Gradient descent

Bang-Shien Chen*

Gradient descent is widely used for solving unconstrained nonlinear optimization problems. If the problem is linear, one can simply apply linear program techniques such as simplex method. As for nonlinear programs, gradient descent tries to find a linear approximation of the nonlinear objective. While the algorithm itself is relatively simple, its convergence analysis can be quite complex. This note focuses on presenting a range of convergence results.

1 Descent Method

Consider the optimization problem:

$$\min_x f(x) \; ; \quad x \in \mathbb{R}^n. \tag{1.1}$$

We first introduce an abstract algorithm (like the abstract class while coding). Almost 80% of algorithms (according to ChatGPT) follow this structure:

$$x^{k+1} = x^k + \alpha^k d^k, \tag{1.2}$$

where d^k is a *descent direction* such that $f(x^k + \alpha^k d^k) < f(x^k)$ for some *step size* $\alpha^k > 0$. We start from some initial point x_0 and generate an iterative process

$$f(x^0) \geq f(x^1) \geq \cdots \geq f(x^k) \geq f(x^{k+1}) \geq \cdots,$$

until it attains a local minimum. This general framework raises two fundamental questions: how to find a descent direction d^k, and how to determine a suitable step size α^k? In the following sections, we will explore different strategies for choosing d^k to develop effective descent methods. Note that we assume that f has a minimum, the sequence $\{f(x^k)\}$ is bounded below.

1.1 Exact Line Search

To proceed, we first introduce some *line search* methods for searching α^k. As suggested by the name, we search for the exact α^k such that

$$\alpha^k = \arg\min_{\alpha \geq 0} f(x^k + \alpha d^k). \tag{1.3}$$

However, solving this one-dimensional optimization problem at each iteration can be computationally expensive.

1.2 Backtracking Line Search (Armijo Rule)

Instead of solving for the exact α^k, we could use heuristic methods to find an α^k that is sufficiently good without excessive computation. Let $\alpha^k = \beta^m s$, the *Armijo rule* provides a sufficient decrease in the objective function for α^k:

$$f(x^k) - f(x^k + \beta^m s d^k) \geq -\sigma \beta^m s \nabla f(x^k)^\top d^k, \tag{1.4}$$

*https://dgbshien.com/

where $\beta \in (0,1)$ and $\sigma \in (0,1)$ are fixed scalars, m is a non-negative integer. This condition guarantees that the function decrease is at least proportional to the linear approximation given by the gradient. *Backtracking line search* is an inexact method based on the Armijo rule. Think of s as an initial step size, given fixed β and σ, we try $m = 0, 1, \ldots$, i.e., shrink s by multiplying β, until the decrease is sufficiently large, i.e., the Armijo rule (1.4) is satisfied.

Algorithm 1 Backtracking Line Search

Input: initial step size s, scalars $\beta \in (0,1)$ and $\sigma \in (0,1)$
Output: suitable step size α^k that decreases the function value
1: **if** Equation (1.4) holds **then**
2: **return** $\alpha^k = \beta s$
3: **end if**
4: $s \leftarrow \beta s$, go back to line 1

Other inexact line search methods include: *constant step size*, where $\alpha^k = c$ is a constant for all k, and *diminishing step size*, where $\alpha^k \to 0$ and $\sum_{k=1}^{\infty} \alpha^k = \infty$.

2 Gradient Descent

Suppose that f is continuously differentiable, a natural selection for the descent direction is the negative gradient $-\nabla f(x)$. This leads to the *gradient descent* method:

$$x^{k+1} = x^k - \alpha^k \nabla f(x^k). \tag{2.1}$$

In general [1], we call (1.2) the gradient descent if $\nabla f(x^k)^\top d < 0$, meaning that the angle between $-\nabla f(x^k)$ and d is less than 90 degrees. To obtain this result, let $x^{k+1} = x^k + \alpha^k d^k$, by first-order approximation, we have

$$\begin{aligned} f(x^{k+1}) &= f(x^k) + \nabla f(x^k)^\top \|x^{k+1} - x^k\| + \mathcal{O}(\|x^{k+1} - x^k\|) \\ &= f(x^k) + \alpha^k \nabla f(x^k)^\top d + \mathcal{O}(\alpha^k). \end{aligned} \tag{2.2}$$

