Optimization

Muchang Bahng

Spring 2025

Contents

Optimization is such an important tool that it deserves a set of notes in itself. All problems in model training essentially stems from non-ideal optimization. Knowing the strengths and weaknesses of each optimizer allows you to diagnose which ones to use.

Generally, we (non-exclusively) categorize optimization algorithms as such:

1. *Convex*? Convex optimization is pretty easy to solve and has been studied extensively.
2. *Constrained*? Is the parameter space constrained to a certain manifold?
3. *Order*. Do we use derivatives at all? First-order derivatives (gradient)? Second-order (Hessian)?

These algorithms try to solve the following potential problems.

1. *Convergence*. Do we converge to some point?
2. *Optimality*. Is this point close to the true global minima?
3. *Efficiency*. Can we iterate efficiently?

As a benchmark test, the following function will be used a lot.

Definition 0.1 (Rosenbrock Function)

The **Rosenbrock function** is defined

$$f(x, y) = (a - x)^2 + b(y - x^2)^2 \tag{1}$$

which has a global minimum at (a, b).

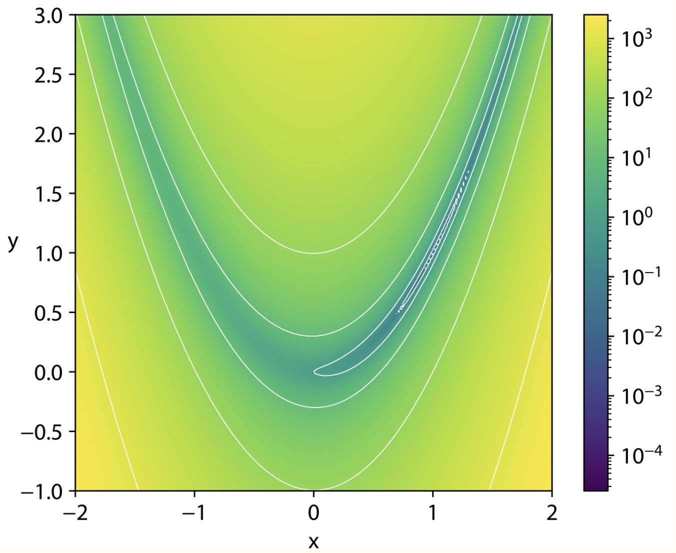


Figure 1: Typically, we set $a = 1, b = 100$.

1 Gradient Methods

The first thing you learn about gradients in multivariate calculus is that they point in the step of steepest ascent. Therefore,

1.1 Newton-Raphson Method

1.2 Stochastic Gradient Descent

Note that gradient computation is generally very expensive and not scalable as n gets high. Given a dataset $\mathcal{D} = \{d_i\}_i$ of D points, our posterior is of the form $p(\theta \mid \mathcal{D}) \propto p(\mathcal{D} \mid \theta)\, p(\theta)$ and so

$$\nabla_\theta \log p(\theta \mid \mathcal{D}) = \nabla_\theta \log p(\theta) + \nabla_\theta \log p(\mathcal{D} \mid \theta) = \nabla_\theta \log p(\theta) + \sum_i \nabla_\theta \log p(d_i \mid \theta) \tag{2}$$

We can approximate this gradient by taking a minibatch of $\mathcal{D}$. Let us take a minibatch of m samples $M_m(\mathcal{D})$ without replacement, where $m << D$. Then, our approximation of the gradient of the log likelihood is

$$\nabla_\theta \log p(\mathcal{D} \mid \theta) \approx \nabla_\theta \log p(M_m(\mathcal{D}) \mid \theta) \coloneqq \frac{D}{m} \sum_{d \in M_m(\mathcal{D})} \nabla_\theta \log p(d \mid \theta) \tag{3}$$

and thus our noisy gradient approximation of the gradient of the log posterior is

$$\nabla_\theta \log p(\theta \mid \mathcal{D}) \approx \nabla_\theta \log p(\theta \mid M_m(\mathcal{D})) \coloneqq \nabla_\theta \log p(\theta) + \nabla_\theta \log p(M_m(\mathcal{D}) \mid \theta) \tag{4}$$

The classical gradient ascent algorithm simply optimizes a concave function, or if f is multimodal, finds a local maxima. When we use the entire $\mathcal{D}$ to compute the gradient, we call this a *batch gradient descent*, and if the minibatch estimate of the gradient is used, then this is called *stochastic gradient descent*. Ideally, we would want to have a variable step size $h(t)$ so that $h \to 0$ as $t \to +\infty$.

