Gradient Methods for Convex Optimization

YoonHaeng Hur

February 24, 2026

In this note, we consider the following minimization problem:

$$\begin{aligned} \text{minimize} \quad & f(x), \\ \text{subject to} \quad & x \in C, \end{aligned}$$

where $f\colon \mathbb{R}^d \to (-\infty, \infty]$ and $C \subset \mathrm{dom}(f) := \{x \in \mathbb{R}^d : f(x) < \infty\}$. Let $f^* = \inf_{x \in C} f(x)$ denote the optimal value of this minimization problem. In this note, we study fundamental methods that use the gradient of f to approximate f^*.

Contents

1 Gradient Descent under Smoothness

Throughout this section, we consider the unconstrained case, namely, $C = \mathbb{R}^d$; we also assume that $f\colon \mathbb{R}^d \to \mathbb{R}$ is differentiable. We study a method called gradient descent defined as follows: given an initial point $x_0 \in \mathbb{R}^d$, for $k \geq 0$, iterate

$$x_{k+1} \leftarrow x_k - t_k \nabla f(x_k) \quad \text{for some suitable } t_k > 0,$$

where we call $t_k > 0$ the step size at x_k. In other words, gradient descent updates a point $x_k \in \mathbb{R}^d$ by moving it along the direction parallel to $-\nabla f(x_k)$ by a suitable amount according to the step size $t_k > 0$. The rationale behind using such an update rule for minimizing f is motivated by the fact that $-\nabla f(x_k)$ is a direction that decreases f at point x_k.

Descent directions Given $\nabla f(x_k) \neq 0$, we call $v \in \mathbb{R}^d$ a descent direction of f at x_k if the directional derivative $D_v f(x_k) = \langle \nabla f(x_k), v \rangle$ is negative. For any descent direction $v \in \mathbb{R}^d$,

$$f(x_k + tv) = f(x_k) + t\langle \nabla f(x_k), v \rangle + o(t), \tag{1}$$

which implies $f(x_k + tv) < f(x_k)$ for sufficiently small $t > 0$ such that $|o(t)| < -t\langle \nabla f(x_k), v \rangle$. Clearly, $-\nabla f(x_k)$ is a descent direction; in fact, it is a steepest descent direction, which provides a direction in which f decreases most rapidly in the following sense: provided $\nabla f(x_k) \neq 0$,

$$\operatorname*{arg\,min}_{v \in \mathbb{S}^{d-1}} D_v f(x_k) = \operatorname*{arg\,min}_{v \in \mathbb{S}^{d-1}} \langle \nabla f(x_k), v \rangle = -\frac{\nabla f(x_k)}{\|\nabla f(x_k)\|_2}.$$

In particular, taking $-\nabla f(x_k)$ as a descent direction, (1) becomes

$$f(x_k - t\nabla f(x_k)) = f(x_k) - t\|\nabla f(x_k)\|_2^2 + o(t). \tag{2}$$

Again, as long as x_k is not a stationary point, i.e., $\nabla f(x_k) \neq 0$, we can always (monotonically) decrease the value of f by taking $t_k > 0$ sufficiently small so that $o(t_k) \leq t_k \|\nabla f(x_k)\|_2^2$ and updating x_k to the point $x_{k+1} \leftarrow x_k - t_k \nabla f(x_k)$. Therefore, gradient descent is indeed a descent method, namely, $f(x_{k+1}) \leq f(x_k)$ for a sufficiently small step size t_k, which leads to the name "gradient descent".

Advantages of gradient descent Gradient descent is implementable as long as $\nabla f(x)$ is computable for any $x \in \mathbb{R}^d$. Therefore, gradient descent is suitable for applications where the computation ∇f, which we often call the first-order oracle, is inexpensive. Especially, gradient descent requires less memory than other algorithms based on higher order oracles, e.g., the second derivative $\nabla^2 f \in \mathbb{R}^{d \times d}$.

Remark 1. Note that gradient descent stops if $\nabla f(x_k) = 0$ for some $k \geq 0$, i.e., it stops if it reaches a stationary point. In practice, one may specify a stopping criterion so that gradient descent terminates once $\nabla f(x_k)$ is sufficiently close to 0. For instance, gradient descent terminates if $\|\nabla f(x_k)\|_2 \leq \varepsilon$, namely, x_k is an ε-stationary point of f, where ε is a user-specified stopping tolerance.

Remark 2 (Other interpretations)**.** There are other ways to interpret gradient descent update rule $x_{k+1} \leftarrow x_k - t_k \nabla f(x_k)$.

- Quadratic approximation: for any $t_k > 0$, note that $x_k - t_k \nabla f(x_k)$ is the unique minimizer of the following quadratic function

$$x \mapsto f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2t_k}\|x - x_k\|_2^2.$$

  In other words,

$$x_{k+1} = \operatorname*{arg\,min}_{x \in \mathbb{R}^d} \left(f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2t_k}\|x - x_k\|_2^2 \right).$$

- Local first-order approximation: one can also view it as the unique minimizer of the following local first-order approximation

$$x_{k+1} = \operatorname*{arg\,min}_{\substack{x \in \mathbb{R}^d \\ \|x - x_k\|_2 \le r_k}} \left(f(x_k) + \langle \nabla f(x_k), x - x_k \rangle \right),$$

  where $r_k = t_k \|\nabla f(x_k)\|_2$.

1.1 Gradient descent: sufficient decrease under smoothness

We have mentioned that gradient descent is a descent method under the differentiability of f, namely, $f(x_{k+1}) \le f(x_k)$ if the step size $t_k > 0$ is sufficiently small. This means that $f(x_k)$ will converge as $k \to \infty$ provided f is bounded below, i.e., $f^* > -\infty$. Of course, there is no guarantee that the limit $\lim_{k\to\infty} f(x_k)$ is the minimum f^*; in other words, it is impossible to quantify the gap between $\lim_{k\to\infty} f(x_k)$ and f^* without any further assumptions.

In this section, we focus on the most fundamental assumption (smoothness of f) for gradient descent to produce sufficient decrease in f. We will later see that sufficient decrease enables gradient descent to find a stationary point, which will play a crucial role in finding the minimum f^* under convexity.

