Descent Methods

Karl Stratos

1 Example

We can learn a great deal about descent methods by carefully studying the simple example in Boyd and Vandenberghe (2004). Here, we wish to minimize a quadratic function $f : \mathbb{R}^2 \to \mathbb{R}$ defined by

$$f(x) := \frac{1}{2}\left(x_1^2 + \gamma x_2^2\right)$$

where we assume $\gamma \geq 1$ for simplicty. It's a special instantiation of the general quadratic program $(1/2)x^\top Qx - b^\top x$ where $Q = \text{diag}(1, \gamma)$ and $b = (0, 0)$. The gradient is $\nabla f(x) = Qx$ and the Hessian is $\nabla^2 f(x) = Q \succ 0$. Thus $f(x)$ is strongly convex and minimized as $f(x^*) = 0$ at the unique minimum $x^* := (0, 0)$. This allows us to evaluate the suboptimality of x by simply checking how large $f(x)$ is.

The (always symmetric) Hessian $\nabla^2 f(x)$ reveals the shape of the function at point x. It has positive eigenvalues when it's positive definite (like here), and its condition number

$$\kappa\left(\nabla^2 f(x)\right) := \frac{\lambda_{\max}\left(\nabla^2 f(x)\right)}{\lambda_{\min}\left(\nabla^2 f(x)\right)} \geq 1$$

upper bounds the "assymmetry" of $f(x)$'s upward acceleration. The Hessian in this example is constant Q with $\kappa(Q) = \gamma$.

Gradient descent Let's consider minimizing $f(x)$ using gradient descent with exact step sizes. Clearly, the choice of an initial point $x^{(0)}$ affects the number of iterations t we need until $x^{(t)}$ is close to optimal (i.e., zero). If $x^{(0)} = (0, 0)$ there's no work to be done. If $x^{(0)} = (1, 0)$, then $\nabla f\left(x^{(0)}\right) = (1, 0)$ so we obtain $x^{(1)} = x^{(0)} - \eta \nabla f\left(x^{(0)}\right) = x^*$ in a single iteration with step size $\eta = 1$. Now consider $x^{(0)} = (\gamma, 1)$ with the initial objective value $f\left(x^{(0)}\right) = \gamma(\gamma + 1)/2$. Then it can be shown that the exact step size is

$$\eta_t^{\text{exact}} := \arg\min_{\eta \geq 0} f\left(x^{(t)} - \eta \nabla f\left(x^{(t)}\right)\right) = \frac{2}{1+\gamma} \qquad \forall t \geq 1$$

and each update and its objective value are

$$x^{(t)} = \left(\frac{\gamma - 1}{\gamma + 1}\right)^t (-1)^t x^{(0)} \qquad f\left(x^{(t)}\right) = \left(\frac{\gamma - 1}{\gamma + 1}\right)^{2t} f\left(x^{(0)}\right) \tag{1}$$

What we're observing is the dependence of the convergence rate of gradient descent on the condition number γ of the Hessian Q. If $\gamma = 1$, the algorithm converges in a single iteration. If $\gamma = 10^6$, then the algorithm crawls down to a halt, progressing by a factor of $(10^6 - 1)/(10^6 + 1) \approx 1$ in each iteration.

Newton's method If a small condition number is so important for fast convergence, can achieve it by preprocessing? Let's apply a change of coordinates $y := Q^{1/2}x$. Then $x = Q^{-1/2}y$ and

$$f(x) = f(Q^{-1/2}y) = \frac{1}{2}\left(y^\top Q^{-1/2}\right) Q \left(Q^{-1/2}y\right) = \frac{1}{2}y^\top y =: \bar{f}(y)$$

and we can minimize $\bar{f}(y)$ over y using gradient descent. It's the same setting in (1) only now we're optimizing a function whose Hessian has a condition number $\kappa\left(\nabla^2 \bar{f}(y)\right) = 1$, so the algorithm immediately converges to $y^{(1)} = 0$ no matter what $x^{(0)}$ is. The original solution can be recovered by $x^{(1)} = Q^{-1/2}y^{(1)} = 0$.

What we're observing is a property of the Newton update $x' = x - \nabla^2 f(x)^{-1} \nabla f(x)$, namely that it's equivalent to gradient descent after the change of coordinates $y := \nabla^2 f(x)^{1/2} x$. See Appendix B for details. The Newton step $v = -\nabla^2 f(x)^{-1} \nabla f(x)$ also minimizes a second-order approximation of $f(x+v)$ around x,

$$f(x+v) \approx f(x) + v^\top \nabla f(x) + \frac{1}{2} v^\top \nabla^2 f(x) v$$

which is exact in this example (hence the immediate solution). In contrast, the gradient step $v = -\eta \nabla f(x)$ minimizes a regularized first-order approximation.

Granted, this example is a simple quadratic function so the behavior of gradient descent and Newton's method is not completely representative. However, it provides right intuition about these descent methods, for instance

- Gradient descent is sensitive to the choice of coordinates; Newton isn't.
- Newton converges really fast if the function looks like quadratic.