Algorithm 1 Stochastic Gradient Ascent

```
Require: Initial $\boldsymbol{\theta}_0$, Stepsize function $h(t)$, Minibatch size $m$
  for $t = 0$ to $T$ until convergence, do
    $\hat{g}(\theta_t) \leftarrow \nabla_\theta \log p(\theta_t \mid M_m(\mathcal{D}))$
    $\theta_{t+1} \leftarrow \theta_t + h(t) \cdot \hat{g}(\theta_t)$
  end for
```

SGD with momentum.

We have assumed knowledge of gradient descent in the back propagation step in the previous section, but let's revisit this by looking at linear regression. Given our dataset $\mathcal{D} = \{\mathbf{x}^(n), y^{(n)}\}$, we are fitting a linear model of the form

$$f(\mathbf{x}; \mathbf{w}, b) = \mathbf{w}^T \mathbf{x} + b \tag{5}$$

The squared loss function is

$$\mathcal{L}(\mathbf{w}, b) = \frac{1}{2} \sum_{n=1}^N \big(y - f(\mathbf{x}; \mathbf{w}, b)\big)^2 = \frac{1}{2} \sum_{n=1}^N \big(y - (\mathbf{w}^T \mathbf{x} + b)\big)^2 \tag{6}$$

If we want to minimize this function, we can visualize it as a d-dimensional surface that we have to traverse. Recall from multivariate calculus that the gradient of an arbitrary function $\mathcal{L}$ points in the steepest direction in which $\mathcal{L}$ increases. Therefore, if we can compute the gradient of $\mathcal{L}$ and step in the *opposite direction*, then we would make the more efficient progress towards minimizing this function (at least locally). The gradient can be solved using chain rule. Let us solve it with respect to $\mathbf{w}$ and b separately first. Beginners might find it simpler to compute the gradient element-wise.

$$\begin{aligned}
\frac{\partial}{\partial w_j}\mathcal{L}(\mathbf{w}, b) &= \frac{\partial}{\partial w_j}\bigg(\frac{1}{2}\sum_{n=1}^{N}\Big(f(\mathbf{x}^{(n)}; \mathbf{w}, b) - y^{(n)}\Big)^2\bigg) && (7)\\
&= \frac{1}{2}\sum_{n=1}^{N}\frac{\partial}{\partial w_j}\Big(f(\mathbf{x}^{(n)}; \mathbf{w}, b) - y^{(n)}\Big)^2 && (8)\\
&= \frac{1}{2}\sum_{n=1}^{N} 2\Big(f(\mathbf{x}^{(n)}) - y^{(n)}\Big)\cdot\frac{\partial}{\partial w_j}\big(f(\mathbf{x}^{(n)}; \mathbf{w}, b) - y^{(n)}\big) && (9)\\
&= \frac{1}{2}\sum_{n=1}^{N} 2\Big(f(\mathbf{x}^{(n)}) - y^{(n)}\Big)\cdot\frac{\partial}{\partial w_j}\big(\mathbf{w}^T\mathbf{x}^{(n)} + b - y^{(n)}\big) && (10)\\
&= \sum_{n=1}^{N}\big(f(\mathbf{x}^{(n)}; \mathbf{w}, b) - y^{(n)}\big)\cdot x_j^{(n)} \qquad (\text{for } j = 0, 1, \ldots, d) && (11)
\end{aligned}$$

As for getting the derivative w.r.t. b, we can redo the computation and get

$$\frac{\partial}{\partial w_j}\mathcal{L}(\mathbf{w}, b) = \sum_{n=1}^{N}\big(f(\mathbf{x}^{(n)}; \mathbf{w}, b) - y^{(n)}\big) \tag{12}$$

and in the vector form, setting $\boldsymbol{\theta} = (\mathbf{w}, b)$, we can set

$$\begin{aligned}
\nabla\mathcal{L}(\mathbf{w}) &= \mathbf{X}^T(\hat{\mathbf{y}} - \mathbf{y}) && (13)\\
\nabla\mathcal{L}(b) &= (\hat{\mathbf{y}} - \mathbf{y})\cdot\mathbf{1} && (14)
\end{aligned}$$

where $\hat{\mathbf{y}}_n = f(\mathbf{x}^{(n)}; \mathbf{w}, b)$ are the predictions under our current linear model and $\mathbf{X} \in \mathbb{R}^{n\times d}$ is our design matrix. This can easily be done on a computer using a package like `numpy`. Remember that GD is really just an algorithm that updates $\boldsymbol{\theta}$ repeatedly until convergence, but there are a few problems.

1. The algorithm can be susceptible to local minima. A few countermeasures include shuffling the training set or randomly choosing initial points θ
2. The algorithm may not converge if α (the step size) is too high, since it may overshoot. This can be solved by reducing the α with each step, using *schedulers*.
3. The entire training set may be too big, and it may therefore be computationally expensive to update $\boldsymbol{\theta}$ as a whole, especially if $d >> 1$. This can be solved using stochastic gradient descent.

