3 Gradient method

In this lecture we are interested in minimizing a smooth convex function f on $\mathbb{R}^n$:

$$f^* = \min_{x \in \mathbb{R}^n} f(x).$$

We assume minimum is finite and attained at some x^*. The gradient method we study has the form: Starting with any $x_0 \in \mathbb{R}^n$, iterate:

$$x_{k+1} = x_k - t_k \nabla f(x)$$

where t_k is the step size. See below for strategies to choose t_k. We now state a convergence result for the gradient method.

Theorem 3.1 (Convergence of gradient method)**.** *Assuming f is convex and has L-Lipschitz continuous gradient (wrt $\|\cdot\|_2$ norm), and assuming the step size is constant with $t_k = t \in (0, 1/L]$, we have $f(x_k) - f^* \leq \frac{1}{2tk}\|x_0 - x^*\|_2^2$ for all $k \geq 1$.*

The theorem tells us that to reach accuracy ϵ, it suffices to run the gradient method for $k = \frac{\|x_0 - x^*\|_2^2}{2t} \cdot \frac{1}{\epsilon}$.

Proof. For any $x \in \mathbb{R}^n$ we denote $x^+ = x - t\nabla f(x)$. By Lipschitz property of ∇f we know that

$$f(x^+) \leq f(x) + \nabla f(x)^T (x^+ - x) + \frac{L}{2}\|x^+ - x\|_2^2.$$

Now since $\nabla f(x) = -\frac{1}{t}(x^+ - x)$ we get

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

where in the last inequality we used the fact that $0 < t \leq 1/L$. Inequality (1) already tells us that the gradient method with $0 < t \leq 1/L$ is a *descent* method, i.e., the value of f decreases at each iteration.

Our goal is to analyze the accuracy $f(x_k) - f^*$ as the algorithm progresses. Convexity of f immediately tells us that $f(x) - f^* \leq \nabla f(x)^T (x - x^*)$. We combine this with inequality (1) above to understand how $f(x^+) - f^*$ evolves:

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

where in the last equality we used the fact that $\nabla f(x) = -(1/t)(x^+ - x)$. Using the identity $\|a\|_2^2 - 2a^T b = \|a - b\|_2^2 - \|b\|_2^2$ note that the right-hand side above is equal to $-\frac{1}{2t}\left[\|x^+ - x^*\|_2^2 - \|x - x^*\|_2^2\right]$. We have thus proved for any i:

$$f(x_{i+1}) - f^* \leq \frac{1}{2t}\left[\|x_i - x^*\|_2^2 - \|x_{i+1} - x^*\|_2^2\right].$$

We sum this inequality for $i = 0, \dots, k-1$ to get

$$\sum_{i=0}^{k-1}(f(x_{i+1}) - f^*) \leq \frac{1}{2t}\left[\|x_0 - x^*\|_2^2 - \|x_k - x^*\|_2^2\right] \leq \frac{1}{2t}\|x_0 - x^*\|_2^2.$$

Now since the function value decreases at each step we have $f(x_k) \leq f(x_{i+1})$ for all $i = 0, \dots, k-1$ and so

$$f(x_k) - f^* \leq \frac{1}{k}\sum_{i=0}^{k-1}(f(x_{i+1}) - f^*) \leq \frac{1}{2kt}\|x_0 - x^*\|_2^2.$$

□

Remark 1. *Nowhere in the proof did we actually use that x^* is a minimizer of f, and f^* is the minimum value! In fact, the proof gives an upper bound on $f(x_k) - f(u)$ for any choice of $u \in \mathbb{R}^n$. It's just that $f(x_k) - f(u)$ is not necessarily nonnegative so the theorem in this case only tells us that, "in the limit", $f(x_k) - f(u)$ will become ≤ 0.*

Line search In practice, we don't usually keep the step size t constant, but we operate a so-called *line search*. There are two main strategies for line search:

- Exact line search: at iteration k, search for the value of $t > 0$ that minimizes $f(x_k - t\nabla f(x_k))$. This is a one-dimensional minimization problem. Finding the exact minimum can be expensive, and so often it is enough to use:
- Backtracking line search: starting from large enough $t \leftarrow \hat{t}$ we keep $t \leftarrow \beta t$ for some $0 < \beta < 1$ until we satisfy a "*sufficient-decrease*" condition (typically called Armijo condition)

  $$f(x_k - t\nabla f(x_k)) \leq f(x_k) - \alpha t\|\nabla f(x_k)\|_2^2$$

  where α is a chosen constant $\in (0, 1)$, say $\alpha = 1/2$. Note that taking $\alpha = 0$ just asks for a t that decreases the value of f.

Analysis for strongly convex functions For strongly convex functions, the gradient method has a linear convergence rate.

Theorem 3.2. *Assume f is convex and has L-Lipschitz continuous gradient, and is m-strongly convex with $m > 0$. Then gradient method with constant step size $t = 2/(m + L)$ produces iterates (x_k) that satisfy*

$$\|x_k - x^*\|_2 \leq \left(\frac{1-\kappa}{1+\kappa}\right)^k \|x_0 - x^*\|_2 \quad \text{and} \quad f(x_k) - f^* \leq \frac{L}{2}\left(\frac{1-\kappa}{1+\kappa}\right)^{2k} \|x_0 - x^*\|_2^2 \tag{2}$$

where $\kappa = m/L \in (0, 1]$.

Theorem above tells us that if we want to reach accuracy ϵ on $f(x_k) - f^*$, it suffices to run the gradient method for $k \gtrapprox \frac{L}{m}\log\left(\frac{1}{\epsilon}\right)$ iterations.

Proof. We are going to assume that f is twice differentiable for convenience (there are proofs that do not require this assumption). Also note that the bound on $f(x_k) - f^*$ in (2) follows directly from the bound on $\|x_k - x^*\|_2$ since, by our smoothness assumption on f we have

$$f(x_k) - f(x^*) \leq \nabla f(x^*)^T(x_k - x^*) + \frac{L}{2}\|x_k - x^*\|_2^2 = \frac{L}{2}\|x_k - x^*\|_2^2.$$

We thus focus on proving the bound on $\|x_k - x^*\|_2$. By Taylor formula applied to ∇f we know that

$$\nabla f(x) = \underbrace{\nabla f(x^*)}_{=0} + \int_0^1 \nabla^2 f(x^* + \alpha(x - x^*))(x - x^*) d\alpha.$$

Recalling that $x^+ = x - t\nabla f(x)$, it thus follows that

$$\|x^+ - x^*\|_2 = \|(I - tM)(x - x^*)\|_2 \leq \|I - tM\|_2 \|x - x^*\|_2$$

where $M = \int_0^1 \nabla^2 f(x^* + \alpha(x - x^*)) d\alpha$ is a symmetric matrix. It suffices now to analyze the eigenvalues of $I - tM$. We know that the eigenvalues of M are all between $[m, L]$ by our assumption on f. Thus the eigenvalues of $I - tM$ are all in $[1 - tL, 1 - tm]$ and the spectral norm of $I - tM$ is $\gamma = \max\{|1 - tL|, |1 - tm|\}$. The best choice of t is when $1 - tL = -(1 - tm)$ which gives $t = \frac{2}{m+L}$ and then $\gamma = \frac{L-m}{L+m} = \frac{1-\kappa}{1+\kappa}$ where $\kappa = m/L$. It then follows that $\|x_k - x^*\|_2 \leq \left(\frac{1-\kappa}{1+\kappa}\right)^k \|x_0 - x^*\|_2$. $\square$