2 Descent Algorithm

More generally, let $f : \mathbb{R}^d \to \mathbb{R}$ be a smooth function with a finite minimizer $x^* := \inf_{x \in \mathbb{R}^d} f(x)$. The sublevel set is denoted by $S := \{x \in \mathbb{R}^d : f(x) \le f(x^{(0)})\}$ and the optimum is denoted by $p^* := f(x^*)$. We assume that f is **strongly convex**; then it follows that

- There exist $m, M > 0$ such that $mI \preceq \nabla^2 f(x) \preceq MI$ for all $x \in S$.
- If the gradient is small at x, $f(x)$ is almost optimal (if zero, $f(x) = p^*$):

$$f(x) - p^* \le \frac{1}{2m} ||\nabla f(x)||_2^2 \tag{2}$$

All descent methods use the following algorithm.

DESCEND
Input: f, number of iterations T, descent directions $\{\Delta x\}_{x \in \mathbb{R}^d}$, step sizes $\{\eta_t\}_{t=0}^{T-1}$

- Initialize $x^{(0)}$ somehow.
- For $t = 0, 1, \ldots, T-1$, set $x^{(t+1)} \leftarrow x^{(t)} + \eta_t \Delta x^{(t)}$.
- Return $x^{(T)}$.

We now discuss some suitable choices of

- **Descent directions** Δx defined for all $x \in \mathbb{R}^d$
- **Step sizes** η_t defined for all $t = 0 \dots T-1$

2.1 Descent Directions

The rate of change of f at x along a vector $v \in \mathbb{R}^d$ can be measured by the **directional derivative**:

$$\nabla_v f(x) := \lim_{\epsilon \to 0} \frac{f(x+\epsilon v) - f(x)}{\epsilon} = \langle v,\ \nabla f(x) \rangle$$

where the second equality can be verified (Appendix A). Thus we obtain the **steepest descent direction of f at x** by

$$\Delta_{||\cdot||} x := \underset{v:\ ||v|| \leq 1}{\arg\min} \langle v,\ \nabla f(x) \rangle \tag{3}$$

where we constrain the length of v with some norm $||\cdot||$. We obtain various directions depending on the choice of the norm. For a symmetric $Q \succ 0$, we define the **Q-norm** by $||v||_Q := \sqrt{v^\top Q v}$. Note that $||v||_Q = \left|\left|Q^{1/2} v\right|\right|_2$.

Proposition 2.1. *Assuming $\nabla f(x) \neq 0$, we have*

$$\Delta_{||\cdot||_2} x = -\left|\left|\nabla f(x)\right|\right|_2^{-1} \nabla f(x) \tag{4}$$

$$\Delta_{||\cdot||_1} x = -\text{sign}\left(\frac{\partial f(x)}{\partial x_l}\right) e_l \qquad l = \underset{i \in \{1,\dots,d\}}{\arg\max} \left|\frac{\partial f(x)}{\partial x_i}\right| \tag{5}$$

$$\Delta_{||\cdot||_Q} x = -\left|\left|\nabla f(x)\right|\right|_{Q^{-1}}^{-1} Q^{-1} \nabla f(x) \tag{6}$$

(Partial proof in Appendix C.) Since we usually just care about the direction, we consider the following "unnormalized" definitions obtained by multiplying $\left|\left|\nabla f(x)\right|\right|_*$ where $\left|\left|\cdot\right|\right|_*$ is the dual norm.[1]

- The **2-norm direction** of f at x is given by

  $$\Delta^{(2)} x := -\nabla f(x)$$

  Gradient descent is **DESCEND** using the 2-norm direction.

- The **1-norm direction** of f at x is given by

  $$\Delta^{(1)} x := -\frac{\partial f(x)}{\partial x_l} e_l$$

  where $l = \arg\max_{i \in \{1,\dots,d\}} \left|\frac{\partial f(x)}{\partial x_i}\right|$. Coordinate descent is **DESCEND** using the 1-norm direction.

[1] Equivalently, we can modify the definition of the steepest direction in Eq. (3) with the dual norm constraint to directly obtain unnormalized directions (e.g., see Appendix C):

$$\min_{v:\ ||v|| \leq ||v||_*} \langle v,\ \nabla f(x) \rangle$$

- The Q-**norm direction** of f at x is given by

$$\Delta^{(Q)}x := -Q^{-1}\nabla f(x)$$

Newton's method is **DESCEND** using the Q-norm direction where $Q = \nabla^2 f(x)$.

If $\bar{x} = Q^{1/2}x$ denotes a new coordinate system, then it can be shown that

$$\Delta^{(Q)}x = Q^{-1/2}\Delta^{(2)}\bar{x}$$

Thus pursuing the Q-norm direction can be seen as pursuing the 2-norm direction in the new coordinate system $x \mapsto Q^{1/2}x$ (Corollary B.2).

