University of California, Berkeley Lecturer: Benjamin Recht

CS227C/STAT260 Convex Optimization and Approximation: Optimization for Modern Data Analysis

January 22, 2015 Notes: Ashia Wilson and Benjamin Recht

Lecture 2: the gradient method

In this lecture, we take a tour of the gradient method, providing several different perspectives on this fundamental algorithm. The gradient method follows the simple algorithmic procedure:

1. Choose $x_0 \in \mathbb{R}^d$ and set $k = 0$
2. Choose $t_k > 0$ and set $x_{k+1} = x_k - t_k \nabla f(x_k)$ and $k = k + 1$,
3. Repeat 2 until converged.

This simple iterative procedure forms the basis for every algorithm we will study between now and the midterm. So we're going to do a deep dive on its properties for the next lecture or two.

1 Descent directions and optimality conditions

Let's suppose we want to minimize a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$. Most of the algorithms we will consider start at some point x_0 and then aim to find a new point x_1 with a lower function value. The simplest way to do so is to find a direction v such that f is decreasing when moving along the direction v. This notion can be formalized by the following definition:

Definition 1 *v is a descent direction for f at x_0 if $f(x_0 + tv) < f(x_0)$ for some $t > 0$.*

A simple characterization of descent directions is given by the following proposition.

Proposition 1 *For a continuous differentiable function f on a neighborhood of x_0, if $v^T \nabla f(x) < 0$ then v is a descent direction.*

Proof By continuity, there exists a T such that $\nabla f(x_0 + tv)^T v < 0$ for all $t \in [0, T]$. By Taylor's theorem, $f(x_0 + tv) = f(x_0) + t\nabla f(x_0 + \tilde{t}v)^T v$ for some $\tilde{t} \in [0, t]$.Therefore $f(x_0 + tv) < f(x_0)$ and v is a descent direction. ∎

Note that among all directions with unit norm,

$$\inf_{||v||=1} v^T \nabla f(x) = -||\nabla f||$$

which is achieved when $v = -\frac{\nabla f(x)}{||\nabla f(x)||}$. This means that $-\nabla f(x)$ is the direction of *steepest descent*. This characterization allows us to provide conditions as to when x minimizes f.

Definition 2 *1. x_* is a* global minimizer *of f if $f(x_*) \leq f(x)$ $\forall x \in \mathbb{R}^d$.*

2. x_* is a local minimizer *of f if there is a neighborhood $\mathcal{N}$ around x such that $f(x_*) \leq f(x)$ for all $x \in \mathcal{N}$.*

The first conditions concern local optimality.

Proposition 2 (Optimality Conditions) *1. $x_\star$ is a local minimizer only if $\nabla f(x_\star) = 0$*

2. *If $\nabla^2 f$ is continuous and $x_\star$ is a local minimzer, then $\nabla^2 f(x_\star) \succeq 0$*

3. *If f is twice continuously differentiable, $\nabla f(x_\star) = 0$, $\nabla^2 f(x_\star) \succ 0$ then $x_\star$ is a local minimizer.*

Proof

1. Since $-\nabla f(x_\star)$ is always a descent direction, the gradient must vanish.

2. If $x_\star$ is a local minimizer, $f(x_\star + td) \geq f(x_\star)$ for all d and some t sufficiently small. Using part 1 and Taylor's theorem,

$$f(x_\star + td) = f(x_\star) + \frac{1}{2} t^2 d^T \nabla^2 f(x_\star + \hat{t} d) d$$

   for some $\hat{t} \in [0, t]$. Therefore $d^T \nabla^2 f(x_\star) d \geq 0$ for all d.

3. There exists an $r > 0$ such that $\nabla^2 f(x) \succ 0$ for all $x \in B(x_\star, r)$. Pick d with $||d|| < r$. Then we have

$$\begin{aligned} f(x_\star + d) &= f(x_\star) + d^T \nabla f(x_\star) + \frac{1}{2} d^T \nabla^2 f(x + td) d \qquad (\text{ for some } t \in [0,1]) \\ &\geq f(x_\star) \end{aligned}$$

   proving $x_\star$ is a local minimizer.

■

For convex f, the situation is dramatically simpler. This is part of the reason why convexity is so appealing.