Definition 1. $f \colon \mathbb{R}^d \to \mathbb{R}$ is said to be L-smooth if

$$\|\nabla f(x) - \nabla f(y)\|_2 \le L\|x - y\|_2 \quad \forall x, y \in \mathbb{R}^d.$$

In other words, ∇f is L-Lipschitz.

In (2), we have already seen that $o(t_k) \le t_k \|\nabla f(x_k)\|_2^2$ leads to $f(x_{k+1}) \le f(x_k)$. From this, one can deduce that the key to the sufficient decrease, namely, $f(x_{k+1})$ is sufficiently smaller than $f(x_k)$, is to control the error of the first-order approximation, i.e., $o(t)$ in (2). The next lemma shows that L-smoothness yields $o(t) = O(t^2)$ so that $o(t)$ converges to 0 much faster than t as $t \to 0$.

Lemma 1. *If $f \colon \mathbb{R}^d \to \mathbb{R}$ is L-smooth, we have*

$$|f(y) - f(x) - \langle \nabla f(x), y - x \rangle| \le \frac{L}{2}\|y - x\|_2^2 \quad \forall x, y \in \mathbb{R}^d.$$

Proof. Let $g(t) = f(x + t(y - x))$ for $t \in [0, 1]$; then, $g'(t) = \langle \nabla f(x + t(y - x)), y - x \rangle$, $g(0) = f(x)$, and $g(1) = f(y)$. Since

$$f(y) - f(x) = g(1) - g(0) = \int_0^1 g'(t)\,\mathrm{d}t = \int_0^1 \langle \nabla f(x + t(y - x)), y - x \rangle\,\mathrm{d}t,$$

we have

$$\begin{aligned}
|f(y) - f(x) - \langle \nabla f(x), y - x \rangle| &= \left| \int_0^1 \langle \nabla f(x + t(y - x)) - \nabla f(x), y - x \rangle\,\mathrm{d}t \right| \\
&\leq \int_0^1 \|\nabla f(x + t(y - x)) - \nabla f(x)\|_2 \|y - x\|_2\,\mathrm{d}t \\
&\leq \int_0^1 Lt\|y - x\|_2^2\,\mathrm{d}t \\
&= \frac{L}{2}\|y - x\|_2^2,
\end{aligned}$$

where the first inequality follows from the Cauchy-Schwarz inequality and the second inequality follows from the L-smoothness of f. □

By Lemma 1, we have the following concrete upper bound on the error of the first-order approximation $o(t)$ in (2):

$$|o(t)| = \left| f(x_k - t\nabla f(x_k)) - f(x_k) + t\|\nabla f(x_k)\|_2^2 \right| \leq \frac{Lt^2}{2}\|\nabla f(x_k)\|_2^2,$$

Therefore,

$$f(x_k - t\nabla f(x_k)) \leq f(x_k) - t\left(1 - \frac{Lt}{2}\right)\|\nabla f(x_k)\|_2^2,$$

implying that updating from x_k to $x_k - t\nabla f(x_k)$ monotonically decreases the value of f provided $0 < t \leq \frac{2}{L}$; in other words, gradient descent is a descent method.

Particularly, we have

$$f(x_k - t\nabla f(x_k)) \leq f(x_k) - \frac{t}{2}\|\nabla f(x_k)\|_2^2 \quad \text{if } 0 < t \leq \frac{1}{L}, \tag{3}$$

which means that $x_k \to x_k - t\nabla f(x_k)$ decreases the value of f by at least $\frac{t}{2}\|\nabla f(x_k)\|_2^2$. We often say that (3) provides sufficient decrease.

Choosing the step size When L is known, letting $t_k = \frac{1}{L}$ for all $k \geq 0$, namely, taking the constant step size, is the simplest way to enjoy the above sufficient decrease. In practice, however, it may be difficult to estimate the constant L. In such a case, one may repeat decreasing t until the sufficient decrease is satisfied, i.e.,

$$f(x_k - t\nabla f(x_k)) \leq f(x_k) - \frac{t}{2}\|\nabla f(x_k)\|_2^2.$$

More generally, one may use the following backtracking line search method to achieve a slightly different form of sufficient decrease.

Backtracking line search

- require $\bar{t} > 0$, $\alpha \in (0,1)$, and $\beta \in (0,1)$,
- set $t_k \leftarrow \bar{t}$,
- repeat $t_k \leftarrow \beta t_k$ until $f(x_k - t_k \nabla f(x_k)) \le f(x_k) - \alpha t_k \|\nabla f(x_k)\|_2^2$.

1.2 Finding a stationary point under smoothness

Gradient descent is all about local movement hoping to decrease the value of f at every iteration based on the first-order oracle. However, this first-order information reveals nothing about the optimal value or global minimizers. Even if we have sufficient decrease under smoothness as earlier, gradient descent may not reach the optimal value f^* without further assumptions on f; all we can do with gradient descent is to keep moving as long as it is not a stationary point.

It turns out, however, that sufficient decrease allows us to find a stationary point under smoothness. More concretely, if f is L-smooth, we can quantify the smallest possible norm of the gradient $\|\nabla f\|_2$ after N iterations as $O(1/\sqrt{N})$.