2.2 Step Sizes

Here are some choices of the step size η_t at each iteration $t \in \{0, \ldots, T-1\}$. A **fixed step size** η_t^{fixed} is simply a constant $\eta > 0$:

$$\eta_t^{\text{fixed}} = \eta \tag{7}$$

The **exact step size** η_t^{exact} is the optimum step size for $\Delta x^{(t)}$:

$$\eta_t^{\text{exact}} = \underset{\eta \in \mathbb{R}}{\arg\min}\ f\left(x^{(t)} + \eta \Delta x^{(t)}\right) \tag{8}$$

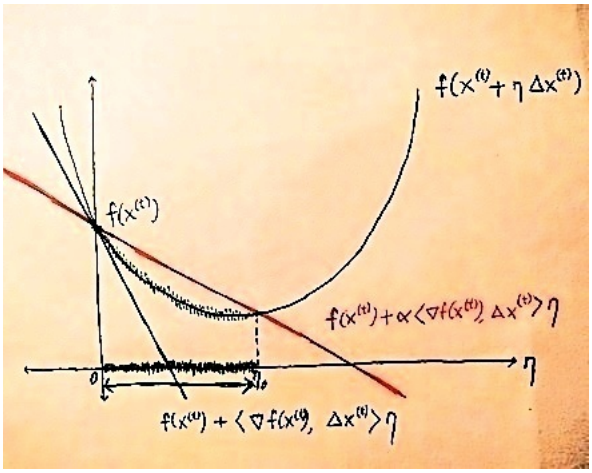
A popular choice in practice is the **backtracking step size** η_t^{back}:

$$\eta_t^{\text{back}} \in \left\{\eta :\ f\left(x^{(t)} + \eta\Delta x^{(t)}\right) \leq f(x^{(t)}) + \alpha\left\langle \nabla f(x^{(t)}),\ \Delta x^{(t)}\right\rangle \eta\right\} \tag{9}$$

where $\alpha \in (0, 0.5)$. To derive this, note that the first-order Taylor polynomial of the objective $f(x^{(t)} + \eta\Delta x^{(t)})$ at $\eta = 0$ (when viewed as a function of η) is $f(x^{(t)}) + \langle \nabla f(x^{(t)}),\ \Delta x^{(t)}\rangle \eta$. Since $\Delta x^{(t)}$ is a descent direction, we have $\langle \nabla f(x^{(t)}),\ \Delta x^{(t)}\rangle < 0$, so this is a line with negative slope tagent to $f(x^{(t)} + \eta\Delta x^{(t)})$ at $\eta = 0$. The line with a reduced slope $f(x^{(t)}) + \alpha\langle \nabla f(x^{(t)}),\ \Delta x^{(t)}\rangle \eta$ will intersect the objective first at some $\eta_0 > 0$. This means

$$f(x + \eta\Delta x) \leq f(x) + \alpha\langle \nabla f(x),\ \Delta x\rangle \eta \qquad\qquad \forall \eta \in [0, \eta_0]$$

The following is a usual illustration of the backtracking step size:



A step size satisfying (9) can be computed by the following algorithm:

Backtrack
Input: f, current location $x^{(t)}$, descent direction $\Delta x^{(t)}$, $\alpha \in (0, 0.5)$, $\beta \in (0, 1)$

- $\eta_t^{\text{back}} \leftarrow 1$
- Until $f(x^{(t)} + \eta_t^{\text{back}} \Delta x^{(t)}) \leq f(x^{(t)}) + \alpha \langle \nabla f(x^{(t)}),\ \Delta x^{(t)} \rangle \eta_t^{\text{back}}$,

$$\eta_t^{\text{back}} \leftarrow \beta \eta_t^{\text{back}}$$

- Return η_t^{back}.

The algorithm eventually terminates since as $\eta \to 0$,

$$f(x + \eta \Delta x) \approx f(x) + \langle \nabla f(x),\ \Delta x \rangle \eta \leq f(x) + \alpha \langle \nabla f(x),\ \Delta x \rangle \eta$$

3 Gradient Descent

Let's look at gradient descent in more details, which specifies the update

$$x' \leftarrow x - \eta \nabla f(x)$$

As we saw, the gradient step $v = -\eta \nabla f(x)$ is given by the 2-norm direction $\Delta^{(2)} x = -\nabla f(x)$ multiplied by a step size η. Alternatively, v minimizes a regularized first-order approximation of $f(x + v)$ around x,

$$f(x + v) \approx f(x) + v^\top \nabla f(x) + \frac{1}{2\eta} ||v||_2^2$$

3.1 Exact Step Size

Let's first consider using the exact step size η_t^{exact} in (8). The line search forces the algorithm to "go all the way" along the direction until it's impossible to reduce f any further, so it yields the minimizer x^* in a single iteration if the 2-norm direction points precisely to x^*. An interesting consequence is that the direction at time $t + 1$ is orthogonal to the direction at time t.

Lemma 3.1. *Given $x \in \mathbb{R}^d$, define $\eta^* := \arg\min_{\eta \in \mathbb{R}} f(x - \eta \nabla f(x))$ and $x^+ := x - \eta^* \nabla f(x)$. Then $\langle \nabla f(x^+),\ \nabla f(x) \rangle = 0$.*

Proof. Since η^* is a stationary point of $g(\eta) := f(x - \eta \nabla f(x))$,

$$\frac{\partial g(\eta^*)}{\partial \eta} = \langle \nabla f(x - \eta^* \nabla f(x)),\ -\nabla f(x) \rangle = -\langle \nabla f(x^+),\ \nabla f(x) \rangle = 0$$

□

Now we have a result generalizing the phonomenon in our earlier example (1): convergence depends on the condition number of the Hessian. Note that M/m is an upper bound on the condition number.

