4 Gradient method

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

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

We assume the 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. We now state a convergence result for the gradient method, with constant step size, under the assumption the function f is L-smooth with respect to the Euclidean norm.

Theorem 4.1 (Convergence of gradient method)**.** *Assuming f is convex and L-smooth (wrt $\|\cdot\|_2$ norm). 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 $k = \frac{\|x_0 - x^*\|_2^2}{2t} \cdot \frac{1}{\epsilon}$ iterations.

Proof. For any $x \in \mathbb{R}^n$ we denote $x^+ = x - t\nabla f(x)$. By L-smoothness of f and the descent lemma we have

$$f(x^+) \leq f(x) + \left\langle \nabla f(x), x^+ - x \right\rangle + \frac{L}{2}\|x^+ - x\|_2^2.$$

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 $f(x) - f^* \leq \langle \nabla f(x), x - x^* \rangle$. We combine this with inequality (1) above to understand how $f(x^+) - f^*$ evolves:

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

where in the last equality we used the fact that $\nabla f(x) = -(1/t)(x^+ - x)$. Using the identity $\|a\|_2^2 - 2\langle a, b \rangle = \|a - b\|_2^2 - \|b\|_2^2$ 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*. A popular approach is so-called *backtracking line search*: starting from large enough $t \leftarrow \hat{t}$ we keep decreasing t by $t \leftarrow \beta t$ for some $0 < \beta < 1$ until we satisfy a "*sufficient-decrease*" condition (typically called Armijo condition)

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

Note that the condition above is precisely the inequality obtained by the descent lemma (1) that is needed for the analysis of the iterations.

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

Theorem 4.2. *Assume f is m-strongly convex and has L-Lipschitz continuous gradient with respect to the Euclidean norm $\|\cdot\|_2$. Then gradient method with constant step size $t = 2/(m+L)$ produces iterates (x_k) that satisfy*

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

where $\kappa = L/m \geq 1$.

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 $k \gtrsim \frac{L}{m}\log\left(\frac{1}{\epsilon}\right)$ iterations.

Proof. We are going to assume that f is twice continuously 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

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

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

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

where $M = \int_0^1 \nabla^2 f(x^* + \alpha(x - x^*))d\alpha$ is a symmetric matrix. Recalling that $x^+ = x - t\nabla f(x)$, it thus follows that

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

It suffices now to analyze the eigenvalues of $I - tM$. Our assumption on f tells us that $mI \preceq \nabla^2 f(y) \preceq LI$ for all y, and so, in particular all the eigenvalues of M are in $[m, L]$. 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 $\gamma = \frac{L-m}{L+m} = \frac{\kappa-1}{\kappa+1}$ where $\kappa = L/m$. It then follows that $\|x_k - x^*\|_2 \le \left(\frac{\kappa-1}{\kappa+1}\right)^k \|x_0 - x^*\|_2$. □

Illustration

- Consider the gradient method applied to the function $f(x) = \frac{1}{2}\|Ax - b\|_2^2$ with $A \in \mathbb{R}^{N \times n}$ and $N > n$, and A is full rank. We have $\nabla f(x) = A^T(Ax - b)$ and $\nabla^2 f(x) = A^T A$. We see that f is m-strongly convex and L-smooth with $m = \lambda_{\min}(A^T A) > 0$ (since A is full column rank) and $L = \lambda_{\max}(A^T A)$. Figure 1 below shows the convergence of the gradient method to the optimal value f^* ($N = 400, n = 200$). We observe a linear convergence rate, i.e., a straight line in a log plot of $f(x_k) - f^*$ vs. iteration number k.

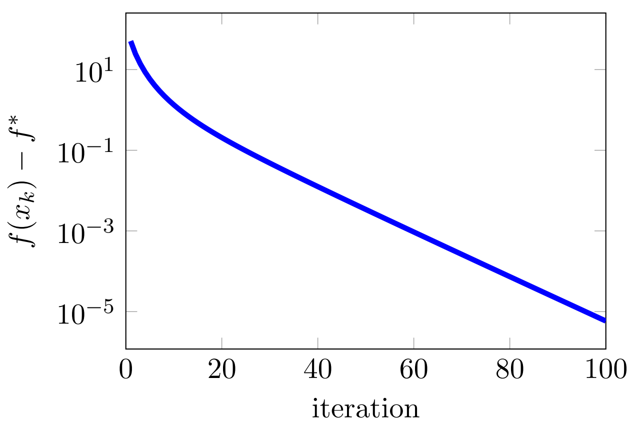

Figure 1: Gradient method for $f(x) = (1/2)\|Ax - b\|_2^2$ where $A \in \mathbb{R}^{N \times n}$ is full column rank. We observe linear convergence.

- Consider now the function $f(x) = \sum_{i=1}^N \log\left(1 + e^{a_i^T x + b_i}\right)$. This function is *not* strongly convex (note that $\log(1 + e^t) \approx t$ for t large). The plot in Figure 2 shows $f(x_k) - f^*$ as a function of k, and we clearly see a sublinear convergence rate.

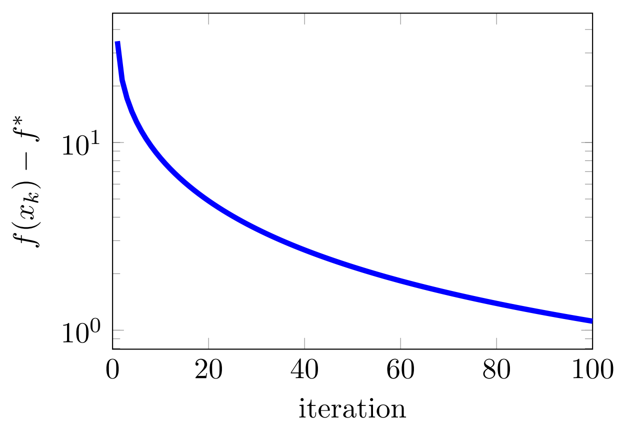

Figure 2: Gradient method for $f(x) = \sum_{i=1}^{N} \log\left(1 + e^{a_i^T x + b_i}\right)$. We observe a sublinear rate of convergence. Note that the function f is *not* strongly convex.