Proposition 1. *Suppose $f\colon \mathbb{R}^d \to \mathbb{R}$ is L-smooth and $f^* = \inf_{x\in\mathbb{R}^d} f(x) > -\infty$. Then, gradient descent with constant step size $t \in (0, 1/L]$ yields*

$$\min_{0\le k\le N} \|\nabla f(x_k)\|_2 \le \sqrt{\frac{2(f(x_0) - f^*)}{t(N+1)}}.$$

Proof. From (3) and $x_{k+1} = x_k - t\nabla f(x_k)$, we have

$$\frac{t}{2}\|\nabla f(x_k)\|_2^2 \le f(x_k) - f(x_{k+1}),$$

which verifies that gradient descent is a descent method.

$$\min_{0\le k\le N} \|\nabla f(x_k)\|_2^2 \le \frac{1}{N+1}\sum_{k=0}^{N} \|\nabla f(x_k)\|_2^2 \le \frac{2(f(x_0) - f(x_{N+1}))}{t(N+1)} \le \frac{2(f(x_0) - f^*)}{t(N+1)}.$$

□

Remark 3. Another way to write Proposition 1 is as follows: one can find an ε-stationary point of f, i.e., a point $x \in \mathbb{R}^d$ such that $\|\nabla f(x)\|_2 \le \varepsilon$, in $O(\varepsilon^{-2})$ iterations. To see this, observe that $\min_{0\le k\le N} \|\nabla f(x_k)\|_2 \le \varepsilon$ holds provided

$$N + 1 \ge \frac{2(f(x_0) - f^*)}{t\varepsilon^2}.$$

A caveat here is that one needs to keep track of the values $\|\nabla f(x_k)\|_2$ for all $k \ge 0$ and find the smallest one.

Remark 4 (Backtracking line search)**.** In Proposition 1, suppose we use backtracking line search instead. To simplify the analysis, let $\alpha = \frac{1}{2}$. Then, for any $k \geq 0$, we still have

$$\frac{t_k}{2}\|\nabla f(x_k)\|_2^2 \leq f(x_k) - f(x_{k+1}),$$

which leads to

$$\min_{0\leq k\leq N} \|\nabla f(x_k)\|_2^2 \leq \frac{1}{N+1}\sum_{k=0}^{N} \|\nabla f(x_k)\|_2^2 \leq \frac{2}{N+1}\sum_{k=0}^{N} \frac{f(x_k) - f(x_{k+1})}{t_k}.$$

Moreover, by (3), one can deduce that $t_k = \bar{t}$ or $\frac{t_k}{\beta} > \frac{1}{L}$, which implies $t_k \geq t_* := \bar{t} \wedge \frac{\beta}{L}$ for all $k \geq 0$. Therefore,

$$\min_{0\leq k\leq N} \|\nabla f(x_k)\|_2^2 \leq \frac{2(f(x_0) - f^*)}{t_*(N+1)}.$$

Hence, we reach the same conclusion with the backtracking line search method. See also Theorem 3.2 of [NW06].

Note that Proposition 1 says nothing about the convergence of $f(x_k)$ or x_k; we have solely relied on smoothness on top of the assumption that $f^* > -\infty$ to find a stationary point. To find the minimum f^*, we will impose convexity, under which any stationary point is a global minimizer.

2 Gradient Descent under Smoothness and Convexity

We keep analyzing the smooth unconstrained case: $C = \mathbb{R}^d$ and $f\colon \mathbb{R}^d \to \mathbb{R}$ is L-smooth. We impose an additional assumption that f is convex and study how to approximate the minimum $f^* = \inf_{x\in\mathbb{R}^d} f(x) > -\infty$ using gradient descent.

2.1 Finding the minimum under convexity

If f is convex, any stationary point is a global minimizer. Though we have shown in Proposition 1 that gradient descent can find a stationary point under smoothness, that result does not explicitly quantify the gap between $f(x_k)$ and the minimum f^*. To find the minimum, we need a related yet different approach to analyze gradient descent. Using the additional convexity assumption, it turns out that we can explicitly quantify the gap between the value of f after N iterations and the minimum f^* as $O(1/N)$.

Proposition 2. *Suppose $f\colon \mathbb{R}^d \to \mathbb{R}$ is L-smooth and convex, and $f^* = f(x^*)$ for some $x^* \in \mathbb{R}^d$. Then, gradient descent with constant step size $t \in (0, 1/L]$ yields*

$$f(x_N) - f^* \leq \frac{\|x_0 - x^*\|_2^2}{2tN}.$$

Proof. To derive an upper bound on the gap $f(x_{k+1}) - f^*$, let us decompose the gap as follows:

$$f(x_{k+1}) - f^* = \underbrace{f(x_{k+1}) - f(x_k)}_{\text{sufficient decrease (3)}} + \underbrace{f(x_k) - f^*}_{\text{convexity of } f},$$

where the two terms on the right-hand side can be further bounded by the sufficient decrease (3) and convexity of f, namely, we can upper bound $f(x_k) - f^*$ by the linear approximation $\langle \nabla f(x_k), x_k - x^*\rangle$. Accordingly, we have

$$f(x_{k+1}) - f^* \leq \langle \nabla f(x_k), x_k - x^*\rangle - \frac{t}{2}\|\nabla f(x_k)\|_2^2.$$

Using $2\langle a, b\rangle - \|a\|_2^2 = \|b\|_2^2 - \|b-a\|_2^2$, note that

$$\begin{aligned}\langle \nabla f(x_k), x_k - x^*\rangle - \frac{t}{2}\|\nabla f(x_k)\|_2^2 &= \frac{2\langle x_k - x_{k+1}, x_k - x^*\rangle - \|x_k - x_{k+1}\|_2^2}{2t} \\ &= \frac{\|x_k - x^*\|_2^2 - \|x_{k+1} - x^*\|_2^2}{2t}.\end{aligned} \tag{4}$$

Therefore, we have the following bound via a telescoping term:

$$f(x_{k+1}) - f^* \leq \frac{\|x_k - x^*\|_2^2 - \|x_{k+1} - x^*\|_2^2}{2t}. \tag{5}$$

As $f(x_{k+1}) \leq f(x_k)$, we have

$$f(x_N) - f^* \leq \frac{1}{N}\sum_{k=0}^{N-1}(f(x_{k+1}) - f^*) \leq \frac{\|x_0 - x^*\|_2^2 - \|x_N - x^*\|_2^2}{2tN} \leq \frac{\|x_0 - x^*\|_2^2}{2tN}.$$

$\square$

Remark 5. Another way to write Proposition 2 is as follows: one can find an ε-suboptimal point of f, i.e., a point $x \in \mathbb{R}^d$ such that $f(x) - f^* \le \varepsilon$, in $O(\varepsilon^{-1})$ iterations.

