $f(x)$, where the input x x is in some high-dimensional space and corresponds to model parameters. In other words, if you solve:

$$\mathrm{x}^* = \arg\min_{\mathrm{x}} \mathrm{f(x)}$$
$$x^* = \arg\min_{x} f(x)$$

then x^* x^* is the 'best' choice for model parameters according to how you've set your objective.[1]

In this post, I'll focus on the motivation for the L-BFGS algorithm for unconstrained function minimization, which is very popular for ML problems where 'batch' optimization makes sense. For larger problems, online methods based around stochastic gradient descent have gained popularity, since they require fewer passes over data to converge. In a later post, I might cover some of these techniques, including my personal favorite AdaDelta.

Note: Throughout the post, I'll assume you remember multivariable calculus. So if you don't recall what a gradient or Hessian is, you'll want to bone up first.



Newton's Method

Most numerical optimization procedures are iterative algorithms which consider a sequence of 'guesses' $\mathrm{x_n}\, x_n$ which ultimately converge to $\mathrm{x}^*\, x^*$ the true global minimizer of $\mathrm{f}\, f$. Suppose, we have an estimate $\mathrm{x_n}\, x_n$ and we want our next estimate $\mathrm{x_{n+1}}\ x_{n+1}$ to have the property that $\mathrm{f(x_{n+1}}) < \mathrm{f(x_n})\quad f(x_{n+1}) < f(x_n)$.

Newton's method is centered around a quadratic approximation of $\mathrm{f}\, f$ for points near $\mathrm{x_n}\, x_n$. Assuming that $\mathrm{f}\, f$ is twice-differentiable, we can use a quadratic approximation of $\mathrm{f}\, f$ for points 'near' a fixed point $\mathrm{x}\, x$ using a Taylor expansion:

$$\mathrm{f(x+\Delta x)} \approx \mathrm{f(x)} + \mathrm{\Delta x^T \nabla f(x)} + \frac{1}{2}\mathrm{\Delta x^T\ (\nabla^2 f(x))\, \Delta x}$$

$$f(x+\Delta x) \quad \approx f(x) + \Delta x^T \nabla f(x) + \frac{1}{2}\Delta x^T \left(\nabla^2 f(x)\right) \Delta x$$

where $\nabla f(x)$ and $\nabla^2 f(x)$ are the gradient and Hessian of f at the point x_n. This approximation holds in the limit as $||\Delta x|| \to 0$. This is a generalization of the single-dimensional Taylor polynomial expansion you might remember from Calculus.

In order to simplify much of the notation, we're going to think of our iterative algorithm of producing a sequence of such quadratic approximations h_n. Without loss of generality, we can write $x_{n+1} = x_n + \Delta x$ and re-write the above equation,

$$h_n(\Delta x) = f(x_n) + \Delta x^T g_n + \frac{1}{2}\Delta x^T H_n \Delta x$$

where g_n and H_n represent the gradient and Hessian of f at x_n.

We want to choose Δx to minimize this local quadratic approximation of f at x_n. Differentiating with respect to Δx above yields:

$$\frac{\partial h_n(\Delta x)}{\partial \Delta x} = g_n + H_n \Delta x$$

Recall that any Δx which yields $\frac{\partial h_n(\Delta x)}{\partial \Delta x} = 0$ is a local extrema of $h_n(\cdot)$. If we assume that H_n is [postive definite] (psd) then we know this Δx is also a global minimum for $h_n(\cdot)$. Solving for Δx:[2]

$$\Delta x = -H_n^{-1} g_n$$

This suggests $H_n^{-1} g_n$ as a good direction to move x_n towards. In practice, we set $x_{n+1} = x_n - \alpha(H_n^{-1} g_n)$ for a value of α where $f(x_{n+1})$ is 'sufficiently' smaller than $f(x_n)$.