Proposition 3 *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable convex function. Then $x_\star$ is a global minimizer of f if and only if $\nabla f(x_\star) = 0$.*

Proof What is particularly remarkable about the proof of this proposition is that it is almost tautological: if f is differentiable, then f is convex if and only if

$$f(x) \geq f(x_\star) + \nabla f(x_\star)^T (x - x_\star)$$

for all x. Using this equivalence, if $\nabla f(x_\star) = 0$, then $f(x) \geq f(x_\star)$ for all x. Conversely, if $f(x) \geq f(x_\star)$ for all x, we also have by the first-order convexity condition that

$$f(x_\star + t \nabla f(x_\star)) \geq f(x_\star) + t \|\nabla f(x_\star)\|^2$$

for all $t > 0$. Subtracting $f(x_\star)$ from both sides shows that $\|\nabla f(x_\star)\|^2 = 0$, and thus $\nabla f(x_\star) = 0$. ■

2 Fixed point iteration

Our first view of the gradient method is as a fixed point iteration. In order to solve for our optimal $x_\star$ it suffices to solve the (typically nonlinear) equation $\nabla f(x_\star) = 0$. A popular method for solving such an equation is by a fixed point iteration. Come up with some mapping $\Phi : \mathbb{R}^d \to \mathbb{R}^d$ such that $x_\star = \Phi(x_\star)$ if and only if $\nabla f(x_\star) = 0$.

A simple candidate is $\Phi(x) = x - \alpha \nabla f(x)$. Let's assume that

1. There exits an $x_\star \in \mathcal{D}$ with $\nabla f(x_\star) = 0$.
2. $\Phi(x) = x - \alpha \nabla f(x)$ is *contractive* on $\mathcal{D}$ for some $\alpha > 0$: i.e., there is a $\beta \in [0, 1)$ such that

$$||\Phi(x) - \Phi(z)|| \leq \beta ||x - z|| \quad \forall x, z \in \mathcal{D}$$

.

Then if we run the gradient method starting at $x_0 \in \mathcal{D}$,

$$\begin{aligned} ||x_{k+1} - x_\star|| &= ||x_k - \alpha \nabla f(x_k) - x_\star|| \\ &= ||\psi(x_k) - \psi(x_\star)|| \\ &\leq \beta ||x_k - x_\star|| \\ &\vdots \\ &\leq \beta^{k+1} ||x_0 - x_\star|| \,. \end{aligned}$$

This derivation reveals that x_k converges *linearly* to $x_\star$. That is, at every iteration, the distance to the optimal solution is decreased by a constant factor.

As an aside, we say that this is linear convergence because

$$||x_{k+1} - x_\star|| \leq \beta ||x_k - x_\star||$$

So the iterates are related by a linear recursion. If we had instead

$$||x_{k+1} - x_\star|| \leq \beta ||x_k - x_\star||^2$$

we'd say that the convergence was quadratic.

OK, back to the gradient method. How many iterations must we run to guarantee that $\|x_k - x_\star\| \leq \epsilon$? A simple calculation reveals that

$$\boxed{k \geq -\frac{\log\left(\frac{||x_0 - x_\star||}{\epsilon}\right)}{\log(\beta)}}$$

suffice.

Quick check:

$$\begin{aligned}
\beta^k ||x_0 - x_\star|| &\leq \epsilon \\
k \log \beta &\leq -\log\left(\frac{||x_0 - x_\star||}{\epsilon}\right) \\
-k \log \beta &\geq \log\left(\frac{||x_0 - x_\star||}{\epsilon}\right) \\
-k &\leq \frac{\log\left(\frac{||x_0 - x_\star||}{\epsilon}\right)}{\log \beta} \qquad\qquad (\log \beta < 0) \\
k &\geq -\frac{\log\left(\frac{||x_0 - x_\star||}{\epsilon}\right)}{\log \beta}
\end{aligned}$$

We are left with a few questions. First, when can we guarantee that Φ is a contractive map? This can be summarized by the following proposition:

Proposition 4 *If f is twice continuously differentiable and Φ is contractive then f must be convex.*

Proof First, by the definition of contractivity, we have for all $t > 0$ that

$$\frac{1}{t}||\Phi(x + t\Delta x) - \Phi(x)|| \leq \beta ||\Delta x||$$

provided $x + t\Delta x \in \mathcal{D}$. Taking the limit as t goes to zero yields

$$\begin{aligned}
\beta ||\Delta x|| &\geq \lim_{t \to 0} \frac{1}{t} ||\psi(x + t\Delta x) - \psi(x)|| \\
&= \lim_{t \to 0} ||\Delta x - \frac{\alpha}{t}(\nabla f(x + t\Delta x) - \nabla f(x))|| \\
&= || \left[I - \alpha \nabla^2 f(x)\right] \Delta x|| \,.
\end{aligned}$$