Rather than updating the vector $\boldsymbol{\theta}$ in batches, we can apply **stochastic gradient descent** that works incrementally by updating $\boldsymbol{\theta}$ with each term in the summation. That is, rather than updating as a batch by performing the entire matrix computation by multiplying over N dimensions,

$$\nabla\mathcal{L}(\mathbf{w}) = \underbrace{\mathbf{X}^T}_{D\times N}\underbrace{(\hat{\mathbf{y}} - \mathbf{y})}_{N\times 1} \tag{15}$$

we can reduce this load by choosing a smaller subset $\mathcal{M} \subset \mathcal{D}$ of $M < N$ elements, which gives

$$\nabla\mathcal{L}_{\mathcal{M}}(\mathbf{w}) = \underbrace{\mathbf{X}_{\mathcal{M}}^T}_{D\times M}\underbrace{(\hat{\mathbf{y}_{\mathcal{M}}} - \mathbf{y})}_{\mathcal{M}}{}_{M\times 1} \tag{16}$$

The reason we can do this is because of the following fact.

Theorem 1.1 (Unbiasedness of SGD)

$\nabla \mathcal{L}_{\mathcal{M}}(\mathbf{w})$ is an *unbiased estimator* of the true gradient. That is, setting $\mathcal{M}$ as a random variable of samples over $\mathcal{D}$, we have

$$\mathbb{E}_{\mathcal{M}}[\nabla \mathcal{L}_{\mathcal{M}}(\mathbf{w})] = \nabla \mathcal{L}(\mathbf{w}) \tag{17}$$

Proof.

We use linearity of expectation for all $\mathcal{M} \subset \mathcal{D}$ of size M.

Even though these estimators are noisy, we get to do much more iterations and therefore have a faster net rate of convergence. By using repeated chain rule, or a fancier term is automatic differentiation, as shown before, SGD can be used to optimize neural networks.

Extending beyond SGD, there are other optimizers we can use. Essentially, we are doing a highly nonconvex optimization, which doesn't have a straightforward answer, so the best we can do is play around with some properties. 0th order approximations are hopeless since the dimensions are too high, and second order approximations are hopeless either since computing the Hessian is too expensive for one run. Therefore, we must resort to some first order methods, which utilize the gradient. Some other properties to consider are:

1. Learning rate
2. Momentum
3. Batch Size

1.3 Momentum and Nesterov

1.4 Block Coordinate Descent

2 Subgradient Methods

Definition 2.1 (Convex Function)

A function $f : U \subset \mathbb{R}^n \to \mathbb{R}$ defined on a convex set U is convex if and only if for any $\mathbf{x}, \mathbf{y} \in U$

$$f\big(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}\big) \leq \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}) \tag{18}$$

Now if f is differentiable, then convexity is equivalent to

$$f(x) \geq f(y) + \nabla f(y)^T \cdot (x - y) \tag{19}$$

for all $x, y \in U$. That is, its local linear approximation always underestimates f.

It is well known that the mean square error of a linear map is convex. However, when we impose the L1 penalty, the loss function is now not differentiable at $\mathbf{0}$. Therefore, we must introduce the notion of a subgradient.

Definition 2.2 (Subgradient)

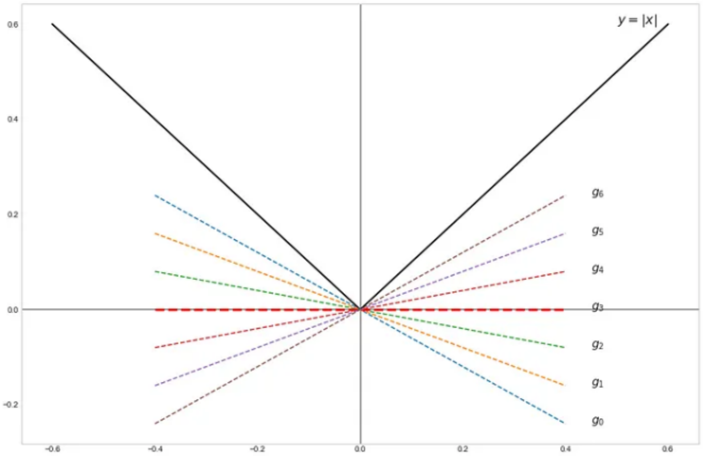
The subgradient of a convex function $f : U \subset \mathbb{R}^n \to \mathbb{R}$ is any linear map $\mathbf{A}(x) : \mathbb{R}^n \to \mathbb{R}$ such that