Remark 6. In the proof of Proposition 2, convexity is used to derive the following upper bound on the difference $f(x_k) - f^*$, namely, the primal error at x_k:

$$f(x_k) - f^* \le \langle \nabla f(x_k), x_k - x^* \rangle.$$

In words, the primal error at x_k is at most $\ell_k(x_k) - \ell_k(x^*)$, where $\ell_k(x) := \langle \nabla f(x_k), x \rangle$ is the local linear approximation of f at x_k. This essentially means that for a convex function f, we may analyze $f(x_k) - f^*$ for the worst case as if f was a linear function ℓ_k instead. Then, for such a linear function, one can obtain (4) by definition, which leads to the telescoping bound (5).

Remark 7 (Backtracking line search)**.** In Proposition 2, suppose we use backtracking line search instead. To simplify the analysis, let $\alpha = \frac{1}{2}$. As in Remark 4, one can still derive (5) with t replaced by t_k, which leads to

$$f(x_N) - f^* \le \frac{1}{N} \sum_{k=0}^{N-1} (f(x_{k+1}) - f^*) \le \frac{1}{N} \sum_{k=0}^{N-1} \frac{\|x_k - x^*\|_2^2 - \|x_{k+1} - x^*\|_2^2}{2t_k}.$$

As in Remark 4, recall that $t_k \ge t_* := \bar{t} \wedge \frac{\beta}{L}$ for all $k \ge 0$. Therefore,

$$f(x_N) - f^* \le \frac{\|x_0 - x^*\|_2^2}{2t_* N}.$$

Hence, we reach the same conclusion with the backtracking line search method.

Remark 8 (Stationary point)**.** Before stating Proposition 2, we have briefly mentioned that Proposition 1, i.e., small $\|\nabla f(x_k)\|_2$, does not directly translate into small $f(x_k) - f^*$. To this end, we need to employ the convexity again. In the proof of Proposition 2, for any $m < N$, applying (3) yields

$$f(x_m) - f(x_N) = \sum_{k=m}^{N-1} (f(x_k) - f(x_{k+1})) \ge \frac{t}{2} \sum_{k=m}^{N-1} \|\nabla f(x_k)\|_2^2.$$

Therefore,

$$\min_{0 \le k \le N} \|\nabla f(x_k)\|_2^2 \le \frac{1}{N-m} \sum_{k=m}^{N-1} \|\nabla f(x_k)\|_2^2 \le \frac{2(f(x_m) - f(x_N))}{t(N-m)}.$$

By Proposition 2,

$$f(x_m) - f(x_N) \le f(x_m) - f^* \le \frac{\|x_0 - x^*\|_2^2}{2tm}.$$

Therefore,

$$\min_{0 \le k \le N} \|\nabla f(x_k)\|_2^2 \le \frac{\|x_0 - x^*\|_2^2}{t^2 (N-m) m}.$$

By letting $m = \lfloor N/2 \rfloor$, we conclude that

$$\min_{0 \le k \le N} \|\nabla f(x_k)\|_2 \le O(1/N),$$

meaning that we can find an ε-stationary point in $O(\varepsilon^{-1})$ iterations. Compare this with Proposition 1; convexity provides a faster convergence of $\min_{0\le k\le N} \|\nabla f(x_k)\|_2$. The above argument is based on [Nes12].

Remark 9. It is important to remember that Proposition 2 does not guarantee the convergence of x_k to x^*. Nevertheless, we can show that $\|x_k - x^*\|_2$ decreases for any minimizer $x^* \in \mathrm{Opt}_f := \{x \in \mathbb{R}^d : f(x) = f^*\}$, where Opt_f denotes the set of all minimizers; in fact, we have already shown this by (5), which is true for $t \in (0, 1/L]$. More generally, we can prove this for any $t \in (0, 2/L)$ by using the co-coercivity of ∇f implied by convexity and L-smoothness of f; see Lemma 2 below. To see this, invoking (4) again (notice that (4) is true for any $t > 0$), provided $t \in (0, 2/L)$, we have

$$\begin{aligned} \|x_k - x^*\|_2^2 - \|x_{k+1} - x^*\|_2^2 &= 2t\langle \nabla f(x_k), x_k - x^*\rangle - t^2\|\nabla f(x_k)\|_2^2 \\ &\ge t\left(\frac{2}{L} - t\right)\|\nabla f(x_k)\|_2^2 \\ &\ge 0, \end{aligned} \tag{6}$$

which shows that $\|x_k - x^*\|_2$ decreases. As this is true for any minimizer x^*, one can deduce that the distance from x_k to the set of all minimizers Opt_f must decrease as well; in other words, the distance $\inf_{x\in \mathrm{Opt}_f} \|x_k - x\|_2$ decreases.

Lemma 2. *Let $f\colon \mathbb{R}^d \to \mathbb{R}$ be a convex function and $L > 0$. Then, the following are equivalent.*

(i) f is L-smooth.

(ii) f satisfies

$$|f(y) - f(x) - \langle \nabla f(x), y - x\rangle| \le \frac{L}{2}\|y - x\|_2^2 \quad \forall x, y \in \mathbb{R}^d.$$

(iii) ∇f is co-coercive, namely,

$$\langle \nabla f(x) - \nabla f(y), x - y\rangle \ge \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|_2^2 \quad \forall x, y \in \mathbb{R}^d. \tag{7}$$

Proof. (i) implies (ii) by Lemma 1. (iii) implies (i) by the Cauchy-Schwarz inequality. For these two directions, convexity of f is not used. We show that (ii) implies (iii); here, convexity is necessary. To this end, fix $x \in \mathbb{R}^d$. Define $g(z) = f(z) - \langle \nabla f(x), z\rangle$ for all $z \in \mathbb{R}^d$; then, g is convex, and x is a minimizer of g as $\nabla g(x) = 0$. Also, from (ii), one can verify that g satisfies

$$|g(z) - g(y) - \langle \nabla g(y), z - y\rangle| \le \frac{L}{2}\|z - y\|_2^2 \quad \forall z, y \in \mathbb{R}^d.$$