This inequality means

$$||I - \alpha \nabla^2 f(x)|| \leq \beta \,,$$

and hence we all of the eigenvalues of $I - \alpha \nabla^2 f(x)$ are between $-\beta$ and β. We can write this in terms of the semidefinite ordering as

$$\frac{1-\beta}{\alpha} I \preceq \nabla^2 f(x) \preceq \frac{1+\beta}{\alpha} I \,.$$

The first term in the partial order implies that f is convex. It actually implies that f is *strongly covnex*, a concept we will revisit shortly. The latter partial ordering states that the curvature of f must be globally bounded. In this next section we will discuss this curvature condition and its consequences. We will indeed show that, for differentiable functions, the gradient method converges at a linear rate if and only if f is strongly convex and has bounded curvature. ■

3 Lipschitz Continuity

Let's dive a bit more into the curvature condition that popped out of our analysis of the fixed point iteration. Unconstrained optimization algorithms should be scale invariant. For any $a > 0$ and $b \in \mathbb{R}$, $af(x) + b$ has the same optimal solution as $f(x)$. Our algorithms should respect this symmetry. One convenient way to set a scale is to define the Lipschitz constants associated with f and its gradients. We will see throughout the course that our precision should scale proportional to these constants.

Definition 3 (Lipschitz Continuity) *A mapping $\phi : \mathbb{R}^d \to \mathbb{R}^n$ is Lipschitz continuous on Ω if $\exists L \geq 0$ such that $\forall (x, y) \in \Omega$*

$$||\phi(x) - \phi(y)|| \leq L||x - y||$$

Note that if $\phi(x)$ is L-Lipschitz continuous and $a > 0$ then $a\phi(x)$ is aL-Lipschitz continuous.

For real-valued functions, the Lipschitz constant gives us a scale of how quickly f can vary from point to point. In particular, if the gradient is bounded, then so is the Lipschitz constant.

Proposition 5 *If $||\nabla f||$ is bounded on Ω, $M = \sup_z ||\nabla f(z)||$, then*

$$|f(x) - f(y)| \leq M||x - y||$$

Proof By Taylor's theorem,

$$f(x) - f(y) = \int_0^1 \nabla f(tx + (1-t)y)^T (x - y) dt\,.$$

Therefore,

$$\begin{aligned}
|f(x) - f(y)| &= \left| \int_0^1 \nabla f(tx + (1-t)y)^T (x - y) dt \right| \\
&\leq \int_0^1 \left| \nabla f(tx + (1-t)y)^T (x - y) dt \right| \\
&\leq \int_0^1 \left|\left| \nabla f(tx + (1-t)y)^T \right|\right| \, ||(x - y)|| \, dt \\
&= \left(\int_0^1 ||\nabla f(tx + (1-t)y)|| dt \right) ||x - y|| \\
&\leq M \, ||x - y||
\end{aligned}$$

Here, the first inequality is the triangle inequality. The second inequality is Cauchy-Schwartz, and the final inequality uses our bound on the gradient. ■

In a very similar fashion, the Lipschitz constant of the gradient controls how quickly the curvature of the function f can change. It is upper bounded by the operator norm of the Hessian:

Proposition 6 *If $||\nabla^2 f||$ is bounded in the operator norm on Ω, $L = \sup_x ||\nabla^2 f(x)||$, then*

$$|\nabla f(x) - \nabla f(y)| \leq L||x - y||$$

Proof The proof is more or less identical to the proof of Proposition 5. By Taylor's theorem.

$$\nabla f(x) - \nabla f(y) = \int_0^1 \nabla^2 f(tx + (1-t)y)^T (x-y) dt$$

And then we have the same chain of inequalities:

$$\begin{aligned}
|\nabla f(x) - \nabla f(y)| &= \left| \int_0^1 \nabla^2 f(tx + (1-t)y)^T (x-y) dt \right| \\
&\leq \int_0^1 |\nabla^2 f(tx + (1-t)y)^T (x-y)| dt \\
&\leq \left(\int_0^1 ||\nabla^2 f(tx + (1-t)y)|| dt \right) ||x-y|| \\
&\leq L||x-y|| \, .
\end{aligned}$$

As above, the first inequality follows from the triangle inequality and the second from Cauchy-Schwarz. ■

We can derive an even stronger coupling between the Hessian and the Lipschitz constant of the gradient via the following chain of equivalences. Again, almost everything is a simple consequence of Taylor's theorem.