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \mathbf{A}(\mathbf{x})(\mathbf{y} - \mathbf{x}) \tag{20}$$

for any $\mathbf{y} \in U$. The set of all subgradients at $\mathbf{x}$ is called the **subdifferential** defined

$$\partial f(\mathbf{x}) = \{\mathbf{A} \in \mathbb{R}^n \mid \mathbf{A} \text{ is a subgradient of } f \text{ at } \mathbf{x}\} \tag{21}$$

The subgradient also acts as a linear approximation of f, but now at nondifferentiable points of convex functions, we have a set of linear approximations. It is clear that the subgradient at a differentiable point is uniquely the gradient ($\partial f(\mathbf{x}) = \{\nabla f(\mathbf{x})\}$), but for places like the absolute value, we can have infinite linear approximations.



Given the subdifferential, thus the optimality condition for any convex $\mathbf{f}$ (differentiable or not) is

$$f(\mathbf{x}^*) = \min_{\mathbf{x}} f(\mathbf{x}) \iff \mathbf{0} \in \partial f(\mathbf{x}^*) \tag{22}$$

known as the subgradient optimality condition, which clearly implies

$$f(\mathbf{y}) \geq f(\mathbf{x}^*) + \mathbf{0}^T(\mathbf{y} - \mathbf{x}^*) = f(\mathbf{x}^*) \tag{23}$$

Example 2.1 ()

The subdifferential of the absolute value function $f(x) = |x|$ at any given x is

$$\partial f(x) = \begin{cases} 1 & \text{if } x > 0 \\ [-1, 1] & \text{if } x = 0 \\ -1 & \text{if } x < 0 \end{cases} \tag{24}$$

2.1 Proximal Gradient Descent

Definition 2.3 (Proximal Operator)

Given a lower semicontinuous convex function f mapping from Hilbert space X to $[-\infty, +\infty]$, its **proximal operator** associated with a point u is defined

$$\text{prox}_{f,\tau}(u) = \underset{x}{\text{argmin}} \left(f(x) + \frac{1}{2\tau} ||x - u||^2 \right) \tag{25}$$

where $\tau > 0$ is a parameter that scales the quadratic term. This is basically the point that minimizes the sum of $f(x)$ and the square of the Euclidean distance between x and u, scaled by $1/2\tau$.

Now given the loss function $L(\boldsymbol{\theta}) = L_{\text{obj}}(\boldsymbol{\theta}) + L_{\text{reg}}(\boldsymbol{\theta})$, we want to compute the proximal operator on the regularization loss and update that with the gradient of the smooth objective loss.

$$\boldsymbol{\theta}^{(k+1)} = \text{prox}_{L_{\text{reg}},\tau} \left[\boldsymbol{\theta}^{(k)} - \tau \nabla L_{\text{obj}}(\boldsymbol{\theta}^{(k)}) \right] \tag{26}$$

Let's compute the proximal operator of the L1 loss $h(\boldsymbol{\theta}) = \lambda ||\boldsymbol{\theta}||_1$. We can parameterize this loss by the λ, so we will use the notation $\text{prox}_{\lambda,\tau}$ rather than $\text{prox}_{h,\tau}$.

$$\begin{aligned} \text{prox}_{\lambda,\tau}(\mathbf{u}) &= \underset{\boldsymbol{\theta}}{\text{argmin}} \left(\lambda ||\boldsymbol{\theta}||_1 + \frac{1}{2\tau} ||\boldsymbol{\theta} - \mathbf{u}||_2^2 \right) \\ &= \underset{\boldsymbol{\theta}}{\text{argmin}} \left(\sum_{i=1}^n \lambda |\theta_i| + \frac{1}{2\tau} (\theta_i - u_i)^2 \right) \end{aligned}$$

These are separable functions that can be decoupled and optimized component-wise. So, we really just want to find

$$\theta_i^* = \underset{\theta_i}{\text{argmin}} \left(\lambda |\theta_i| + \frac{1}{2\tau} (\theta_i - u_i)^2 \right) \tag{27}$$

The sum of convex functions is convex, and so we should differentiate it and find where the gradient is 0 to optimize it.

1. When $\theta_i > 0$, then we minimize $\lambda \theta_i + \frac{1}{2\tau}(\theta_i - u_i)^2$, so taking the gradient and setting to 0 gives

$$\theta_i = u_i - \lambda \tau \tag{28}$$

subject to the constraint that $\theta_i > 0$, or equivalently, that $u_i > \lambda \tau$.

2. When $\theta_i < 0$, then we minimize $-\lambda \theta_i + \frac{1}{2\tau}(\theta_i - u_i)^2$, so taking the gradient and setting to 0 gives

$$\theta_i = u_i + \lambda \tau \tag{29}$$

subject to the constraint that $\theta_i < 0$, or equivalently, that $u_i < -\lambda \tau$.