Therefore,

$$g(x) = \inf_{z\in\mathbb{R}^d} g(z) \le \inf_{z\in\mathbb{R}^d} \underbrace{\left(g(y) + \langle \nabla g(y), z-y\rangle + \frac{L}{2}\|z-y\|_2^2\right)}_{=:Q(z)},$$

where the quadratic function Q is minimized by $z = y - \frac{\nabla g(y)}{L}$. Hence,

$$g(x) \le \inf_{z\in\mathbb{R}^d} Q(z) = g(y) - \frac{\|\nabla g(y)\|_2^2}{2L}.$$

Accordingly,

$$f(y) - f(x) - \langle \nabla f(x), y-x\rangle = g(y) - g(x) \ge \frac{\|\nabla g(y)\|_2^2}{2L} = \frac{\|\nabla f(y) - \nabla f(x)\|_2^2}{2L}.$$

Reversing the role of x and y,

$$f(x) - f(y) - \langle \nabla f(y), x-y\rangle \ge \frac{\|\nabla f(x) - \nabla f(y)\|_2^2}{2L}.$$

Combining the above inequalities, we have (7). □

2.2 Convergence to the minimizer under strong convexity

As noted in Remark 9, Proposition 2 does not guarantee the convergence of x_k to x^*. The main catch is that (6) is insufficient to prove such a result. In order to prove convergence, we want to modify (6) as follows: there exists a constant $\gamma \in (0,1)$ such that

$$\|x_k - x^*\|_2^2 - \|x_{k+1} - x^*\|_2^2 = 2t\langle \nabla f(x_k), x_k - x^*\rangle - t^2\|\nabla f(x_k)\|_2^2 \overset{\text{want}}{\ge} \gamma\|x_k - x^*\|_2^2.$$

This essentially means that we want a lower bound on $\langle \nabla f(x_k), x_k - x^*\rangle$ that involves the term $\|x_k - x^*\|_2^2$. It turns out that such a lower bound is obtainable if f is strongly convex.

We first recall the definition of strong convexity and show that strong convexity and smoothness together imply a stronger version of co-coercivity than (7), which is the key to proving the convergence of x_k to x^*.

Definition 2. A function $f\colon \mathbb{R}^d \to (-\infty,\infty]$ is μ-strongly convex for some $\mu > 0$ if $f - \frac{\mu}{2}\|\cdot\|_2^2$ is convex, or equivalently,

$$f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda) f(y) - \frac{\mu}{2}\lambda(1-\lambda)\|x-y\|_2^2 \quad \forall x,y\in\mathbb{R}^d, \lambda\in[0,1].$$

If f is μ-strongly convex and is differentiable, then

$$f(y) \ge f(x) + \langle \nabla f(x), y-x\rangle + \frac{\mu}{2}\|y-x\|_2^2 \quad \forall x,y\in\mathbb{R}^d.$$

Lemma 3. *If $f\colon \mathbb{R}^d \to \mathbb{R}$ is L-smooth and μ-strongly convex, then $\mu \le L$ must hold and*

$$\langle \nabla f(x) - \nabla f(y), x-y\rangle \ge \frac{\mu L}{\mu+L}\|x-y\|_2^2 + \frac{1}{\mu+L}\|\nabla f(x) - \nabla f(y)\|_2^2 \quad \forall x,y\in\mathbb{R}^d. \quad (8)$$

Proof. Let $g(x) = f(x) - \frac{\mu}{2}\|x\|_2^2$ for all $x \in \mathbb{R}^d$; then, $\nabla g(x) = \nabla f(x) - \mu x$. The μ-strong convexity of f implies that g is convex. Combined with the L-smoothness of f, we have for any $x, y \in \mathbb{R}^d$,

$$0 \le \langle \nabla g(x) - \nabla g(y), x - y \rangle = \langle \nabla f(x) - \nabla f(y), x - y \rangle - \mu\|x-y\|_2^2 \le (L-\mu)\|x-y\|_2^2.$$

Hence, we have $\mu \le L$. If $\mu = L$, we must have $\langle \nabla f(x) - \nabla f(y), x - y \rangle = \mu\|x-y\|_2^2$ for all $x, y \in \mathbb{R}^d$, which, combined with (7) with $L = \mu$, yield (8). Consider the case $\mu < L$. From $\langle \nabla g(x) - \nabla g(y), x - y \rangle \le (L-\mu)\|x-y\|_2^2$, one can deduce, by mimicking the proof of Lemma 3, that

$$|g(y) - g(x) - \langle \nabla g(x), y - x \rangle| \le \frac{L-\mu}{2}\|y-x\|_2^2 \quad \forall x, y \in \mathbb{R}^d,$$

which, together with the convexity of g, implies that ∇g is co-coercive by Lemma 2, namely,

$$\begin{aligned} \frac{1}{L-\mu}\|\nabla g(x) - \nabla g(y)\|_2^2 &\le \langle \nabla g(x) - \nabla g(y), x - y \rangle \\ &= \langle \nabla f(x) - \nabla f(y), x - y \rangle - \mu\|x-y\|_2^2. \end{aligned}$$

Plugging in $\nabla g(x) - \nabla g(y) = \nabla f(x) - \nabla f(y) - \mu(x-y)$, we have (8). $\square$

Now, we show that gradient descent converges linearly for strongly convex and smooth functions. Here, linear convergence means that the gap decreases by a constant factor in each iteration, leading to an exponential decrease of the gap as the number of iterations increases.

Proposition 3. *Suppose $f\colon \mathbb{R}^d \to \mathbb{R}$ is L-smooth and μ-strongly convex, and $f^* = f(x^*)$ for some $x^* \in \mathbb{R}^d$ so that x^* is the unique minimizer of f. Then, gradient descent with constant step size $t \in (0, 2/(\mu+L)]$ yields*

$$\|x_N - x^*\|_2^2 \le \left(1 - \frac{2t\mu L}{\mu + L}\right)^N \|x_0 - x^*\|_2^2$$

and

$$f(x_N) - f^* \le \frac{L}{2}\left(1 - \frac{2t\mu L}{\mu + L}\right)^N \|x_0 - x^*\|_2^2,$$

meaning that one can find an ε-suboptimal point in $O(\log(1/\varepsilon))$ iterations. Particularly, if we choose $t = 2/(\mu+L)$, then

$$\|x_N - x^*\|_2^2 \le \left(\frac{\kappa-1}{\kappa+1}\right)^N \|x_0 - x^*\|_2^2$$

and

$$f(x_N) - f^* \le \frac{L}{2}\left(\frac{\kappa-1}{\kappa+1}\right)^N \|x_0 - x^*\|_2^2,$$

where $\kappa := L/\mu$ is often called the condition number of f.