Proposition 7 *Suppose f is twice differentiable on Ω. Then the following are equivalent.*

1. *∇f is Lipschitz with constant L on Ω*
2. *$\forall x, y \in \Omega, \ f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}||x-y||^2$*
3. *$\forall x, y \in \Omega, \ \langle \nabla f(x) - \nabla f(y), x - y \rangle \leq L||x-y||^2$*
4. *$\langle y - x, \nabla^2 f(x)(y-x) \rangle \leq L||x-y||^2$*

Proof ((1) $\Rightarrow$ (2)): Apply Taylor's theorem and Cauchy-Schwartz,

$$\begin{aligned}
f(y) - f(x) - \nabla f(x)^T (y-x) &= \int_0^1 \left(\nabla f(ty + (1-t)x) - \nabla f(x) \right)^T (y-x) dt \\
&\leq ||y-x|| \int_0^1 ||\nabla f(ty + (1-t)x) - \nabla f(x)|| dt \\
&\leq ||y-x|| \int_0^1 Lt||y-x|| dt \\
&= L||y-x||^2 \int_0^1 t dt \\
&= \frac{L}{2}||y-x||^2
\end{aligned}$$

Here, the third inequality follows by applying Proposition 6. Rearranging terms proves the assertion.

((2) $\Rightarrow$ (3)): Note that switching the roles of x and y, we have two inequalities.

$$f(y) \leq f(x) + \nabla f(x)^T(y-x) + \frac{L}{2}||y-x||^2$$
$$f(x) \leq f(y) + \nabla f(y)^T(x-y) + \frac{L}{2}||x-y||^2$$

adding the inequalities proves this implication.
((3) $\Rightarrow$ (4)): Substituting $x = z + t(u-z)$ and $y = z$ gives

$$\langle \nabla f(x + t(y-x)) - \nabla f(x), t(y-x) \rangle \leq Lt^2||y-x||^2$$

Dividing by t^2 gives

$$\left\langle \frac{\nabla f(x + t(y-x)) - \nabla f(x)}{t}, y-x \right\rangle \leq L||y-x||^2$$

Finally, taking the limit as $t \to 0$ proves the assertion.
((4) $\Rightarrow$ (1)): Condition 4 is the same as saying that $||\nabla^2 f||$ is bounded in operator norm. We established from the previous proposition that bounds on the operator norm of the Hessian implies the gradients are Lipschitz. ■

4 Line search methods

We now turn to our second interpretation of gradient descent: that it is a line search method. The main idea here is to find a descent direction, and then minimize f—exactly or approximately—along that direction. That is, we will study the procedure

1. Pick v_k such that $\nabla f(x_k)^T v_k < 0$
2. Pick t_k to decrease f in the direction of v_k (1-D search).
3. Repeat 1 and 2 until converged.

There are a variety of ways to choose v_k. The most obvious choice is $-\nabla f(x_k)$. This is the gradient method. As we discussed in Section 1, $-\nabla f(x_k)$ is the direction of *steepest* descent. Though it seems odd that we would pick any direction other than the steepest descent direction, we will find many examples where being a bit less greedy can result in considerably faster convergence.

The line search part is a bit more nebulous. When we analyzed the gradient method as a fixed point iteration, we pulled the α parameter out of a hat. But there are far more systematic choices when we view the gradient method as a descent method. Some examples include

1. *Exact Line Search*: We choose $t_k = \arg\min_{t\geq 0}\{f(x_k + tv_k)\}$. This method relies on being able to solve differentiable, one-dimensional optimization problems quickly. While this is not generically easy, there are many cases where exact line search is straightforward. The most notable example is when f is a multivariate polynomial. In this case, $f(x_k + tv_k$ is a polynomial in t, and the minimum can be found by computing the derivative and root finding.

2. *Constant Step Size*: As we saw in section 2, constant step sizes can yield rapid convergence rates. The main drawback with these methods is one often needs some prior information about f to properly choose the stepsize.

3. *Diminishing stepsize*: Another canonical choice is to pick a stepsize sequence that tends to zero but whose sum diverges. For example $t_k = C/k$ for some C. We will return to this more when we analyze nonsmooth and stochastic optimization.

4. *Goldstein-Armijo Condition*: This condition has two parameters $0 < \alpha < \beta < 1$. We choose t such that

$$\begin{aligned} f(x_k + tv_k) &\leq f(x_k) + \alpha t \nabla f(x_k)^T v \\ f(x_k + tv_k) &\geq f(x_k) + \beta t \nabla f(x_k)^T v \end{aligned}$$

   The idea behind the Armijo condition is displayed in Figure 1. When $\alpha = 1$ and f is convex, the linear approximation lies below the curve. As we shrink α to zero, the approximation becomes more and more of an over approximation of f. The first condition guarantees that we lie *below* the over-approximation and hence are not moving too far. The second condition guarantees that we make some progress on this iteration. The Armijo-Goldstein condtions are the main principle behind back-tracking line search.