Theorem 3.2. *If $x^{(t)}$ is the output of **DESCEND** with exact step sizes, then*

$$f(x^{(t)}) - p^* \leq \left(1 - \frac{m}{M}\right)^t \left(f\left(x^{(0)}\right) - p^*\right)$$

3.2 Fixed Step Size

Let's also consider using the fixed step size η_t^{fixed} in (7). Here is a classical result on the convergence of gradient descent in this case; see Appendix D for a proof. The analysis assumes the gradient is Lipschitz; ∇f is **L-Lipschitz** if ∇f doesn't "change too quickly":

$$||\nabla f(x) - \nabla f(y)||_2 \leq L\,||x - y||_2 \qquad \forall x, y \in \mathbb{R}^d$$

It can be verified that this is equivalent to $\nabla^2 f(x) \preceq LI$ for all $x \in \mathbb{R}^d$.

Theorem 3.3. *Assume that ∇f is L-Lipschitz. Pick $\eta \in (0, 1/L]$. The output $x^{(T)}$ of gradient descent with step size η satisfies*

$$f(x^{(T)}) - f(x^*) \leq \frac{\left|\left|x^{(0)} - x^*\right|\right|_2^2}{2\eta T}$$

3.3 Stochastic Gradient Descent (SGD)

Proposition 3.1. *Let* **pop** *denote a distribution over $\mathcal{X}$. For $x \in \mathcal{X}$, let $J_x : \mathbb{R}^d \to \mathbb{R}$ be a twice differentiable and strongly convex function. Define $J : \mathbb{R}^d \to \mathbb{R}$ by*

$$J(\theta) := \mathop{\boldsymbol{E}}_{x \sim \mathbf{pop}} [J_x(\theta)]$$

*and let $\theta^* = \arg\min_{\theta \in \mathbb{R}^d} J(\theta)$. Let $\Theta \subset \mathbb{R}^d$ denote a sufficiently large closed subset of $\mathbb{R}^d$ and assume $\theta^* \in \Theta$. Pick any $\theta_0 \in \Theta$ and define a sequence $\theta_1, \theta_2, \ldots$ by drawing $x_t \sim$ **pop** and setting*

$$\theta_{t+1} = \theta_t - \eta_t \nabla J_{x_t}(\theta_t) \tag{10}$$

where $\eta_t \geq 0$. Then $\lim_{t\to\infty} \theta_t = \theta^*$ with probability 1, provided that

$$\sum_{t=0}^{\infty} \eta_t = \infty \qquad\qquad \sum_{t=0}^{\infty} \eta_t^2 < \infty \tag{11}$$

Proof. The update (10) is element-wise, so without loss of generality assume $d = 1$. $J' : \mathbb{R} \to \mathbb{R}$ is nondecreasing and $J''(\theta^*) > 0$ by premise. $J'_x(\theta)$ is an unbiased estimator of $J'(\theta)$, that is $J'(\theta) = \mathbf{E}_{x \sim \mathbf{pop}}[J'_x(\theta)]$. Furthermore, $|J'_x(\theta)| \leq G$ for all $\theta \in \Theta$ for some constant G, so it is uniformly bounded. Since θ^* is the unique solution to the equation $J'(\theta) = 0$, by Robbins-Monro θ_t converges to θ^* almost surely with the conditions in (11). □

The conditions in (11) are necessary as follows. By a telescoping sum on (10) we have

$$\theta_0 - \theta_T = \sum_{t=0}^{T} \eta_t \nabla J_{x_t}(\theta_t)$$

Thus if $\sum_{t=0}^{\infty} \eta_t$ is finite, then the difference between θ_0 and $\lim_{t\to\infty} \theta_t$ is finite. In this case there is no hope of approaching θ^* if θ^* happens to be too far away from θ_0. An extreme case is $\eta_t = 0$ which clearly breaks the algorithm, but even with a strictly positive η_t this can be a problem. For instance, if $\eta_t = 1/2^{t+1}$, then $\sum_{t=0}^{\infty} \eta_t \nabla J_{x_t}(\theta_t)$

is the expected value of $\nabla J_{x_t}(\theta_t)$ over $x_1, x_2, \ldots \sim \mathbf{pop}$ weighted by probabilities $1/2, 1/4, \ldots$ which is finite since ∇J_{x_t} is bounded.

On the other hand, for the algorithm to converge at all we must have

$$\begin{aligned}(\theta_{t+1} - \theta^*)^2 - (\theta_t - \theta^*)^2 &= \eta_t^2 \left|\left|\nabla J_{x_t}(\theta_t)\right|\right|^2 - 2\eta_t \nabla J_{x_t}(\theta_t)^\top (\theta_t - \theta^*) \\ &\leq \eta_t^2 \left|\left|\nabla J_{x_t}(\theta_t)\right|\right|^2\end{aligned}$$

converge as $t \to \infty$. (The first equality can be easily checked by using (10), and $\nabla J_{x_t}(\theta_t)^\top (\theta_t - \theta^*) \geq 0$ for any finite t since J_{x_t} is convex.) The second condition $\sum_{t=0}^{\infty} \eta_t^2 < \infty$, together with the boundedness of ∇J_{x_t}, ensure that $\sum_{t=0}^{\infty} \eta_t^2 \left|\left|\nabla J_{x_t}(\theta_t)\right|\right|^2$ is finite, and this implies the RHS converges in *conditional* expectation. We need to condition on the past otherwise we allow for degenerate scenarios such as using the same $x \in \mathcal{X}$ at every step. This involves the Martingale convergence theorem and details can be found in Bertsekas (2011) and Bottou (1998).

