Consider the problem

$$\min_{x \in \text{dom} f} (x)$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is convex and twice continuously differentiable. We assume that there exists a minimizer $x^*$, with $f(x^*) = p^*$. A necessary and sufficient condition for optimality is

$$\nabla f(x^*) = 0.$$

Assume also that $f$ is closed, i.e., $\text{epi} f$, the epigraph of $f$, is closed, so that

$$S = \{x \in \text{dom} f : f(x) \leq f(x(0))\}$$

is closed for all $x(0) \in \text{dom} f$. (See BV, p. 640, for examples of when $f$ is or is not closed.) Assume further that $f$ is strongly convex,\(^1\) which means $\exists m > 0$ such that the Hessian of $f$ satisfies

$$\nabla^2 f(x) \succeq mI \quad \forall x \in S.$$

Equivalently, the least eigenvalue of $\nabla^2 f(x)$ is uniformly bounded below by $m$. By Taylor’s theorem in one variable, given $x, y \in S$, we have

$$f(y) = f(x) + \nabla f(x)^T(y - x) + \frac{1}{2}(y - x)\nabla^2 f(z)(y - x)$$

\(^1\)Not to be confused with strictly convex, which means that the inequality in the convexity definition holds strictly. The function $e^x$ is strictly convex but not strongly convex.
for some \( z \) in the line segment \([x, y]\). Thus
\[
f(y) \geq f(x) + \nabla f(x)^T(y - x) + \frac{m}{2}(y - x)^T(y - x). \tag{1}
\]
Setting \( m = 0 \) gives the first-order property of convex functions that we proved in Lecture 2 (BV eq. (3.2)).

The right-hand side of (1) is a convex function of \( y \) (for fixed \( x \)). Let’s set its gradient (w.r.t. \( y \)) to zero:
\[
\nabla f(x) + m(y - x) = 0,
\]
so the right-hand side of (1) is minimized by
\[
\bar{y} = x - \frac{1}{m} \nabla f(x).
\]
Thus we have
\[
f(y) \geq f(x) + \nabla f(x)^T(\bar{y} - x) + \frac{m}{2}(\bar{y} - x)^T(\bar{y} - x)
\]
\[
= f(x) + \nabla f(x)^T(-\frac{1}{m} \nabla f(x)) + \frac{m}{2} \frac{1}{m^2} \nabla f(x)^T \nabla f(x)
\]
\[
= f(x) - \frac{1}{2m} \| \nabla f(x) \|^2.
\]
This is true for all \( y \in S \), so
\[
p^* \geq f(x) - \frac{1}{2m} \| \nabla f(x) \|^2. \tag{2}
\]
Also, inequality (1) implies that \( S \) is bounded (otherwise \( \|y - x\| \) can be arbitrarily large, violating the condition that \( x, y \in S \)). So, since \( f \) is twice continuously differentiable, \( \| \nabla^2 f(x) \| \) is bounded above by some \( M \) on \( S \), and hence strong convexity actually implies
\[
MI \succeq \nabla^2 f(x) \succeq mI \quad \forall x \in S.
\]
So, similarly to (2), we get
\[
f(y) \leq f(x) + \nabla f(x)^T(y - x) + \frac{M}{2}(y - x)^T(y - x), \tag{3}
\]
and, taking the gradient of the right-hand side w.r.t. \( y \) as before and setting it to zero, we find
\[
f(y) \leq f(x) - \frac{1}{2M} \| \nabla f(x) \|^2
\]
so
\[
p^* \leq f(x) - \frac{1}{2M} \| \nabla f(x) \|^2. \tag{4}
\]
Descent Methods

For $k = 0, 1, 2, \ldots$,

- Choose a “descent direction” $\Delta x$, with $\nabla f(x^{(k)})^T \Delta x < 0$
- Do a “line search”: find $x^{(k+1)} = x^{(k)} + t \Delta x$ satisfying

$$f(x^{(k+1)}) \leq f(x^{(k)}) + \alpha t \nabla f(x^{(k)})^T (\Delta x)$$

(5)

where $\alpha$ is the “Armijo” parameter. Setting $\alpha = 0$ will not guarantee convergence. Assume $0 < \alpha < \frac{1}{2}$.

Exact line search

Choose $t = t_{ELS}$ where $\nabla f(x^{(k)}) + t_{ELS} \Delta x)^T (\Delta x) = 0$. Usually not practical.

Backtracking line search

- Start with $t = 1$
- While $f(x^{(k)} + t \Delta x) > f(x^{(k)}) + \alpha t \nabla f(x^{(k)})^T (\Delta x)$ do: $t \leftarrow \beta t$

where $\beta$ is the “backtracking” parameter with $0 < \beta < 1$. (Normally $\beta = \frac{1}{2}$.)

It’s not hard to prove using convexity (and is easy to see from a picture) that the Armijo descent condition is satisfied on a nontrivial interval $[0, t_0]$. Thus, eventually the Armijo condition must be satisfied. In fact, the backtracking line search must either set $t = 1$ (Armijo condition satisfied immediately) or $t \in [\beta t_0, t_0]$ (the final step where the Armijo condition failed for some $t > t_0$ led to the current value $t > \beta t_0$). So, the final step satisfies $t \geq \min(1, \beta t_0)$.