Iterative Algorithm

The above suggests an iterative algorithm:

$$
\begin{aligned}
&\text{NewtonRaphson}(\mathrm{f}, \mathrm{x}_0):\\
&\quad \text{For } \mathrm{n} = 0, 1, \ldots \text{ (until converged)}:\\
&\qquad \text{Compute } \mathrm{g}_\mathrm{n} \text{ and } \mathrm{H}_\mathrm{n}^{-1} \text{ for } \mathrm{x}_\mathrm{n}\\
&\qquad \mathrm{d} = \mathrm{H}_\mathrm{n}^{-1}\,\mathrm{g}_\mathrm{n}\\
&\qquad \alpha = \min_{\alpha \geq 0} \mathrm{f}(\mathrm{x}_\mathrm{n} - \alpha \mathrm{d})\\
&\qquad \mathrm{x}_{\mathrm{n}+1} \leftarrow \mathrm{x}_\mathrm{n} - \alpha \mathrm{d}
\end{aligned}
$$

$$
\begin{aligned}
&NewtonRaphson(f, x_0):\\
&\quad \text{For } n = 0, 1, \ldots \text{ (until converged)}:\\
&\qquad \text{Compute } g_n \text{ and } H_n^{-1} \text{ for } x_n\\
&\qquad d = H_n^{-1} g_n\\
&\qquad \alpha = \min_{\alpha \geq 0} f(x_n - \alpha d)\\
&\qquad x_{n+1} \leftarrow x_n - \alpha d
\end{aligned}
$$

The computation of the α step-size can use any number of line search algorithms. The simplest of these is backtracking line search, where you simply try smaller and smaller values of α until the function value is 'small enough'.

In terms of software engineering, we can treat NewtonRaphson $NewtonRaphson$ as a blackbox for any twice-differentiable function which satisfies the Java interface:

```
public interface TwiceDifferentiableFunction {
  // compute f(x)

  public double valueAt(double[] x);

  // compute grad f(x)

  public double[] gradientAt(double[] x);

  // compute inverse hessian H^-1

  public double[][] inverseHessian(double[] x);
}
```

With quite a bit of tedious math, you can prove that for a convex function, the above procedure will converge to a unique global minimizer $\mathrm{x}^{*} x^{*}$, regardless of the choice of $\mathrm{x}_{0} x_{0}$. For non-convex functions that arise in ML (almost all latent variable models or deep nets), the procedure still works but is only guranteed to converge to a local minimum. In practice, for non-convex optimization, users need to pay more attention to initialization and other algorithm details.

Huge Hessians

The central issue with $\mathrm{NewtonRaphson}$ $NewtonRaphson$ is that we need to be able to compute the inverse Hessian matrix.[3] Note that for ML applications, the dimensionality of the input to $\mathrm{f} f$ typically corresponds to model parameters. It's not unusual to have hundreds of millions of parameters or in some vision applications even billions of parameters. For these reasons, computing the hessian or its inverse is often impractical. For many functions, the hessian may not even be analytically computable, let along representable.

Because of these reasons, $\mathrm{NewtonRaphson}$ $NewtonRaphson$ is rarely used in practice to optimize functions corresponding to large problems. Luckily, the above algorithm can still work even if $\mathrm{H}_{\mathrm{n}}^{-1}$ H_n^{-1} doesn't correspond to the exact inverse hessian at $\mathrm{x}_{\mathrm{n}} x_n$, but is instead a good approximation.

Quasi-Newton

Suppose that instead of requiring $\mathrm{H}_{\mathrm{n}}^{-1}$ H_n^{-1} be the inverse hessian at $\mathrm{x}_{\mathrm{n}} x_n$, we think of it as an approximation of this information. We can generalize $\mathrm{NewtonRaphson}$ $NewtonRaphson$ to take a $\mathrm{QuasiUpdate}$ $\mathrm{QuasiUpdate}$ policy which is responsible for producing a sequence of $\mathrm{H}_{\mathrm{n}}^{-1}$ H_n^{-1}.