5. *Wolfe Conditions* The Wolfe conditions are primarily geared towards quasi-Newton methods, and we will spend more time on these when we revisit that topic. They are

$$\begin{aligned} f(x_k + tv_k) &\leq f(x_k) + \alpha t \nabla f(x_k)^T v \\ \nabla f(x_k + tv_k)^T v_k &\geq \gamma \nabla f(x_k)^T v \end{aligned}$$

   where $\gamma \in (\alpha, 1)$. The first condition is the sufficient decrease condition from the Armijo condtitions. The second states that the derivative of $\phi(t) = f(x_k + tv_k)$ after taking our step is sufficiently larger than at $t = 0$. This makes sense as if the slope is smaller, we should keep moving along the direction v!

Let's now provide a simple analysis of the gradient method as a line search method. Suppose f has L-Lipschitz gradients. Observe that for any t, the Lipschitz continuity of the gradient implies

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) - t||\nabla f(x_k)||^2 + \frac{Lt^2}{2}||\nabla f(x_k)||^2 \\ &= f(x_k) - t\left(1 - \frac{tL}{2}\right)||\nabla f(x_k)||^2 . \end{aligned}$$

Note that if $0 < t < \frac{2}{L}$, then $t\left(1 - \frac{tL}{2}\right) > 0$, and the function value is decreasing. In particular, if you choose $t = 1/L$, we have

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2L}||\nabla f(x_k)||^2 \tag{1}$$

Also note that if you perform exact line search, (1) holds. This is because

$$\min_{t>0} f(x_k - t\nabla f(x_k)) \leq f(x_k - 1/L\nabla f(x_k))$$

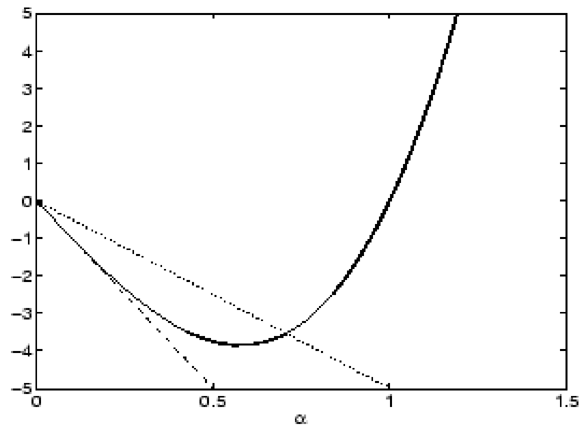

Figure 1: A graphical display of the Armijo condition. The dashed line is the tangent curve defined by the gradient ($\alpha = 1$). As we shrink α, the curve now lies partially above f. Note that if we like under the stronger dashed curve, then we are likely to not overshoot the minimum.

Define the quantity

$$\eta := \begin{cases} 2L & \text{for exact line search} \\ \frac{1}{t\left(1-\frac{TL}{2}\right)} & \text{for constant step size} \end{cases}$$

We can rearrange (1) and sum over the iterates of the algorithm to find that

$$\begin{aligned} \sum_{k=0}^{N} ||\nabla f(x_k)||^2 &\leq \eta \sum_{k=0}^{N} f(x_k) - f(x_{k+1}) \\ &= \eta[f(x_0) - f(x_N)] \\ &\leq \eta[f(x_0) - f(x_\star)]\,. \end{aligned}$$

The second line follows because the sum telescopes.

This implies that $\lim_{N\to\infty} ||\nabla f(x_N)|| = 0$. More concretely

$$\begin{aligned} \min_{0\leq k\leq N} ||\nabla f(x_k)|| &\leq \sqrt{\frac{\eta[f(x_0) - f(x_N)]}{N}} \\ &\leq \sqrt{\frac{\eta[f(x_0) - f(x_\star)]}{N}} \\ &\leq \sqrt{\frac{\eta\frac{L}{2}||x_0 - x_\star||^2}{N}}\,. \end{aligned}$$

For exact line search, this guarantees that we find a point x

$$\|\nabla f(x)\| \leq \frac{L||x_0 - x_\star||}{N^{1/2}}\,.$$

For constant step size we are guaranteed to find a point with

$$\|\nabla f(x)\| \leq \sqrt{\frac{1}{2\beta\left(1-\frac{\beta}{2}\right)}}\frac{L||x_0 - x_\star||}{N^{1/2}}$$

when our stepsize is $t = \frac{\beta}{L}$.

Note that this convergence rate is very slow, and only tells us that we will find a stationary point. We need more structure about f to guarantee faster convergence and global optimality.