Proof. Now, provided $t \in (0, 2/(\mu + L))$,

$$\begin{aligned}
\|x_k - x^*\|_2^2 - \|x_{k+1} - x^*\|_2^2 &= 2t\langle \nabla f(x_k), x_k - x^* \rangle - t^2 \|\nabla f(x_k)\|_2^2 \\
&\geq \frac{2t\mu L}{\mu + L}\|x_k - x^*\|_2^2 + t\left(\frac{2}{\mu + L} - t\right)\|\nabla f(x_k)\|_2^2 \\
&\geq \frac{2t\mu L}{\mu + L}\|x_k - x^*\|_2^2,
\end{aligned}$$

where the equality is (4) and the first inequality follows from (8) with $\nabla f(x^*) = 0$. Hence,

$$\|x_N - x^*\|_2^2 \leq \left(1 - \frac{2t\mu L}{\mu + L}\right)^N \|x_0 - x^*\|_2^2.$$

Also, we have

$$f(x_N) - f^* \leq \frac{L}{2}\|x_N - x^*\|_2^2 \leq \frac{L}{2}\left(1 - \frac{2t\mu L}{\mu + L}\right)^N \|x_0 - x^*\|_2^2,$$

where the first inequality follows from Lemma 1 with $\nabla f(x^*) = 0$. □

2.3 Convergence under the Polyak-Łojasiewicz condition

In many optimization problems, strong convexity may not hold. In the literature, there has been a line of research that studies the convergence of gradient descent under weaker conditions than strong convexity. One such condition is the Polyak-Łojasiewicz (PL) condition, which is strictly weaker than strong convexity but still guarantees linear convergence of gradient descent.

Definition 3. A function $f\colon \mathbb{R}^d \to \mathbb{R}$ satisfies the Polyak-Łojasiewicz (PL) condition with parameter $\mu > 0$ if

$$\frac{1}{2}\|\nabla f(x)\|_2^2 \geq \mu(f(x) - f^*) \quad \forall x \in \mathbb{R}^d, \tag{9}$$

where $f^* = \inf_{x \in \mathbb{R}^d} f(x)$.

Lemma 4. *If $f\colon \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex and differentiable, then f satisfies the PL condition with parameter μ.*

Proof. From

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|_2^2 \quad \forall x, y \in \mathbb{R}^d,$$

let us minimize the right-hand side over $y \in \mathbb{R}^d$ while keeping x fixed, namely,

$$\min_{y \in \mathbb{R}^d}\left(f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|_2^2\right) = f(x) - \frac{\|\nabla f(x)\|_2^2}{2\mu},$$

where the minimum is attained at $y = x - \frac{\nabla f(x)}{\mu}$. Hence,

$$f(y) \geq f(x) - \frac{\|\nabla f(x)\|_2^2}{2\mu} \quad \forall y \in \mathbb{R}^d,$$

which leads to (9). □

From (9), observe that the PL condition implies that any stationary point of f must be a global minimizer. Unlike strong convexity, the PL condition does not require the uniqueness of the minimizer; in fact, there can be multiple minimizers of f under the PL condition. Also, the PL condition does not imply the convexity of f; in fact, there exist non-convex functions that satisfy the PL condition.

Now, we show that gradient descent converges linearly under the PL condition. The key is to combine (3) and (9) to obtain a linear decrease of the gap $f(x_k) - f^*$.

Proposition 4. *Suppose $f\colon \mathbb{R}^d \to \mathbb{R}$ is L-smooth and satisfies the PL condition with parameter $\mu > 0$. Then, $\mu \le L$ must hold, and gradient descent with constant step size $t \in (0, 1/L]$ yields*

$$f(x_N) - f^* \le (1 - \mu t)^N (f(x_0) - f^*),$$

meaning that one can find an ε-suboptimal point in $O(\log(1/\varepsilon))$ iterations.

Proof. To see $\mu \le L$, recall from (3) that

$$f(x - \nabla f(x)/L) \le f(x) - \frac{1}{2L}\|\nabla f(x)\|_2^2 \quad \forall x \in \mathbb{R}^d.$$

Hence,

$$f^* \le f(x) - \frac{1}{2L}\|\nabla f(x)\|_2^2 \quad \forall x \in \mathbb{R}^d,$$

which, combined with (9), implies $\mu \le L$. Now, with $t \in (0, 1/L]$, we have for any $k \ge 0$,

$$f(x_{k+1}) - f^* \le f(x_k) - f^* - \frac{t}{2}\|\nabla f(x_k)\|_2^2 \le (1 - \mu t)(f(x_k) - f^*),$$

where the first inequality follows from (3) and the second inequality follows from (9). Hence,

$$f(x_N) - f^* \le (1 - \mu t)^N (f(x_0) - f^*).$$

□

3 Projection for Constrained Convex Optimization

In this section, we assume $C \subset \mathbb{R}^d$ is closed and convex. Also, we assume f is L-smooth and convex on C; here, L-smoothness on C means that ∇f is well-defined on C and is L-Lipschitz. In this setting, the gradient descent update rule may produce a point outside of C, namely, one may encounter the situation where $x - t\nabla f(x) \notin C$ for $x \in C$ and $t > 0$. A simple remedy for this situation is to project the point $x - t\nabla f(x)$ back to the set C. It turns out that such a projection is well-defined as long as C is closed and convex.

Definition 4. Let $C \subset \mathbb{R}^d$ be a closed convex set. For any $x \in \mathbb{R}^d$, define

$$P_C(x) := \arg\min_{z \in C} \|x - z\|_2.$$

We call $P_C \colon \mathbb{R}^d \to C$ the projection operator onto C.