It is also shown that in the ideal setting of Proposition 3.1 in which J is smooth and strongly convex, the convergence rate of SGD is as good as gradient descent (e.g., Theorem 3.3) in that the expected suboptimality is inversely proportional to the number of iterations:

$$\mathop{\mathbf{E}}_{x_1 \ldots x_{T-1} \sim \mathbf{pop}} [J(\theta_T) - J(\theta^*)] = O\left(\frac{1}{T}\right)$$

SGD also has a quite different interpretation under the framework of online convex optimization. Here, we aim to minimize the "regret" with respect to the best hypothesis in an online setting. It can be shown that SGD optimizes the sum of linearized past losses with l_2 regularization which upper bounds this regret.

3.3.1 General convergence analysis

Let f be smooth function bounded below by f^*. Fix x_1 arbitrarily and consider

$$x_{t+1} = x_t - \eta g_t$$

where g_t is a stochastic estimator of $\nabla f(x_t)$ and $\eta > 0$. By Taylor's theorem,

$$f(x_{t+1}) = f(x_t) - \eta \left\langle g_t, \nabla f(x_t) \right\rangle + \frac{\eta^2 \left|\left|g_t\right|\right|^2}{2} \left|\left|\nabla^2 f(\xi)\right|\right|$$

where $\xi = (1-t)x_t + t x_{t+1}$ for some $t \in (0,1)$. Assume $||g_t|| \leq M$ and $\left|\left|\nabla^2 f(x)\right|\right| \leq L$. Crucially, suppose

$$\mathbf{E}\left[\langle g_t, \nabla f(x_t)\rangle\right] \geq C \left|\left|\nabla f(x_t)\right|\right|^p \tag{12}$$

for some $C > 0$ and $p \geq 1$. Then we can take an expectation over g_t to have

$$\mathbf{E}\left[f(x_{t+1})\right] \leq f(x_t) - \eta C \left|\left|\nabla f(x_t)\right|\right|^p + \frac{\eta^2 M^2 L}{2}$$

Rearranging and averaging both sides over $t = 1 \ldots T$, we have

$$\frac{1}{T} \sum_{t=1}^{T} \left|\left|\nabla f(x_t)\right|\right|^p \leq \frac{f(x_1) - f^*}{\eta C T} + \frac{\eta M L}{2C}$$

In particular, given any $\epsilon > 0$, we can choose $\eta = \frac{2C\epsilon}{ML}$ to have the asymptotic convergence

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} ||\nabla f(x_t)||^p \leq \epsilon$$

In classical SGD, we have $f(x) = (1/B) \sum_{i=1}^{B} f_i(x)$ (e.g., x is the model parameter and f_i is the loss function defined on the i-th batch) and use $g_t = \nabla f_{i_t}(x) \approx \nabla f(x)$ where $i_t \in \{1 \dots B\}$ is uniformly random. Then (12) holds with $C = 1$ and $p = 2$.

More generally, we can consider any stochastic gradient estimator g_t such that

$$\mathbf{E}\left[\cos\theta_t\right] \geq \delta \tag{13}$$

where θ_t is the angle between g_t and $\nabla f(x_t)$.[2] If we additionally assume $||g_t|| \geq m$,

$$\mathbf{E}\left[\langle g_t, \nabla f(x_t)\rangle\right] = \mathbf{E}\left[||g_t|| \cos\theta_t\right] ||\nabla f(x_t)|| \geq m\delta \, ||\nabla f(x_t)||$$

thus (12) holds with $C = m\delta$ and $p = 1$.

4 Newton's Method

As we saw, Newton's method is given by choosing the Q-norm direction $\Delta^{(Q)}x = -Q^{-1}\nabla f(x)$ where

$$Q = \nabla^2 f(x)$$

in the **DESCEND** algorithm. The algorithm can also be interepreted as minimizing the second-order Taylor polynomial around x:

$$f(x+v) \approx f(x) + v^\top \nabla f(x) + \frac{1}{2} v^\top \nabla^2 f(x) v$$

See Boyd and Vandenberghe (2004) for more details. Some main takeaways for Newton's Methods are:

- Its convergence speed is unaffected by any change of coordinates $x \mapsto Tx$ (Appendix B).
- It's an exact algorithm for quadratic functions. It's a great algorithm for quadratic-like functions, which is formalized by the L-Lipschitzness of the Hessian

  $$\left|\left|\nabla^2 f(x) - \nabla^2 f(y)\right|\right|_2 \leq L \, ||x - y||_2 \qquad \forall x, y \in S$$

  Note that L can be taken zero for quadratic functions. It can be shown that there is some $\tau \in (0, m^2/L]$ such that once $||\nabla f(x)||_2 < \tau$ then convergence of Newton's method is extremely rapid.