$$\text{QuasiNewton}(f, x_0, H_0^{-1}, \text{QuasiUpdate}):$$
$$\text{For } n = 0, 1, \ldots \text{ (until converged)}:$$
$$\text{// Compute search direction and step-size}$$
$$d = H_n^{-1} g_n$$
$$\alpha \leftarrow \min_{\alpha \geq 0} f(x_n - \alpha d)$$
$$x_{n+1} \leftarrow x_n - \alpha d$$
$$\text{// Store the input and gradient deltas}$$
$$g_{n+1} \leftarrow \nabla f(x_{n+1})$$
$$s_{n+1} \leftarrow x_{n+1} - x_n$$
$$y_{n+1} \leftarrow g_{n+1} - g_n$$
$$\text{// Update inverse hessian}$$
$$H_{n+1}^{-1} \leftarrow \text{QuasiUpdate}(H_n^{-1}, s_{n+1}, y_{n+1})$$

$$QuasiNewton(f, x_0, H_0^{-1}, \text{QuasiUpdate}):$$
$$\text{For } n = 0, 1, \ldots \text{ (until converged)}:$$
$$\text{// Compute search direction and step-size}$$
$$d = H_n^{-1} g_n$$
$$\alpha \leftarrow \min_{\alpha \geq 0} f(x_n - \alpha d)$$
$$x_{n+1} \leftarrow x_n - \alpha d$$
$$\text{// Store the input and gradient deltas}$$
$$g_{n+1} \leftarrow \nabla f(x_{n+1})$$
$$s_{n+1} \leftarrow x_{n+1} - x_n$$
$$y_{n+1} \leftarrow g_{n+1} - g_n$$
$$\text{// Update inverse hessian}$$
$$H_{n+1}^{-1} \leftarrow \text{QuasiUpdate}(H_n^{-1}, s_{n+1}, y_{n+1})$$

We've assumed that QuasiUpdate QuasiUpdate only requires the former inverse hessian estimate as well tas the input and gradient differences (s_n s_n and y_n y_n respectively). Note that if QuasiUpdate QuasiUpdate just returns $\nabla^2 f(x_{n+1})$ $\nabla^2 f(x_{n+1})$, we recover exact NewtonRaphson NewtonRaphson.

In terms of software, we can blackbox optimize an arbitrary differentiable function (with no need to be able to compute a second derivative) using QuasiNewton $QuasiNewton$ assuming we get a quasi-newton approximation update policy. In Java this might look like this,

```
public interface DifferentiableFunction {
  // compute f(x)

  public double valueAt(double[] x);

  // compute grad f(x)

  public double[] gradientAt(double[] x);
}

public interface QuasiNewtonApproximation {
  // update the H^{-1} estimate (using x_{n+1}-x_n and grad_{n+1}-grad_n)

  public void update(double[] deltaX, double[] deltaGrad);

  // H^{-1} (direction) using the current H^{-1} estimate

  public double[] inverseHessianMultiply(double[] direction);
}
```

Note that the only use we have of the hessian is via it's product with the gradient direction. This will become useful for the L-BFGS algorithm described below, since we don't need to represent the Hessian approximation in memory. If you want to see these abstractions in action, here's a link to a Java 8 and golang implementation I've written.

Behave like a Hessian

What form should $\mathrm{QuasiUpdate}$ QuasiUpdate take? Well, if we have $\mathrm{QuasiUpdate}$ QuasiUpdate always return the identity matrix (ignoring its inputs), then this corresponds to simple gradient descent, since the search direction is always $\nabla \mathrm{f}_{\mathrm{n}}$ ∇f_n. While this actually yields a valid procedure which will converge to x^* x^* for convex f f, intuitively this choice of $\mathrm{QuasiUpdate}$ QuasiUpdate isn't attempting to capture second-order information about f f.

