Note on the Gradient Method for Smooth Convex Minimization

Michael L. Overton
following the derivation in Boyd and Vandenberghe

Spring 2022

Consider the problem
$$\min_{x \in \mathrm{dom} f} f(x)$$
where $f : \mathbb{R}^n \to \mathbb{R}$ is convex and twice continuously differentiable on the domain $\mathrm{dom} f$, which is assumed to be open. 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., $\mathrm{epi} f$, the epigraph of f, is closed, so that

$$S = \{x \in \mathrm{dom} f : f(x) \le f(x^{(0)})\}$$

is closed for all $x^{(0)} \in \mathrm{dom} f$. (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[2]

$$\tilde{y} = x - \frac{1}{m}\nabla f(x).$$

Thus we have, for all $y \in S$,

$$\begin{aligned} f(y) &\geq f(x) + \nabla f(x)^T (\tilde{y} - x) + \frac{m}{2}(\tilde{y} - x)^T (\tilde{y} - x) \\ &= f(x) + \nabla f(x)^T (-\frac{1}{m}\nabla f(x)) + \frac{m}{2}\frac{1}{m^2}\nabla f(x)^T \nabla f(x) \\ &= f(x) - \frac{1}{2m}\|\nabla f(x)\|_2^2. \end{aligned}$$

Hence

$$p^* \geq f(x) - \frac{1}{2m}\|\nabla f(x)\|_2^2. \tag{2}$$

Also, inequality (1) implies that S is bounded (otherwise $\|y - x_0\|$ 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[3]

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

So, similarly to (1), we get, for all $x, y \in S$,

$$f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{M}{2}(y - x)^T (y - x). \tag{3}$$

[2]Note that this is true regardless of whether $\tilde{y} \in S$.

[3]It will be useful later to assume that M and m are respectively the lowest possible upper bound and the largest possible lower bound.

Now we minimize *both sides* of this inequality over $y \in S$. On the left this gives p^*; on the right, we minimize the quadratic form just as we did with the quadratic form on the right-hand side of (1), setting its gradient w.r.t. y to zero and finding [4]

$$p^* \le f(x) - \frac{1}{2M}\|\nabla f(x)\|_2^2. \tag{4}$$

Descent Methods

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

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

$$f(x^{(k+1)}) \le f(x^{(k)}) + \alpha t \nabla f(x^{(k)})^T(\Delta x) \tag{5}$$

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

[4] A former student Zijian Liu pointed out that although BV don't say so, in order for this inequality to be valid, we need to check that the minimizer of the quadratic form on the right, namely $x - (1/M)\nabla f(x)$, is actually in S. Zijian gave a proof but Lieven Vandenberghe gave a slightly simpler proof to me by email, as follows. Let $[0, \delta] = \{t \ge 0 : x + tv \in S\}$ where $v = -\nabla f(x)$ and $x \in S$. The interval $[0, \delta]$ is closed because we assume S is closed. The inequality (3) implies that $f(x + tv) \le f(x) - t(1 - (M/2)t)\|v\|_2^2$ for $t \in [0, \delta]$. We must have $\delta \ge 2/M$ because otherwise this inequality implies that $f(x + \delta v) < f(x) \le f(x_0)$, which is impossible since $x + \delta v$ is on the boundary of S and $\mathrm{dom} f$ is always assumed to be open. So, since $\delta \ge 2/M$, we have $x + (1/M)v \in S$.

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

$$\begin{aligned} f(x^{(k)} + t(\Delta x)) &\leq f(x^{(k)}) - t\|\nabla f(x^{(k)}\|_2^2 + \frac{M}{2}t^2\|\nabla f(x^{(k)}\|_2^2 \qquad (6) \\ &= f(x^{(k)}) + \left(\frac{Mt^2}{2} - t\right)\|\nabla f(x^{(k)}\|_2^2. \end{aligned}$$

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

$$f(x^{(k+1)}) \leq f(x^{(k)}) - \frac{1}{2M}\|\nabla f(x^{(k)}\|_2^2. \qquad (7)$$

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

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

This is called *linear* or *geometric* convergence. This is slow if the *condition number* of the convex function f, defined to be $\frac{M}{m}$ and denoted κ, is large (in this case we say f is *ill-conditioned*). Let $\lambda_{\max}$ and $\lambda_{\min}$ denote largest and smallest eigenvalues. Then since

$$\kappa = \frac{M}{m} = \frac{\max_{x \in S} \lambda_{\max}(\nabla^2 f(x))}{\min_{x \in S} \lambda_{\min}(\nabla^2 f(x))},$$

this is an upper bound on the worst case matrix condition number of $\nabla^2 f(x)$, namely

$$\max_{x \in S} \frac{\lambda_{\max}(\nabla^2 f(x))}{\lambda_{\min}(\nabla^2 f(x))}.$$

Now let $c = 1 - \frac{m}{M}$ and $\epsilon_0 = f(x_0) - p_*$. Then

$$f(x^{(\ell)}) - p^* \leq c^\ell \left(f(x^{(0)}) - p^*\right),$$

so if we want the left-hand side to be at most ϵ, 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 ℓ is small. We sometimes say the number of iterations is $O(\log(1/\epsilon)$, absorbing the information about c and ϵ_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

$$f(x^{(k)} + t(\Delta x)) \leq f(x^{(k)}) - \alpha t \|\nabla f(x^{(k)}\|_2^2$$

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

$$f(x^{(k)} + t(\Delta x)) \leq f(x^{(k)}) - \alpha \|\nabla f(x^{(k)}\|_2^2$$

and otherwise we have

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

so either way we have

$$f(x^{(k)} + t(\Delta x)) \leq f(x^{(k)}) - \min\left(\alpha, \frac{\alpha\beta}{M}\right) \|\nabla f(x^{(k)}\|_2^2.$$

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 α too small.