[2] This assumption may not hold in SGD due to the nonlinearity of cosine. Let $f_i(x) = -x^2$ for $i = 1 \dots 9$ and $f_{10}(x) = 19x^2$. Then $f(x) = x^2$. The gradients are $\nabla f_i(x) = -2x$ for $i = 1 \dots 9$, $\nabla f_{10}(x) = 38x$, and $\nabla f(x) = 2x$. At $x = 1$, $\nabla f_i(1) = -2$ for $i = 1 \dots 9$, $\nabla f_{10}(1) = 38$, and $\nabla f(1) = 2$. The expected gradient is consistent: $(-18 + 38)/10 = 2$. However, letting θ_i denote the angle between $\nabla f_i(1)$ and $\nabla f(1)$,

$$\mathop{\mathbf{E}}_{i \sim \text{Unif}(\{1 \dots 10\})} [\cos\theta_i] = \frac{1}{10}(-9 + 1) = -0.8$$

- In practice, it's impractical to store and invert the Hessian. A successful approach known as BFGS/L-BFGS approximates the multiplication of $\nabla f(x)$ by $\nabla^2 f(x)^{-1}$ without explicitly computing $\nabla^2 f(x)^{-1}$.

A Directional Derivative and Gradient

Lemma A.1. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function. For any $v \in \mathbb{R}^d$,*

$$\nabla_v f(x) = \langle v,\ \nabla f(x) \rangle$$

Proof. Define $g_v : \mathbb{R} \to \mathbb{R}$ to be $f(x + \epsilon v)$ viewed as a function of ϵ:

$$g_v(\epsilon) := f(x + \epsilon v)$$

Express $g_v' : \mathbb{R} \to \mathbb{R}$ in both the limit and the gradient form:

$$g_v'(\epsilon) = \lim_{\rho \to 0} \frac{f(x + (\epsilon + \rho)v) - f(x + \epsilon v)}{\rho}$$
$$g_v'(\epsilon) = \frac{\partial f(x + \epsilon v)}{\partial (x + \epsilon v)} \frac{\partial (x + \epsilon v)}{\partial \epsilon} = \langle \nabla f(x + \epsilon v),\ v \rangle$$

Evaluating g_v' at $\epsilon = 0$ with these forms yields

$$g_v'(0) = \nabla_v f(x) = \langle \nabla f(x),\ v \rangle$$

□

B Affine Invariance

We write $\Delta_x^{f(\tau(x))}$ denote a step for variable x when the target function f is viewed a function of some transformation $\tau(x)$ of x.

Definition B.1. *A descent method is **affine invariant** if for all $A \succ 0$,*

$$\Delta_x^{f(x)} = A \Delta_{A^{-1}x}^{f(x)}$$

Meaning $x \mapsto A^{-1}x$ is a (reversible) change of coordinates where $y = A^{-1}x \in \text{range}(A^{-1})$ is the new variable to optimize. For instance, given $f(x) := x^\top \text{diag}(1, \gamma)x$ where γ is large, we may consider a change of coordinates by $A = \text{diag}(1, \gamma)^{-1/2}$. This yields an optimization of a simple function $g(y) := f(Ay) = y^\top y$ (from which we can recover $x = Ay$). If a descent method is affine invariant, it means that there is no benefit of doing this trick in terms of improving convergence rate. This is because the updates are exactly the same (up to transformation by A). See Lemma B.1.

Lemma B.1. *Let $x_1, x_2, \ldots$ denote the sequence of updates optimizing $f(x)$ from an initial point x_0. Let $y_1, y_2, \ldots$ denote the sequence of updates optimizing $g(y) = f(Ay)$ from $y_0 = A^{-1}x_0$. If the descent method is affine invariant, $x_t = Ay_t$ for all $t \geq 0$.*

Proof. Each step on $g(y)$ is

$$\Delta_y^{g(y)} = \Delta_y^{f(Ay)} = \Delta_{A^{-1}z}^{f(z)} = A^{-1} \Delta_z^{f(z)}$$

The second equality is just renaming variable $z = Ay$. The last equality is affine invariance. $x_0 = Ay_0$ by premise. Assume $x_t = Ay_t$. We have

$$y_{t+1} = y_t - \eta \Delta_{y=y_t}^{g(y)} = y_t - \eta A^{-1} \Delta_{z=Ay_t}^{f(z)} = A^{-1}x_t - \eta A^{-1} \Delta_{z=x_t}^{f(z)}$$

Thus $Ay_{t+1} = x_t - \eta \Delta_{z=x_t}^{f(z)} = x_{t+1}$. □

Proposition B.1. *Newton's method is affine invariant.*

Proof. Define $g(y) := f(Ay)$. Then

$$\begin{aligned} \nabla g(y) &= A^\top \nabla f(Ay) \\ \nabla^2 g(y) &= A^\top \nabla^2 f(Ay) A \end{aligned} \tag{14}$$

Thus the Newton step on $g(y)$ is

$$\begin{aligned} \Delta_y^{g(y)} &= (A^\top \nabla^2 f(Ay) A)^{-1} A^\top \nabla f(Ay) \\ &= A^{-1} \nabla^2 f(Ay)^{-1} \nabla f(Ay) \\ &= A^{-1} \nabla^2 f(x)^{-1} \nabla f(x) \end{aligned}$$

where we simply rename $x = Ay$. Then

$$A\Delta_{A^{-1}x}^{f(x)} = A\Delta_y^{f(Ay)} = A\Delta_y^{g(y)} = \nabla^2 f(x)^{-1} \nabla f(x) = \Delta_x^{f(x)}$$