We show that P_C is well-defined. Pick any element $w \in C$ and let $r := \|w - x\|_2$. If $r = 0$, then $P_C(x)$ must be w. Otherwise, let $B_r(x) := \{z \in \mathbb{R}^d : \|z - x\|_2 \leq r\}$. Then, minimizing $g(z) := \|z - x\|_2$ over C is equivalent to minimizing g over $C \cap B_r(x)$, namely,

$$\min_{z \in C} \|z - x\|_2 = \min_{z \in C \cap B_r(x)} \|z - x\|_2.$$

Since $C \cap B_r(x)$ is compact and g is continuous, g admits a minimizer on $C \cap B_r(x)$, which must be a minimizer of g on C. This shows the existence of minimizers of g on C. We show that there can be only one minimizer. Suppose $z_1, z_2 \in C$ are minimizers of g on C, namely,

$$\|z_1 - x\|_2 = \|z_2 - x\|_2 = \min_{z \in C} \|z - x\|_2 =: \delta.$$

Then, by the parallelogram law,

$$\frac{\|z_1 - z_2\|_2^2}{4} = \frac{\|z_1 - x\|_2^2 + \|z_2 - x\|_2^2}{2} - \left\|\frac{z_1 + z_2}{2} - x\right\|_2^2 = \delta^2 - \left\|\frac{z_1 + z_2}{2} - x\right\|_2^2 \leq 0,$$

where the last inequality follows as $\frac{z_1+z_2}{2} \in C$. Hence, $z_1 = z_2$, which shows the uniqueness of the minimizer.

Lemma 5. *Let $C \subset \mathbb{R}^d$ be a closed convex set. Then, for any $x \in \mathbb{R}^d$,*

$$\langle P_C(x) - x, z - P_C(x) \rangle \geq 0 \quad \forall z \in C.$$

Proof. Let $f(z) = \frac{1}{2}\|z - x\|_2^2$ for any $z \in \mathbb{R}^d$. Now, fix $z \in C$. For $t \in [0, 1]$, define $h(t) = f(P_C(x) + t(z - P_C(x)))$. Then, by definition, $h(t) \geq h(0)$ for all $t \in [0, 1]$. Hence,

$$0 \leq \lim_{t \to 0} \frac{h(t) - h(0)}{t} = h'(0) = \langle P_C(x) - x, z - P_C(x) \rangle.$$

$\square$

Using the projection operator P_C, one may attempt the following projected gradient descent: given an initial point $x_0 \in C$, for $k \geq 0$, iterate

$$x_{k+1} \leftarrow P_C(x_k - t_k \nabla f(x_k)) \quad \text{for some suitable } t_k > 0.$$

Of course, in order for projected gradient descent to be practical, the projection operator P_C should easily be computable; for instance, C is a Euclidean ball.

Using Lemma 5, we can verify that projected gradient descent is a descent method provided $t \leq \frac{2}{L}$. To this end, use Lemma 1 to derive

$$f(x_{k+1}) - f(x_k) \leq \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2} \|x_{k+1} - x_k\|_2^2.$$

Since

$$t \nabla f(x_k) = x_k - (x_k - t \nabla f(x_k)) = (x_k - x_{k+1}) + (x_{k+1} - (x_k - t \nabla f(x_k))), \tag{10}$$

we have

$$\begin{aligned} \langle \nabla f(x_k), x_{k+1} - x_k \rangle &= -\frac{\|x_{k+1} - x_k\|_2^2}{t} + \frac{\langle (x_{k+1} - t \nabla f(x_k), x_{k+1} - x_k) \rangle}{t} \\ &\leq -\frac{\|x_{k+1} - x_k\|_2^2}{t}, \end{aligned}$$

where the inequality follows from Lemma 5. Therefore,

$$f(x_{k+1}) - f(x_k) \leq -\left(\frac{1}{t} - \frac{L}{2} \right) \|x_{k+1} - x_k\|_2^2, \tag{11}$$

which verifies that projected gradient descent is indeed a descent method provided $t \leq \frac{2}{L}$.

Proposition 5. *Suppose $f \colon \mathbb{R}^d \to \mathbb{R}$ is convex and L-smooth on a closed convex set $C \subset \mathbb{R}^d$; also, assume $f^* := \inf_{x \in C} f(x) = f(x^*)$ for some $x^* \in C$. Then, projected gradient descent with constant step size $t \in (0, 1/L]$ yields*

$$f(x_N) - f^* \leq \frac{\|x_0 - x^*\|_2^2}{2tN}.$$

Proof. Using L-smoothness (Lemma 1) and convexity,

$$\begin{aligned} f(x_{k+1}) - f^* &= f(x_{k+1}) - f(x_k) + f(x_k) - f^* \\ &\leq \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2} \|x_{k+1} - x_k\|_2^2 + \langle \nabla f(x_k), x_k - x^* \rangle \\ &= \langle \nabla f(x_k), x_{k+1} - x^* \rangle + \frac{L}{2} \|x_{k+1} - x_k\|_2^2. \end{aligned}$$

Using (10), we have

$$\begin{aligned} \langle \nabla f(x_k), x_{k+1} - x^* \rangle &= \frac{\langle x_k - x_{k+1}, x_{k+1} - x^* \rangle + \langle x_{k+1} - (x_k - t \nabla f(x_k)), x_{k+1} - x^* \rangle}{t} \\ &\leq \frac{\langle x_k - x_{k+1}, x_{k+1} - x^* \rangle}{t}, \end{aligned} \tag{12}$$

where the inequality is due to Lemma 5 (recall that $x^* \in C$). Therefore, we have

$$f(x_{k+1}) - f^* \leq \frac{\langle x_k - x_{k+1}, x_{k+1} - x^* \rangle}{t} + \frac{L}{2}\|x_{k+1} - x_k\|_2^2. \tag{13}$$