3. When $\theta_i = 0$, then we minimize $\lambda|\theta_i| + \frac{1}{2\tau}(\theta_i - u_i)^2$, which doesn't have derivative at $\theta_i = 0$. So, we can compute the subdifferential of it to get

$$0 \in \partial\bigg(\lambda|\theta_i| + \frac{1}{2\tau}(\theta_i - u_i)^2\bigg) = \lambda\partial(|\theta_i|) + \frac{1}{\tau}(\theta_i - u_i)$$

   Now at $\theta_i = 0$, the subdifferential can be any value in $[-1, 1]$, and the above reduces to

$$0 \in \lambda[-1, 1] - \frac{1}{\tau}u_i \tag{30}$$

   this is equivalent to saying that u_i/τ is contained in the interval $[-\lambda, \lambda]$, meaning that $u_i \in [-\lambda\tau, \lambda\tau]$.

Ultimately we get that

$$\text{prox}_{\lambda,\tau}(u) = \begin{cases} u - \lambda\tau & \text{if } u > \lambda\tau \\ 0 & \text{if } |u| \leq \lambda\tau \\ u + \lambda\tau & \text{if } u < -\lambda\tau \end{cases} \tag{31}$$

which can be simplified to

$$\text{prox}_{\lambda,\tau}(u) = \text{sign}(u)\max\{|u| - \lambda\tau, 0) \tag{32}$$

3 Adaptive Gradient Methods

3.1 Adagrad

3.2 RMSProp and Adadelta

3.3 Adam

Adam and AdamW

4 Second-Order Optimizers

4.1 Newton's Method

Newton's method is an iterative algorithm for finding the roots of a differentiable function F. An immediate consequence is that given a convex C^2 function f, we can apply Newton's method to its derivative f' to get the critical points of f (minima, maxima, or saddle points), which is relevant in optimizing f. Given a C^1 function $f : D \subset \mathbb{R}^n \longrightarrow \mathbb{R}$ and a point $\mathbf{x}_k \in D$, we can compute its linear approximation as

$$f(\mathbf{x}_k + \mathbf{h}) \approx f(\mathbf{x}_k) + Df_{\mathbf{x}_k}\, \mathbf{h} = f(\mathbf{x}_k) + \nabla f(\mathbf{x}_k) \cdot \mathbf{h} \tag{33}$$

where $Df_{\mathbf{x}_k}$ is the total derivative of f at $\mathbf{x}_k$ and $\mathbf{h}$ is a small n-vector. Discretizing this gives us our gradient descent algorithm as

$$\mathbf{x}_{k+1} \leftarrow \mathbf{x}_k - \alpha\, f'(\mathbf{x}_k) \tag{34}$$

This linear function is unbounded, so we must tune the step size α accordingly. If α is too small, then convergence is slow, and if α is too big, we may overshoot the minimum. Netwon's method automatically tunes this α using the curvature information, i.e. the second derivative. If we take a second degree Taylor approximation

$$f(\mathbf{x}_k + \mathbf{h}) \approx f(\mathbf{x}_k) + Df_{\mathbf{x}_k}\, \mathbf{h} + \mathbf{h}^T\, Hf_{\mathbf{x}_k}\, \mathbf{h} \tag{35}$$

then we are guaranteed that this quadratic approximation of f has a minimum (existence and uniqueness can be proved), and we can calculate it to find our "approximate" minimum of f. We simply take the total derivative of this polynomial w.r.t. $\mathbf{h}$ and set it equal to the n-dimensional covector $\mathbf{0}$. This is equivalent to setting the gradient as $\mathbf{0}$, so

$$\begin{aligned}
\mathbf{0} &= \nabla_{\mathbf{h}} \big[f(\mathbf{x}_k) + Df_{\mathbf{x}_k}\, \mathbf{h} + \mathbf{h}^T\, Hf_{\mathbf{x}_k}\, \mathbf{h} \big](\mathbf{h}) \\
&= \nabla_{\mathbf{h}} [Df_{x_k} \mathbf{h}](\mathbf{h}) + \nabla_{\mathbf{h}} [\mathbf{h}^T\, Hf_{\mathbf{x}_k}\, \mathbf{h}](\mathbf{h}) \\
&= \nabla_{\mathbf{x}} f(\mathbf{x}_k) + Hf_{\mathbf{x}_k}\, \mathbf{h} \\
&\implies \mathbf{h} = -[Hf_{\mathbf{x}_k}]^{-1} \nabla_{\mathbf{x}} f(\mathbf{x}_k)
\end{aligned}$$

which results in the iterative update

$$\mathbf{x}_{k+1} \leftarrow \mathbf{x}_k - [Hf_{\mathbf{x}_k}]^{-1} \nabla_{\mathbf{x}} f(\mathbf{x}_k) \tag{36}$$

Note that we require $\mathbf{f}$ to be convex, so that Hf is positive definite. Since f is C^2, this implies Hf is also symmetric, implying invertibility by the spectral theorem. Note that Newton's method is very expensive, since we require the computation of the gradient, the Hessian, *and* the inverse of the Hessian, making the computational complexity of this algorithm to be $O(n^3)$. We can also add a smaller stepsize h to control stability.