□

Proposition B.2. *Gradient descent is not affine invariant in general.*

Proof. Define $g(y) := f(Ay)$. The gradient step on $g(y)$ is Eq. (14),

$$\Delta_y^{g(y)} = A^\top \nabla f(Ay) = A^\top \nabla f(x)$$

again renaming $x = Ay$. Then

$$A\Delta_{A^{-1}x}^{f(x)} = A\Delta_y^{g(y)} = AA^\top \nabla f(x) \tag{15}$$

is not equal to $\Delta_x^{f(x)} = \nabla f(x)$ unless $\nabla f(x) \in \text{range}(A)$. □

Thus the convergence rate of gradient descent can change depending on the choice of coordinates, whereas Newton's method does not.[3] While gradient descent is not affine invariant itself, it has a precise connection to Newton's method as the following statement shows.

Corollary B.2. *Let Δ denote a Newton step and $\overline{\Delta}$ denote a gradient step. Then*

$$\Delta_x^{f(x)} = A\overline{\Delta}_{A^{-1}x}^{f(x)}$$

for the choice of $A = \nabla^2 f(x)^{-1/2}$.

Proof. By Eq. (15), the RHS is $AA^\top \nabla f(x) = \nabla^2 f(x)^{-1} \nabla f(x) = \Delta_x^{f(x)}$. □

Therefore, the convergence rate of gradient descent with the change of coordinate $x \mapsto \nabla^2 f(x)^{1/2} x$ is exactly the same as the convergence rate of Newton's method.

[3] Newton's method is still not invariant to general coordinate transformations—only affine.

C Proof of Proposition 2.1

We only prove for the Q-norm constraint. The constrained optimization problem is

$$\min_{v:\ ||v||_Q \le \epsilon} \langle v,\ \nabla f(x)\rangle$$

where $\epsilon > 0$, $\nabla f(x) \neq 0$, and $||v||_Q := \sqrt{v^\top Q v} \geq 0$ is the Q-norm defined with some $Q \succ 0$. If v is any vector with $J(v) = \langle v,\ \nabla f(x)\rangle < 0$ (which exists since $\nabla f(x) \neq 0$), then we can achieve smaller objective by increasing its Q-norm: pick any $C > 1$ and set $v' = Cv$, then $||v'||_Q = C\,||v||_Q$ and $J(v') = CJ(v) < J(v)$. Thus the solution is clearly achieved at v with $||v||_Q = \epsilon$, so without loss of generality we can consider the following Lagrangian relaxation

$$L(v, \lambda) = \langle v,\ \nabla f(x)\rangle + \frac{\lambda}{2}\left(v^\top Q v - \epsilon^2\right)$$

and solve $\max_\lambda\ \min_v\ L(v, \lambda)$ to calculate v that minimizes $\langle v,\ \nabla f(x)\rangle$ while satisfying $||v||_Q = \epsilon$. As usual the solution is found at a saddle point (v, λ) satisfying

$$\begin{aligned}
\frac{\partial}{\partial \lambda} L(v, \lambda) = 0 \qquad &\Leftrightarrow \qquad v^\top Q v = \epsilon^2 \\
\frac{\partial}{\partial v} L(v, \lambda) = 0 \qquad &\Leftrightarrow \qquad Qv = -\frac{1}{\lambda}\nabla f(x)
\end{aligned}$$

This is a system of two equations with two variables thus solvable. One silly trap here is that we might be tempted to multiply the second RHS with v to get $\lambda = -(1/\epsilon^2) v^\top \nabla f(x)$, but this doesn't eliminate v so we run in circles. Instead, we have to write it as $Q^{1/2} v = -\frac{1}{\lambda} Q^{-1/2} \nabla f(x)$ and use the fact that both sides have same squared norm, yielding

$$v^\top Q v = \frac{1}{\lambda^2} \nabla f(x)^\top Q^{-1} \nabla f(x) \qquad \Leftrightarrow \qquad \lambda = \frac{||\nabla f(x)||_{Q^{-1}}}{\epsilon}$$

Thus $v = -(\epsilon / ||\nabla f(x)||_{Q^{-1}}^{-1}) Q^{-1} \nabla f(x)$.

D Proof of Theorem 3.3

We repeat the theorem:

Theorem Assume that ∇f is L-Lipschitz. Pick $\eta \in (0, 1/L]$. The output $x^{(T)}$ of gradient descent with step size η satisfies

$$f(x^{(T)}) - f(x^*) \leq \frac{\left|\left|x^{(0)} - x^*\right|\right|_2^2}{2\eta T}$$

We prove this theorem in small pieces. First, we show that each gradient update will decrease the value of f if the gradient is nonzero. This crucially depends on the quadratic upper bound provided by the L-Lipschitzness assumption on ∇f, but not on the convexity of f.

