1 Steepest descent method

Recall the unconstrained problem: \( \min_x f(x) \)

The first-order Taylor approximation of \( f(x + v) \) around \( x \) is:

\[
f(x + v) \approx f(x) + \nabla f(x)^T v
\]

which gives the approximate change in \( f \) for a small step \( v \) (chosen as a descent direction, i.e. \( \nabla f(x)^T v < 0 \)).

**Definition 1.** We define a normalized steepest descent direction \( \Delta x_{nsd} \) (w.r.t. the norm \( \| \cdot \| \)) as a step of unit norm that gives the largest decrease in the linear approximation of \( f \), i.e. for any norm \( \| \cdot \| \):

\[
\Delta x_{nsd} = \text{argmin}\{\nabla f(x)^T v | \|v\| = 1\}
\]

**Note 1.** A normalized steepest descent direction can be interpreted geometrically as follows. We can just as well define \( \Delta x_{nsd} \) as

\[
\Delta x_{nsd} = \text{argmin}\{\nabla f(x)^T v | \|v\| \leq 1\}
\]

i.e., as the direction in the unit ball of \( \| \cdot \| \) that extends farthest in the direction \( -\nabla f(x) \).

**Problem 1.** Compute \( \Delta x_{sd} \) when the norm \( \| \cdot \| \) in the steepest descent is:

a) the Euclidean norm.

b) the quadratic norm: \( \|z\|_p = (z^T P z)^{1/2} = \|P^{1/2} z\|_2 \) where \( P \in \mathbb{S}_{++}^n \).

![Figure 1: Normalized steepest descent direction for a quadratic norm. The ellipsoid shown is the unit ball of the norm, translated to the point \( x \). The normalized steepest descent direction \( \Delta x_{nsd} \) at \( x \) extends as far as possible in the direction \( -\nabla f(x) \) while staying in the ellipsoid. The gradient and normalized steepest descent directions are shown.](image-url)
Problem 2. We define the change of coordinates $\bar{u} = P^{1/2}u$. We can solve the original problem of minimizing $f$ by solving the equivalent problem of minimizing $\bar{f}$ in the new system of coordinates.

a) We have $\bar{f}(\bar{u}) = f(u)$. Express $\bar{f}(\bar{u})$ in terms of $P$, $f$, and $u$.
b) We apply the gradient method to $\bar{f}$: $\Delta \bar{x} = -\nabla \bar{f}(\bar{x})$. Express $\Delta \bar{x}$ in terms of $P$, $f$, and $x$.
c) Show that the steepest descent method in the quadratic norm $\| \cdot \|_P$ can be thought of as the gradient method applied to the problem after the change of coordinates $\bar{x} = P^{1/2}x$.

2 Newton’s method

For $x \in \text{dom } f$, the vector
$$\Delta x_{nt} = -\nabla^2 f(x)^{-1}\nabla f(x)$$
is called the Newton step. Positive definiteness of $\nabla^2 f(x)$ implies that
$$\nabla f(x)^T \Delta x_{nt} = -\nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x) < 0$$
unless $\nabla f(x) = 0$, so the Newton step is a descent direction (unless $x$ is optimal).

Problem 3. Show that the Newton step $\Delta x_{nt} = -\nabla^2 f(x)^{-1}\nabla f(x)$ is:

a) the minimized of the second-order approximation of $f$ at $x$.
b) the steepest descent direction at $x$, for the quadratic norm defined by the Hessian $\nabla^2 f(x)$, i.e.,
$$\| u \|_{\nabla^2 f(x)} = (u^T \nabla^2 f(x) u)^{1/2}$$
c) Argue why $x + \Delta x_{ns}$ should be a very good estimate of the minimizer of $f$, i.e., $x^*$.

Figure 2: Iterates of steepest descent for the function $f(x_1, x_2) = e^{x_1 + 3x_2 - 0.1} + e^{x_1 - 3x_2 - 0.1} + e^{-x_1 - 0.1}$. Steepest descent method with a quadratic norm $\| \cdot \|_{P_1}$ (left) and with a quadratic norm $\| \cdot \|_{P_2}$ (right). The ellipses are the boundaries of the norm balls $\{ x \mid \| x - x^{(k)} \|_P \leq 1 \}$ at $x^{(0)}$ and $x^{(1)}$. We have $P_1 = \text{diag}(2, 8)$ and $P_2 = \text{diag}(8, 2)$. 


Figure 3: Iterates of steepest descent with norm $\| \cdot \|_{P_1}$ (left) and iterated of steepest descent with norm $\| \cdot \|_{P_2}$ (right) after the change of coordinates. On the left, the change of coordinates reduces the condition number of the sublevel sets, and so speeds up convergence. On the right, the change of coordinates increases the condition number of the sublevel sets, and so slows down convergence.

3 Notes

The choice of norm used to define the steepest descent direction can have a dramatic effect on the convergence rate. For simplicity, we consider the case of steepest descent with quadratic $P$-norm.

We showed that the steepest descent method with quadratic $P$-norm is the same as the gradient method applied to the problem after the change of coordinates $\bar{x} = P^{1/2}x$. We know that the gradient method works well when the condition numbers of the sublevel sets (or the Hessian near the optimal point) are moderate, and works poorly when the condition numbers are large. It follows that when the sublevel sets, after the change of coordinates $\bar{x} = P^{1/2}x$, are moderately conditioned, the steepest descent method will work well.

This observation provides a prescription for choosing $P$: It should be chosen so that the sublevel sets of $f$, transformed by $P^{-1/2}$, are well conditioned. For example if an approximation $\hat{H}$ of the Hessian at the optimal point $H(x^*)$ were known, a very good choice of $P$ would be $P = \hat{H}$, since the Hessian of $\tilde{f}$ at the optimum is then

$$\hat{H}^{-1/2}\nabla^2 f(x^*) \hat{H}^{-1/2} \approx I,$$

and so is likely to have a low condition number.

This same idea can be described without a change of coordinates. Saying that a sublevel set has low condition number after the change of coordinates $\bar{x} = P^{1/2}x$ is the same as saying that the ellipsoid

$$\mathcal{E} = \{ x \mid x^T P x \leq 1 \}$$

approximates the shape of the sublevel set. (In other words, it gives a good approximation after appropriate scaling and translation.)