Gradient descent, also known as steepest descent (in the 2-norm):

$$\Delta x = -\nabla f(x^{(k)})$$
**Convergence analysis**

From (3), with \( x = x^{(k)} \), \( y = x^{(k)} + t(\Delta x) = x^{(k)} - t\nabla f(x^{(k)}) \), we have

\[
f(x^{(k)} + t(\Delta x)) \leq f(x^{(k)}) - t\|\nabla f(x^{(k)})\|^2_2 + \frac{M}{2}t^2\|\nabla f(x^{(k)})\|^2_2
\]

(6)

\[
= f(x^{(k)}) + \left(\frac{Mt^2}{2} - t\right)\|\nabla f(x^{(k)})\|^2_2.
\]

The right-hand side is minimized by \( t = \frac{1}{M} \), so let’s first consider a **fixed-step method** with \( t = \frac{1}{M} \) (which, of course, may not be known). Then

\[
f(x^{(k+1)}) \leq f(x^{(k)}) - \frac{1}{2M}\|\nabla f(x^{(k)})\|^2_2.
\]

(7)

We have \( \|\nabla f(x^{(k)})\|^2_2 \geq 2m(f(x^{(k)}) - p^*) \) from (2), so

\[
f(x^{(k+1)}) - p^* \leq \left(1 - \frac{m}{M}\right)(f(x^{(k)}) - p^*).
\]

This is true for \( k = 1, 2, \ldots \) so

\[
f(x^{(\ell)}) - p^* \leq \left(1 - \frac{m}{M}\right)^\ell(f(x^{(0)}) - p^*).
\]

This is called **linear or geometric convergence**. This is slow if the **condition number** of \( f \), defined to be \( \frac{M}{m} \) and denoted \( \kappa \), is large (in this case we say \( f \) is **ill-conditioned**. (This is not necessarily the worst case condition number of \( \nabla^2 f(x) \) over all \( x \in S \), but it is an upper bound on this.) Let \( c = 1 - \frac{m}{M} \) and \( \epsilon_0 = f(x_0) - p_* \). Then

\[
f(x^{(\ell)}) - p^* \leq c^\ell(f(x^{(0)}) - p^*),
\]

so if we want the left-hand side to be at most \( \epsilon \), we have that guarantee as long as

\[
c^\ell \leq \frac{\epsilon}{\epsilon_0}
\]

i.e.,

\[
\ell \log c \leq \log \frac{\epsilon}{\epsilon_0}
\]

i.e., the worst case number of iterations is bounded by

\[
\ell \geq \frac{\log(\epsilon_0/\epsilon)}{\log(1/c)}.
\]
Note that when $\kappa = M/m$ is big, we have
\[
\log \frac{1}{c} = -\log \left(1 - \frac{m}{M}\right) \approx \frac{m}{M},
\]
so the denominator in the bound on $\ell$ is small. We sometimes say the number of iterations is $O(\log(1/\epsilon))$, absorbing the information about $c$ and $\epsilon_0$ into the constant in the “big $O$”.

An exact line search may do better than this, because it would minimize the left-hand side of (6), while the fixed step $t = 1/M$ minimizes the upper bound on the right-hand side. It cannot do worse.

However, an exact line search is expensive and we may not know $M$, in which case we may want to use the backtracking line search, so let us give a convergence analysis for that. For gradient descent, the Armijo condition (5) becomes
\[
f(x^{(k)} + t(\Delta x)) \leq f(x^{(k)}) - \alpha t \|\nabla f(x^{(k)})\|^2
\]
which holds for $t = 1/M$ by (7) since $\alpha < 1/2$. Furthermore, by convexity it is clear that since the Armijo condition is satisfied for $t = 1/M$, it must be satisfied for all $t < 1/M$ as well. Hence, $t_0 \geq 1/M$, and so since the $t$ computed by the backtracking line search satisfies $t \geq \min(1, \beta t_0)$, it follows that $t \geq \min(1, \beta/M)$. If $t = 1$ we have
\[
f(x^{(k)} + t(\Delta x)) \leq f(x^{(k)}) - \alpha \|\nabla f(x^{(k)})\|^2
\]
and otherwise we have
\[
f(x^{(k)} + t(\Delta x)) \leq f(x^{(k)}) - \frac{\alpha \beta}{M} \|\nabla f(x^{(k)})\|^2
\]
so either way we have
\[
f(x^{(k)} + t(\Delta x)) \leq f(x^{(k)}) - \min \left(\alpha, \frac{\alpha \beta}{M}\right) \|\nabla f(x^{(k)})\|^2.
\]
The rest of the analysis is as earlier: we now get
\[
f(x^{(k+1)}) - p^* \leq (1 - 2m\alpha \min \left(1, \frac{\beta}{M}\right) (f(x^{(k)}) - p^*)
\]
so
\[
f(x^{(\ell)}) - p^* \leq c^\ell (f(x^{(0)}) - p^*)
\]
with $c = 1 - 2m\alpha \min(1, \beta/M)$. If we take $\beta = \frac{1}{2}$ and assume $M \geq \frac{1}{2}$, we have $c = 1 - \alpha m/M$ — instead of $c = 1 - m/M$ for the fixed step analysis: not much different if we avoid talking $\alpha$ too small.