Algorithm 2 Newton's Method

```
Require: Initial $\mathbf{x}_0$, Stepsize $h$ (optional)
  for $t = 0$ to $T$ until convergence do
    $g(\mathbf{x}_t) \leftarrow \nabla f(\mathbf{x}_t)$
    $H(\mathbf{x}_t) \leftarrow Hf_{\mathbf{x}_t}$
    $H^{-1}(\mathbf{x}_t) \leftarrow [H(\mathbf{x}_t)]^{-1}$
    $\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t - h\, H^{-1}(\mathbf{x}_t)\, g(\mathbf{x}_t)$
  end for
```

4.2 Gauss Newton Method

5 Quasi-Newton Methods

5.1 Secant Method

5.2 DFP

5.3 Broyden

5.4 Symmetric Rank 1

5.5 BFGS

Netwon's method is extremely effective for finding the minimum of a convex function, but there are two disadvantages. First, it is sensitive to initial conditions, and second, it is extremely expensive, with a computational complexity of $O(n^3)$ from having to invert the Hessian. An alternative family of optimizers, called *quasi-Newton* methods, try to *approximate* the Hessian (or Jacobian) with $\hat{H}f$, reducing the computational cost without too much loss in convergence rates, and simply use this approximation in the Newton's update:

$$\mathbf{x}_{k+1} \leftarrow \mathbf{x}_k - [\hat{H}f_{\mathbf{x}_k}]^{-1} \nabla_{\mathbf{x}} f(\mathbf{x}_k)$$

The method of the Hessian approximation varies by algorithm, but the most popular is BFGS.

So how do we approximate the Hessian with only the gradient information? With secants. Starting off with $f : \mathbb{R} \longrightarrow \mathbb{R}$, let us assume that we have two points $(x_k, f(x_k))$ and $(x_{k+1}, f(x_{k+1}))$. We can approximate our derivative (gradient in dimension 1) at x_{k+1} using finite differences:

$$f'(x_{k+1})(x_{k+1} - x_k) \approx f(x_{k+1}) - f(x_k)$$

and doing the same for f' gives us the second derivative approximation:

$$f''(x_{k+1})(x_{k+1} - x_k) \approx f'(x_{k+1}) - f'(x_k)$$

which gives us the update:

$$x_{k+1} \leftarrow x_k - \frac{x_k - x_{k-1}}{f'(x_k) - f'(x_{k-1})} f'(x_k)$$

This method of approximating Netwon's method in one dimension by replacing the second derivative with its finite difference approximation is called the *secant method.* In multiple dimensions, given two points $\mathbf{x}_k, \mathbf{x}_{k+1}$ with their respective gradients $\nabla f(\mathbf{x}_k), \nabla f(\mathbf{x}_{k+1})$, we can approximate the Hessian $\hat{H}f_{\mathbf{x}_{k+1}} \approx D(\nabla f)_{\mathbf{x}_{k+1}}$ (which is the total derivative of the gradient) at $\mathbf{x}_{k+1}$ with the equation

$$\hat{H}f_{\mathbf{x}_{k+1}}(\mathbf{x}_{k+1} - \mathbf{x}_k) = \nabla_{\mathbf{x}} f(\mathbf{x}_{k+1}) - \nabla_{\mathbf{x}} f(\mathbf{x}_k)$$