Let's think about our choice of H_{n} H_n as an approximation for f f near x_{n} x_n:

$$\mathrm{h}_{\mathrm{n}}(\mathrm{d}) = \mathrm{f}(\mathrm{x}_{\mathrm{n}}) + \mathrm{d}^{\mathrm{T}}\mathrm{g}_{\mathrm{n}} + \frac{1}{2}\mathrm{d}^{\mathrm{T}}\mathrm{H}_{\mathrm{n}}\mathrm{d}$$

$$h_n(d) = f(x_n) + d^T g_n + \frac{1}{2} d^T H_n d$$

Secant Condition

A good property for $\mathrm{h_n(d)}$ $h_n(d)$ is that its gradient agrees with f f at $\mathrm{x_n}$ x_n and $\mathrm{x_{n-1}}$ x_{n-1}. In other words, we'd like to ensure:

$$\begin{aligned}\nabla \mathrm{h_n(x_n)} &= \mathrm{g_n} \\ \nabla \mathrm{h_n(x_{n-1})} &= \mathrm{g_{n-1}} \\ \nabla h_n(x_n) &= g_n \\ \nabla h_n(x_{n-1}) &= g_{n-1}\end{aligned}$$

Using both of the equations above:

$$\begin{aligned}\nabla \mathrm{h_n(x_n)} - \nabla \mathrm{h_n(x_{n-1})} &= \mathrm{g_n} - \mathrm{g_{n-1}} \\ \nabla h_n(x_n) - \nabla h_n(x_{n-1}) &= g_n - g_{n-1}\end{aligned}$$

Using the gradient of $\mathrm{h_{n+1}}$ $(\cdot)$ $h_{n+1}(\cdot)$ and canceling terms we get

$$\mathrm{H_n(x_n - x_{n-1})} = \mathrm{(g_n - g_{n-1})}$$

$$H_n(x_n - x_{n-1}) = (g_n - g_{n-1})$$

This yields the so-called "secant conditions" which ensures that $\mathrm{H_{n+1}}$ H_{n+1} behaves like the Hessian at least for the diference $\mathrm{(x_n - x_{n-1})}$ $(x_n - x_{n-1})$. Assuming $\mathrm{H_n}$ H_n is invertible (which is true if it is psd), then multiplying both sides by $\mathrm{H_n^{-1}}$ H_n^{-1} yields

$$\begin{aligned}\mathrm{H_n^{-1} y_n} &= \mathrm{s_n} \\ H_n^{-1} y_n &= s_n\end{aligned}$$

where $\mathrm{y_{n+1}}$ y_{n+1} is the difference in gradients and $\mathrm{s_{n+1}}$ s_{n+1} is the difference in inputs.

Symmetric

Recall that the a hessian represents the matrix of 2nd order partial derivatives:
$\mathrm{H^{(i,j)}} = \partial \mathrm{f}/\partial \mathrm{x_i} \partial \mathrm{x_j}$ $H^{(i,j)} = \partial f \, / \, \partial x_i \partial x_j$. The hessian is symmetric since the order of differentiation doesn't matter.

The BFGS Update

Intuitively, we want $\mathrm{H}_{\mathrm{n}}\, H_n$ to satisfy the two conditions above:

- Secant condition holds for $\mathrm{s}_{\mathrm{n}}\, s_n$ and $\mathrm{y}_{\mathrm{n}}\, y_n$
- $\mathrm{H}_{\mathrm{n}}\, H_n$ is symmetric

Given the two conditions above, we'd like to take the most conservative change relative to $\mathrm{H}_{\mathrm{n-1}}\ H_{n-1}$. This is reminiscent of the MIRA update, where we have conditions on any good solution but all other things equal, want the 'smallest' change.

$$\begin{aligned} \min_{\mathrm{H}^{-1}} \quad & \|\mathrm{H}^{-1} - \mathrm{H}_{\mathrm{n-1}}^{-1}\|^2 \\ \text{s.t.} \quad & \mathrm{H}^{-1}\,\mathrm{y}_{\mathrm{n}} = \mathrm{s}_{\mathrm{n}} \\ & \mathrm{H}^{-1} \text{ is symmetric} \end{aligned}$$

$$\begin{aligned} \min_{H^{-1}} \quad & \|H^{-1} - H_{n-1}^{-1}\|^2 \\ \text{s.t.} \quad & H^{-1} y_n = s_n \\ & H^{-1} \text{ is symmetric} \end{aligned}$$

