6 Subgradient method

In this lecture we look at the problem of minimizing a general nonsmooth convex function \( f(x) \).

**Subgradient method** The subgradient method to minimize \( f(x) \) works as follows. Choose \( x_0 \in \mathbb{R}^n \) and iterate, for \( k \geq 0 \):

\[
x_{k+1} = x_k - t_k g_k
\]

where \( g_k \in \partial f(x_k) \) is a subgradient of \( f \) at \( x_k \) and \( t_k > 0 \) is the step size.

Note: A negative subgradient is not necessarily a descent direction, i.e., it is possible that \( f(x - t g) > f(x) \) for all \( t > 0 \) (small enough). For example \( f(x) = |x|, x = 0 \) and \( g = -1 \in \partial f(0) \).

Convergence analysis of subgradient method:

\[
\|x_{k+1} - x^*\|_2^2 = \|x_k - t_k g_k - x^*\|_2^2
\]

\[
= \|x_k - x^*\|_2^2 - 2t_k g_k^T (x_k - x^*) + t_k^2 \|g_k\|_2^2
\]

\[
\leq \|x_k - x^*\|_2^2 + t_k^2 \|g_k\|_2^2 + 2t_k (f^* - f(x_k))
\]

where in the last line we used the fact that \( g_k \in \partial f(x_k) \). Applying this inequality recursively to \( \|x_k - x^*\|_2^2 \), we get at the end:

\[
\|x_{k+1} - x^*\|_2^2 \leq \|x_0 - x^*\|_2^2 + \sum_{i=0}^{k} t_i^2 \|g_i\|_2^2 + 2 \sum_{i=0}^{k} t_i (f^* - f(x_i))
\]

which after rearranging, and using \( \|x_{k+1} - x^*\|_2^2 \geq 0 \), gives us

\[
\sum_{i=0}^{k} t_i (f(x_i) - f^*) \leq \frac{\|x_0 - x^*\|_2^2}{2} + \frac{1}{2} \sum_{i=0}^{k} t_i^2 \|g_i\|_2^2.
\]

Let \( f_{\text{best},k} = \min \{f(x_0), \ldots, f(x_k)\} \). Since \( t_i \geq 0 \) we get

\[
f_{\text{best},k} - f^* \leq \frac{1}{\sum_{i=0}^{k} t_i} \sum_{i=0}^{k} t_i (f(x_i) - f^*) \leq \frac{\|x_0 - x^*\|_2^2}{2 \sum_{i=0}^{k} t_i} + \frac{\sum_{i=0}^{k} t_i^2 \|g_i\|_2^2}{2 \sum_{i=0}^{k} t_i}
\]

\[
\leq \frac{\|x_0 - x^*\|_2^2}{2 \sum_{i=0}^{k} t_i} + \frac{G^2 \sum_{i=0}^{k} t_i^2}{2 \sum_{i=0}^{k} t_i}.
\]

where in the last equation we assumed that \( f \) is \( G \)-Lipschitz, so that \( \|g_i\|_2 \leq G \) (see Exercise sheet 2).

- Constant step size: If \( t_k = t \) and \( f \) is \( G \)-Lipschitz then we get

\[
f_{\text{best},k} - f^* \leq \frac{\|x_0 - x^*\|_2^2}{2(k+1)t} + \frac{G^2 t}{2}.
\]

In this case we do not guarantee convergence: we only guarantee that \( f_{\text{best},k} \) will be at most \( G^2 t / 2 \) sub-optimal, in the limit \( k \to \infty \).
Assume that \( k \) is fixed a priori (i.e., we have a certain number of iterations that we are going to run). What is the choice of \( t \) that minimizes the right-hand side of (4)? The choice of \( t \) is the one that will make the two terms equal, namely \( \|x_0 - x^*\|^2_2/(k+1) = G^2 t^2 \), i.e.,

\[
t = \frac{\|x_0 - x^*\|^2}{G\sqrt{k+1}} \Rightarrow f_{\text{best},k} - f^* \leq \frac{GR}{\sqrt{k+1}}.
\]

- Diminishing step size: consider the choice \( t_i \sim 1/\sqrt{i} \). Then \( \sum_{i=0}^{k} t_i \approx \sqrt{k} \), \( \sum_{i=0}^{k} t_i^2 \approx \ln(k) \), and so we get a convergence like \( \ln(k)/\sqrt{k} \). In fact, one can get rid of the log term by recursing the inequality (1) only up to iterate \( k/2 \) (instead of all the way back to the first iterate), and use the fact that \( \sum_{k/2}^{1} 1/i \leq \text{constant} \).

**Illustration** The figure below shows the subgradient method applied to the problem of minimizing the nonsmooth function \( f(x) = \|Ax - b\|_1 \) where \( A \in \mathbb{R}^{m \times n} \) with \( m > n \), and \( b \in \mathbb{R}^m \). We see that with a constant step size, the method does not converge to \( f^* \), but only to a neighborhood of the optimal value.

![Illustration of subgradient method](image)

**Optimality of subgradient method** One can show that the convergence rate of \( 1/\sqrt{k} \) is the best possible one can get on the class of nonsmooth convex Lipschitz functions. More precisely, fix \( k, G, \) and \( R > 0 \). For any algorithm where the \( k \)'th iterate satisfies

\[
x_k \in x_0 + \text{span}\{g_1, \ldots, g_k\}
\]

where \( g_i \in \partial f(x_i) \) and \( x_0 \) is the starting point, there is a convex function \( f \) that is \( G \)-Lipschitz on \( \{x : \|x - x_0\|_2 \leq R\} \) such that after \( k \) iterations of the algorithm we have

\[
f_{\text{best},k} - f^* \geq \frac{GR}{\sqrt{k+1}}.
\]

See Exercise sheet 2 for a proof.