This is solving the equation of form $A\mathbf{x} = \mathbf{y}$ for some linear map A. Since $\hat{H}f_{\mathbf{x}_{k+1}}$ is a symmetric $n \times n$ matrix with $n(n+1)/2$ components, there are $n(n+1)/2$ unknowns with only n equations, making this an underdetermined system. Quasi-Newton methods have to impose additional constraints, with the BFGS choosing the one where we want $\hat{H}f_{\mathbf{x}_{k+1}}$ to be as close as to $\hat{H}f_{\mathbf{x}_k}$ at each update $k+1$. Luckily, we can formalize this notion as minimizing the distance between $f_{\mathbf{x}_{k+1}}$ and $\hat{H}f_{\mathbf{x}_k}$. Therefore, we wish to find

$$\arg \min_{\hat{H}f_{\mathbf{x}_{k+1}}} ||\hat{H}f_{\mathbf{x}_{k+1}} - \hat{H}f_{\mathbf{x}_k}||_F,$$

where $|| \cdot ||_F$ is the Frobenius matrix norm, subject to the restrictions that $\hat{H}f_{\mathbf{x}_{k+1}}$ be positive definite and symmetric and that $\hat{H}f_{\mathbf{x}_{k+1}}(\mathbf{x}_{k+1} - \mathbf{x}_k) = \nabla_{\mathbf{x}} f(\mathbf{x}_{k+1}) - \nabla_{\mathbf{x}} f(\mathbf{x}_k)$ is satisfied. Since we have to invert it eventually, we can actually just create an optimization problem that directly computes the inverse. So, we wish to find

$$\arg \min_{(\hat{H}f_{\mathbf{x}_{k+1}})^{-1}} ||(\hat{H}f_{\mathbf{x}_{k+1}})^{-1} - (\hat{H}f_{\mathbf{x}_k})^{-1}||_F$$

subject to the restrictions that

1. $(\hat{H}f_{\mathbf{x}_{k+1}})^{-1}$ be positive definite and symmetric. It turns out that the positive definiteness restriction also restricts it to be symmetric.
2. $\mathbf{x}_{k+1} - \mathbf{x}_k = (\hat{H}f_{\mathbf{x}_{k+1}})^{-1}[\nabla_{\mathbf{x}} f(\mathbf{x}_{k+1}) - \nabla_{\mathbf{x}} f(\mathbf{x}_k)]$

After some complicated mathematical derivation, which we will not go over here, the problem ends up being equivalent to updating our approximate Hessian at each iteration by adding two symmetric, rank-one matrices U and V scaled by some constant, which can each be computed as an outer product of vectors with itself.

$$\hat{H}f_{\mathbf{x}_{k+1}} = \hat{H}f_{\mathbf{x}_k} + aU + bV = \hat{H}f_{\mathbf{x}_k} + a\mathbf{u}\mathbf{u}^T + b\mathbf{v}\mathbf{v}^T$$

where $\mathbf{u}$ and $\mathbf{v}$ are linearly independent. This addition of a rank-2 sum of matrices $aU + bV$, known as a rank-2 update, guarantees the "closeness" of $\hat{H}f_{\mathbf{x}_{k+1}}$ to $\hat{H}f_{\mathbf{x}_k}$ at each iteration. With this form, we now impose the quasi-Newton condition. Substituting $\Delta\mathbf{x}_k = \mathbf{x}_{k+1} - \mathbf{x}_k$ and $\mathbf{y}_k = \nabla_{\mathbf{x}} f(\mathbf{x}_{k+1}) - \nabla_{\mathbf{x}} f(\mathbf{x}_k)$, we have

$$\hat{H}f_{\mathbf{x}_{k+1}}\Delta\mathbf{x}_k = \hat{H}f_{\mathbf{x}_{k+1}}\Delta\mathbf{x}_k + a\mathbf{u}\mathbf{u}^T\Delta\mathbf{x}_k + b\mathbf{v}\mathbf{v}^T\Delta\mathbf{x}_k = \mathbf{y}_k$$

A natural choice of vectors turn out to be $\mathbf{u} = \mathbf{y}_k$ and $\mathbf{v} = \hat{H}f_{\mathbf{x}_k}\Delta\mathbf{x}_k$, and substituting this and solving gives us the optimal values

$$a = \frac{1}{\mathbf{y}_k^T\Delta\mathbf{x}_k}, \qquad b = -\frac{1}{\Delta\mathbf{x}_k^T\hat{H}f_{\mathbf{x}_k}\Delta\mathbf{x}_k}$$

and substituting these values back to the Hessian approximation update gives us the BFGS update:

$$\hat{H}f_{\mathbf{x}_{k+1}} = \hat{H}f_{\mathbf{x}_k} + \frac{\mathbf{y}_k\mathbf{y}_k^T}{\mathbf{y}_k^T\Delta\mathbf{x}_k} - \frac{\hat{H}f_{\mathbf{x}_k}\Delta\mathbf{x}_k\Delta\mathbf{x}_k^T\hat{H}f_{\mathbf{x}_k}}{\Delta\mathbf{x}_k^T\hat{H}f_{\mathbf{x}_k}\Delta\mathbf{x}_k}$$