The norm used here $\|\cdot\|$ $\|\cdot\|$ is the weighted frobenius norm.[4] The solution to this optimization problem is given by

$$\mathrm{H}_{\mathrm{n+1}}^{-1} = (\mathrm{I} - \rho_{\mathrm{n}}\,\mathrm{y}_{\mathrm{n}}\,\mathrm{s}_{\mathrm{n}}^{\mathrm{T}})\mathrm{H}_{\mathrm{n}}^{-1}\,(\mathrm{I} - \rho_{\mathrm{n}}\,\mathrm{s}_{\mathrm{n}}\,\mathrm{y}_{\mathrm{n}}^{\mathrm{T}}) + \rho_{\mathrm{n}}\,\mathrm{s}_{\mathrm{n}}\,\mathrm{s}_{\mathrm{n}}^{\mathrm{T}}$$

$$H_{n+1}^{-1} = (I - \rho_n y_n s_n^T) H_n^{-1} (I - \rho_n s_n y_n^T) + \rho_n s_n s_n^T$$

where $\rho_{\mathrm{n}} = (\mathrm{y}_{\mathrm{n}}^{\mathrm{T}}\,\mathrm{s}_{\mathrm{n}})^{-1}$ $\rho_n = (y_n^T s_n)^{-1}$. Proving this is relatively involved and mostly symbol crunching. I don't know of any intuitive way to derive this unfortunately.

This update is known as the Broyden–Fletcher–Goldfarb–Shanno (BFGS) update, named after the original authors. Some things worth noting about this update:

- $\mathrm{H}_{\mathrm{n+1}}^{-1}\ H_{n+1}^{-1}$ is positive definite (psd) when $\mathrm{H}_{\mathrm{n}}^{-1}\ H_{n}^{-1}$ is. Assuming our initial guess of $\mathrm{H}_{0} H_0$ is psd, it follows by induction each inverse Hessian estimate is as well. Since we can choose any $\mathrm{H}_{0}^{-1}\ H_{0}^{-1}$ we want, including the $\mathrm{I}\ I$ matrix, this is easy to ensure.

- The above also specifies a recurrence relationship between $\mathrm{H}_{\mathrm{n+1}}^{-1}\ H_{n+1}^{-1}$ and $\mathrm{H}_{\mathrm{n}}^{-1}\ H_{n}^{-1}$. We only need the history of $\mathrm{s}_{\mathrm{n}}\ s_n$ and $\mathrm{y}_{\mathrm{n}}\ y_n$ to re-construct $\mathrm{H}_{\mathrm{n}}^{-1}\ H_{n}^{-1}$.

The last point is significant since it will yield a procedural algorithm for computing $\mathrm{H}_{\mathrm{n}}^{-1}\ \mathrm{d}H_{n}^{-1}d$, for a direction $\mathrm{d}d$, without ever forming the $\mathrm{H}_{\mathrm{n}}^{-1}\ H_{n}^{-1}$ matrix. Repeatedly applying the recurrence above we have

$$
\begin{aligned}
&\text{BFGSMultiply}(H_0^{-1}, \{s_k\}, \{y_k\}, d):\\
&\quad r \leftarrow d\\
&\quad \text{// Compute right product}\\
&\quad \text{for } i = n, \ldots, 1:\\
&\qquad \alpha_i \leftarrow \rho_i s_i^T r\\
&\qquad r \leftarrow r - \alpha_i y_i\\
&\quad \text{// Compute center}\\
&\quad r \leftarrow H_0^{-1} r\\
&\quad \text{// Compute left product}\\
&\quad \text{for } i = 1, \ldots, n:\\
&\qquad \beta \leftarrow \rho_i y_i^T r\\
&\qquad r \leftarrow r + (\alpha_{n-i+1} - \beta) s_i\\
&\quad \text{return } r
\end{aligned}
$$