Lemma D.1. *Assume that ∇f is L-Lipschitz. If $\eta \leq 1/L$,*

$$f(x^{(t+1)}) \leq f(x^{(t)}) - \frac{\eta}{2}\left|\left|\nabla f(x^{(t)})\right|\right|_2^2 \qquad \forall t \in \{0 \dots T-1\} \tag{16}$$

Proof.

$$\begin{aligned} f(x^{(t+1)}) &= f\left(x^{(t)} - \eta \nabla f(x^{(t)})\right) \\ &\leq f(x^{(t)}) - \eta \left|\left|\nabla f(x^{(t)})\right|\right|_2^2 + \frac{L\eta^2}{2} \left|\left|\nabla f(x^{(t)})\right|\right|_2^2 \\ &\leq f(x^{(t)}) - \eta \left(1 - \frac{L\eta}{2}\right) \left|\left|\nabla f(x^{(t)})\right|\right|_2^2 \end{aligned}$$

If we choose $\eta \leq 1/L$, we have the desired result. □

Next, we show that the suboptimality of $x^{(t)}$ is bounded linearly by the gradient at $x^{(t)}$. The convexity assumption on f provides exactly such a bound.

Lemma D.2. *Assume that f is convex. Then*

$$f(x^{(t)}) - f(x^*) \leq \langle \nabla f(x^{(t)}),\ x^{(t)} - x^* \rangle \qquad \forall t \in \{0 \dots T\} \tag{17}$$

Proof. This follows by rearranging the definition of convexity at $x^{(t)}$:

$$f(x^*) \geq f(x^{(t)}) + \langle \nabla f(x^{(t)}),\ x^* - x^{(t)} \rangle$$

□

The suboptimality bound on $x^{(t)}$ in Eq. (17) depends on the gradient. Somewhat fortuitously, if we bound the suboptimality of its *update* $x^{(t+1)}$ using Lemma D.1 and D.2, we end up removing the dependence on the gradient. The resulting bound is purely in terms of the distance away from x^*.

Lemma D.3. *Assume that f is convex and ∇f is L-Lipschitz. If $\eta \in (0, 1/L]$,*

$$f(x^{(t+1)}) - f(x^*) \leq \frac{1}{2\eta} \left(\left|\left|x^{(t)} - x^*\right|\right|_2^2 - \left|\left|x^{(t+1)} - x^*\right|\right|_2^2 \right) \quad \forall t \in \{0 \dots T-1\} \tag{18}$$

Proof. Successively applying Eq. (16) and (17) gives:

$$f(x^{(t+1)}) \leq f(x^*) + \langle \nabla f(x^{(t)}),\ x^{(t)} - x^* \rangle - \frac{\eta}{2} \left|\left|\nabla f(x^{(t)})\right|\right|_2^2$$

Using the following observation,

$$\frac{1}{2\eta} \left|\left|x^{(t+1)} - x^*\right|\right|_2^2 = \frac{1}{2\eta} \left|\left|x^{(t)} - x^*\right|\right|_2^2 - \langle \nabla f(x^{(t)}),\ x^{(t)} - x^* \rangle + \frac{\eta}{2} \left|\left|\nabla f(x^{(t)})\right|\right|_2^2$$

we can further express

$$\begin{aligned} f(x^{(t+1)}) &\leq f(x^*) + \left(\frac{1}{2\eta} \left|\left|x^{(t)} - x^*\right|\right|_2^2 + \frac{\eta}{2} \left|\left|\nabla f(x^{(t)})\right|\right|_2^2 - \frac{1}{2\eta} \left|\left|x^{(t+1)} - x^*\right|\right|_2^2 \right) \\ &\quad - \frac{\eta}{2} \left|\left|\nabla f(x^{(t)})\right|\right|_2^2 \\ &= f(x^*) + \frac{1}{2\eta} \left(\left|\left|x^{(t)} - x^*\right|\right|_2^2 - \left|\left|x^{(t+1)} - x^*\right|\right|_2^2 \right) \end{aligned}$$

Note that the gradient terms cancel. □

Given these lemmas, the proof of the theorem is easy. We bound the suboptimality of $x^{(T)}$ by the average suboptimality and apply Lemma D.3. The terms telescope and we get the desired result.

Proof of Theorem 3.3.

$$
\begin{aligned}
f(x^{(T)}) - f(x^*) &\leq \frac{1}{T}\sum_{t=0}^{T-1} f(x^{(t+1)}) - f(x^*) && \text{[by Lemma D.1]} \\
&\leq \frac{1}{2\eta T}\sum_{t=0}^{T-1} \left|\left|x^{(t)} - x^*\right|\right|_2^2 - \left|\left|x^{(t+1)} - x^*\right|\right|_2^2 && \text{[by Eq. (18)]} \\
&\leq \frac{\left|\left|x^{(0)} - x^*\right|\right|_2^2 - \left|\left|x^{(T)} - x^*\right|\right|_2^2}{2\eta T} \\
&\leq \frac{\left|\left|x^{(0)} - x^*\right|\right|_2^2}{2\eta T}
\end{aligned}
$$

□

References

Bertsekas, D. P. (2011). Incremental gradient, subgradient, and proximal methods for convex optimization: A survey. *Optimization for Machine Learning*, **2010**(1-38), 3.

Bottou, L. (1998). Online learning and stochastic approximations. *On-line learning in neural networks*, **17**(9), 142.