We still need to invert this, and using the *Woodbury formula*

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}$$

which tells us how to invert the sum of an intertible matrix A and a rank-k correction, we can derive the iterative update of the inverse Hessian as

$$(\hat{H}f_{\mathbf{x}_{k+1}})^{-1} = \left(I - \frac{\Delta\mathbf{x}_k\mathbf{y}^T}{\mathbf{y}_k^T\Delta\mathbf{x}_k}\right)(\hat{H}f_{\mathbf{x}_k})^{-1}\left(I - \frac{\mathbf{y}_k\Delta\mathbf{x}_k^T}{\mathbf{y}_k^T\Delta\mathbf{x}_k}\right) + \frac{\Delta\mathbf{x}_k\Delta\mathbf{x}_k^T}{\mathbf{y}_k^T\Delta\mathbf{x}_k}$$

Remember that this is the iterative step that we want to actually compute, rather than the ones computing the regular Hessian. The whole point of using the Woodbury formula to derive an inverse update step was to do away with the tedious $O(n^3)$ computations of inverting an $n \times n$ matrix. This rank-2 update also preserves positive-definiteness.

Finally, we can choose the initial inverse Hessian approximation $(\hat{H}f_{\mathbf{x}_{k+1}})^{-1}$ to be the identity I or the true inverse Hessian $(Hf_{\mathbf{x}_{k+1}})^{-1}$ (computed just once), which would lead to more efficient convergence. The pseudocode for BFGS is a bit too long and confusing to include here, but most of the time, we won't be implementing BFGS by hand; efficient and established BFGS optimizers are already in numerous packages. Like most optimizers, BFGS is not guaranteed to converge to the true global minimum.

5.6 Limited Memory BFGS (L-BFGS)

6 Gradient Free Methods

6.1 Simulated Annealing

Unlike the previous optimizers, *simulated annealing* is useful in finding *global* optima in the presence of multimodal functions within a usually very large discrete space $\mathcal{S}$. Given some function f defined on $\mathcal{S}$, we would like to find its global maximum. Rather than picking the "best move" using gradient information (like SGD), we propose a random move. Let us start at a state θ_k and propose a random P_{k+1}. We denote $\Delta f = f(P_{k+1}) - f(\theta_k)$.

1. If the selected move improves the solution (i.e. $\Delta f \geq 0$, then it is always accepted and we set $\theta_{k+1} \leftarrow P_{k+1}$.
2. Otherwise, when $\Delta f < 0$ it makes the move with the following acceptance probability

$$p(\theta_{k+1} \leftarrow P_{k+1} \mid \Delta f < 0) = e^{\Delta f / T(t)}$$

We can see that if Δf is very negative (the move is very bad), then this probability of acceptance decreases as well. Furthermore, $T(t)$ represents some sort of "temperature" that we anneal as a function of time, called the *annealing schedule*. T starts off at a high value, increasing the rate at which bad moves are accepted, which promotes exploration of $\mathcal{S}$ and allows the algorithm to travel to suboptimal areas. As T decreases, the vast majority of steps move uphill, promoting exploitation, which means that once the algorithm is in the right search space, there is no need to search other sections of the search space.

Algorithm 3 Simulated Annealing

```
Require: Initial $\theta_0$, Transition kernel $\pi(\theta_{k+1} \mid \theta_k)$, Annealing schedule $T(t)$
  for $t = 0$ to $T$ until convergence do
    $P_{t+1} \sim \pi(\cdot \mid \theta_t)$
    if $f(P_{t+1}) - f(\theta_t) \geq 0$ then
      $\theta_{t+1} \leftarrow P_{t+1}$
    else
      $\delta \sim \text{Uniform}[0, 1]$
      if $\delta < \exp[(f(P_{t+1}) - f(\theta_t))/T(t)]$ then
        $\theta_{t+1} \leftarrow P_{t+1}$
      else
        $\theta_{t+1} \leftarrow \theta_t$
      end if
    end if
  end for
```

This algorithm is very easy to implement and provides optimal solutions to a wide range of problems (e.g. TSP and nonlinear optimization), but it can take a long time to run if the annealing schedule is very long. We can stop either if T reaches a certain threshold or if we have determined convergence.

6.2 Nelder-Mead

Uses simplex.

7 Lagrangian Optimizers for Constraints

7.1 Lagrange Multipliers

For equality.

7.2 KKT Conditions

For inequality.

7.3 ADMM

8 Non-Lagrangian Optimizers for Constraints

8.1 Penalty

8.2 Projection

8.3 Saddle Point Problem in Nonconvex Optimization

[PDGB14]

9 Sparsity-Inducing Optimizers

9.1 Clipping

We can do SGD with clipping.

References

[PDGB14] Razvan Pascanu, Yann N. Dauphin, Surya Ganguli, and Yoshua Bengio. On the saddle point problem for non-convex optimization, 2014.