$$
\begin{aligned}
&BFGSMultiply(H_0^{-1}, \{s_k\}, \{y_k\}, d):\\
&\quad r \leftarrow d\\
&\quad \text{// Compute right product}\\
&\quad \text{for } i = n, \ldots, 1:\\
&\qquad \alpha_i \leftarrow \rho_i s_i^T r\\
&\qquad r \leftarrow r - \alpha_i y_i\\
&\quad \text{// Compute center}\\
&\quad r \leftarrow H_0^{-1} r\\
&\quad \text{// Compute left product}\\
&\quad \text{for } i = 1, \ldots, n:\\
&\qquad \beta \leftarrow \rho_i y_i^T r\\
&\qquad r \leftarrow r + (\alpha_{n-i+1} - \beta) s_i\\
&\quad \text{return } r
\end{aligned}
$$

Since the only use for H_n^{-1} H_n^{-1} is via the product $H_n^{-1} g_n$ $H_n^{-1} g_n$, we only need the above procedure to use the BFGS approximation in QuasiNewton QuasiNewton.

L-BFGS: BFGS on a memory budget

The BFGS quasi-newton approximation has the benefit of not requiring us to be able to analytically compute the Hessian of a function. However, we still must maintain a history of the s_n s_n and y_n y_n vectors for each iteration. Since one of the core-concerns of the NewtonRaphson $NewtonRaphson$ algorithm were the memory requirements associated with

maintaining an Hessian, the BFGS Quasi-Newton algorithm doesn't address that since our memory use can grow without bound.

The L-BFGS algorithm, named for *limited* BFGS, simply truncates the $\mathrm{BFGSMultiply}$ $BFGSMultiply$ update to use the last m m input differences and gradient differences. This means, we only need to store $\mathrm{s_n}, \mathrm{s_{n-1}}, \ldots, \mathrm{s_{n-m-1}}$ $s_n, s_{n-1}, \ldots, s_{n-m-1}$ and $\mathrm{y_n}, \mathrm{y_{n-1}}, \ldots, \mathrm{y_{n-m-1}}$ $y_n, y_{n-1}, \ldots, y_{n-m-1}$ to compute the update. The center product can still use any symmetric psd matrix $\mathrm{H_0^{-1}}$ H_0^{-1}, which can also depend on any $\{\mathrm{s_k}\}$ $\{s_k\}$ or $\{\mathrm{y_k}\}$ $\{y_k\}$.

L-BFGS variants

There are lots of variants of L-BFGS which get used in practice. For non-differentiable functions, there is an othant-wise varient which is suitable for training $\mathrm{L_1}$ L_1 regularized loss.

One of the main reasons to *not* use L-BFGS is in very large data-settings where an online approach can converge faster. There are in fact online variants of L-BFGS, but to my knowledge, none have consistently out-performed SGD variants (including AdaGrad or AdaDelta) for sufficiently large data sets.

1. This assumes there is a unique global minimizer for f f. In practice, in practice unless f f is convex, the parameters used are whatever pops out the other side of an iterative algorithm. ↩

2. We know $-\mathrm{H^{-1}}\,\nabla\mathrm{f}$ $-H^{-1}\nabla f$ is a local extrema since the gradient is zero, since the Hessian has positive curvature, we know it's in fact a local minima. If f f is convex, we know the Hessian is always positive definite and we know there is a single unique global minimum. ↩

3. As we'll see, we really on require being able to multiply by $\mathrm{H^{-1}}\,\mathrm{d}$ $H^{-1}d$ for a direction d d. ↩

4. I've intentionally left the weighting matrix W W used to weight the norm since you get the same solution under many choices. In particular for any positive-definite W W such that $\mathrm{W\,s_n} = \mathrm{y_n}$ $Ws_n = y_n$, we get the same solution. ↩