Using $t \leq \frac{1}{L}$, we have

$$\begin{aligned} f(x_{k+1}) - f^* &\leq \frac{2\langle x_k - x_{k+1}, x_{k+1} - x^* \rangle + \|x_{k+1} - x_k\|_2^2}{2t} \\ &= \frac{\|x_k - x^*\|_2^2 - \|x_{k+1} - x^*\|_2^2}{2t}, \end{aligned} \tag{14}$$

where the last equality uses $2\langle a, b\rangle + \|a\|_2^2 = \|a + b\|_2^2 - \|b\|_2^2$. Therefore,

$$f(x_N) - f^* \leq \frac{1}{N}\sum_{k=0}^{N-1}(f(x_{k+1}) - f^*) \leq \frac{\|x_0 - x^*\|_2^2}{2tN}.$$

□

Remark 10. The implication of Proposition 5 is that projected gradient descent yields the result essentially the same as the unconstrained case (cf. Proposition 2) as long as the L-smoothness and convexity are satisfied on the closed convex set C. Though projected gradient descent is as good as gradient descent for the unconstrained case in theory, one should keep in mind that projected gradient descent is practical only when P_C is easily computable.

Remark 11. The previous results (11) and (13) follow from the following general result: for any $z \in C$,

$$f(x_{k+1}) \leq f(z) + \frac{\langle x_k - x_{k+1}, x_{k+1} - z \rangle}{t} + \frac{L}{2}\|x_{k+1} - x_k\|_2^2. \tag{15}$$

Clearly, one can derive (11) by letting $z = x_k$. To derive (13), let $z = x^*$ and use the last two equalities of (12). To derive (15), observe that

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|_2^2 \\ &\leq f(z) + \langle \nabla f(x_k), x_k - z \rangle + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|_2^2 \\ &= f(z) + \langle \nabla f(x_k), x_{k+1} - z \rangle + \frac{L}{2}\|x_{k+1} - x_k\|_2^2, \end{aligned}$$

where the inequalities are due to L-smoothness and convexity. Hence, to derive (15), it suffices to show

$$\langle \nabla f(x_k), x_{k+1} - z \rangle \leq \frac{\langle x_k - x_{k+1}, x_{k+1} - z \rangle}{t}. \tag{16}$$

One can prove (16) by means of Lemma 5; simply mimic (12) with z instead of x^*.

Remark 12. As in Remark 9, note that (14) implies that the distance from x_k to the set of all minimizers $\mathrm{Opt}_f := \{x \in C : f(x) = f^*\}$ must decrease; in other words, the distance $\inf_{x \in \mathrm{Opt}_f} \|x_k - x\|_2$ decreases.

Proposition 6. *Suppose $f\colon \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex and L-smooth on a closed convex set $C \subset \mathbb{R}^d$; also, assume $f^* := \inf_{x\in C} f(x) = f(x^*)$ for some $x^* \in C$ so that x^* is the unique minimizer of f on C. Then, projected gradient descent with constant step size $t \in (0, 1/L]$ yields*

$$\|x_N - x^*\|_2^2 \le (1-\mu t)^N \|x_0 - x^*\|_2^2$$

and

$$f(x_N) - f^* \le \frac{(1-\mu t)^N}{2t}\|x_0 - x^*\|_2^2.$$

Proof. Using L-smoothness (Lemma 1) and μ-strong convexity,

$$\begin{aligned}
f(x_{k+1}) - f^* &= f(x_{k+1}) - f(x_k) + f(x_k) - f^* \\
&\le \langle \nabla f(x_k), x_{k+1} - x_k\rangle + \frac{L}{2}\|x_{k+1} - x_k\|_2^2 + \langle \nabla f(x_k), x_k - x^*\rangle - \frac{\mu}{2}\|x_k - x^*\|_2^2 \\
&= \langle \nabla f(x_k), x_{k+1} - x^*\rangle + \frac{L}{2}\|x_{k+1} - x_k\|_2^2 - \frac{\mu}{2}\|x_k - x^*\|_2^2.
\end{aligned}$$

Using (12) (or (16)), for $t \le \frac{1}{L}$, we have

$$\begin{aligned}
f(x_{k+1}) - f^* &\le \frac{\langle x_k - x_{k+1}, x_{k+1} - x^*\rangle}{t} + \frac{L}{2}\|x_{k+1} - x_k\|_2^2 - \frac{\mu}{2}\|x_k - x^*\|_2^2 \\
&\le \frac{2\langle x_k - x_{k+1}, x_{k+1} - x^*\rangle + \|x_{k+1} - x_k\|_2^2}{2t} - \frac{\mu}{2}\|x_k - x^*\|_2^2 \\
&= \frac{\|x_k - x^*\|_2^2 - \|x_{k+1} - x^*\|_2^2}{2t} - \frac{\mu}{2}\|x_k - x^*\|_2^2,
\end{aligned}$$

where the last equality uses $2\langle a, b\rangle + \|a\|_2^2 = \|a+b\|_2^2 - \|b\|_2^2$. As $f(x_{k+1}) \ge f^*$, we have

$$\|x_{k+1} - x^*\|_2^2 \le (1-\mu t)\|x_k - x^*\|_2^2.$$

Therefore,

$$\|x_N - x^*\|_2^2 \le (1-\mu t)^N \|x_0 - x^*\|_2^2$$

and

$$f(x_N) - f^* \le \frac{(1-\mu t)\|x_{N-1} - x^*\|_2^2 - \|x_N - x^*\|_2^2}{2t} \le \frac{(1-\mu t)^N}{2t}\|x_0 - x^*\|_2^2.$$

□

Beyond Euclidean projection Though the projection operator P_C is well-defined for any closed convex set $C \subset \mathbb{R}^d$, it may not admit a simple closed form in general; e.g., consider $C = \Delta_d$ or $C = [0,1]^d$. To tackle this issue, one may use the Bregman projection based on the Bregman divergence which serves as a non-Euclidean distance. Or, if solving a linear problem over C is easy, one may use the Frank-Wolfe algorithm.

References

[Nes12] Yurii Nesterov. How to make the gradients small. In *Mathematical Optimization Society Newsletter Optima 88*, 2012.

[NW06] Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer, second edition, 2006.