Since $\alpha^k \nabla f(x^k)^\top d$ dominates $\mathcal{O}(\alpha^k)$, we only need the direction chosen such that

$$\alpha^k \nabla f(x^k)^\top d < 0 \implies \nabla f(x^k)^\top d < 0.$$

However, in this note, we particularly refer (2.1) as gradient descent, i.e.,

$$d^k = -\nabla f(x^k),$$

since it is less complicated (requires less assumptions during convergence analysis) and is more practical in applications[1]. A natural stopping criteria for gradient descent is when we occur at a stationary point, i.e., $\nabla f(x^k) = 0$. A question then arises: is the limit point of $\{x^k\}$ generated by a gradient descent is a stationary point?

2.1 Limit points of Gradient descent

Proposition 2.1

Let $\{x^k\}$ be a sequence generated by gradient descent and α^k is chosen by exact line search. Then every limit point of $\{x^k\}$ is a stationary point.

[1] Most machine learning literature also refer (2.1) as gradient descent.

Proof. By first-order approximation (2.2), we have

$$f(x^{k+1}) = f(x^k) - \alpha^k \|\nabla f(x^k)\|^2 + \mathcal{O}(\alpha^k).$$

We can always choose an α^k by exact line search such that $f(x^{k+1}) < f(x^k)$ if $\nabla f(x^k) \neq 0$. Otherwise, the sequence converges to a stationary point. □

Proposition 2.2

Let $\{x^k\}$ be a sequence generated by gradient descent and α^k is chosen by backtracking line search. Then every limit point of $\{x^k\}$ is a stationary point.

Proof. Let $\bar{x}$ be a limit point of $\{x^k\}$ but is not a stationary point. Since $\{f(x^k)\}$ is monotonically non-increasing and bounded below, by Monotone Convergence Theorem, $\{f(x^k)\}$ converges to some finite value. Since f is continuous, $\{f(x^k)\}$ converges to $f(\bar{x})$, i.e.,

$$f(x^k) - f(x^{k+1}) \to 0.$$

For the case of backtracking line search, by Armijo rule (1.4), we have

$$f(x^k) - f(x^{k+1}) \geq \sigma \alpha^k \|\nabla f(x^k)\|^2,$$

which implies that $\alpha^k \|\nabla f(x^k)\|^2 \to 0$. Since $\bar{x}$ is not a stationary point, we have

$$\limsup_{k\to\infty} -\|\nabla f(x^k)\|^2 < 0,$$

and thus $\alpha^k \to 0$. By taking one iteration back of the backtracking line search such that the Armijo rule (1.4) is not satisfied, i.e., let $\bar{\alpha}^k = \alpha^k/\beta$, we have

$$f(x^k) - f(x^k + \bar{\alpha}^k d^k) < \sigma \bar{\alpha}^k \|\nabla f(x^k)\|^2.$$

Divide both sides by $\bar{\alpha}^k$ and with Mean Value Theorem, we have

$$\frac{f(x^k) - f(x^k - \bar{\alpha}^k \nabla f(x^k))}{\bar{\alpha}^k} = \nabla f(x^k - \tilde{\alpha}^k \nabla f(x^k))^\top \nabla f(x^k) < \sigma \|\nabla f(x^k)\|^2,$$

for some $\tilde{\alpha}^k \in [0, \bar{\alpha}^k]$. Taking the limits as $k \to \infty$ implies $\bar{\alpha}^k \to 0$, we obtain

$$\|\nabla f(\bar{x})\|^2 \leq \sigma \|\nabla f(\bar{x})\|^2,$$

which contradicts $\sigma \in (0, 1)$. □

Definition 2.3: Lipschitz continuity

A function $f : \mathbb{R}^n \to \mathbb{R}$ is *Lipschitz continuous* if there exists $L > 0$, called the *Lipschitz constant*, such that for all $x, y \in \mathbb{R}^n$,

$$\|f(x) - f(y)\| \leq L\|x - y\|. \tag{2.3}$$

A differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is *L-smooth* if its gradient is Lipschitz continuous, i.e., for all $x, y \in \mathbb{R}^n$,

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|. \tag{2.4}$$

Lemma 2.4

If $f : \mathbb{R}^n \to \mathbb{R}$ is L-smooth, then for all $x, y \in \mathbb{R}^n$,

$$f(y) \leq f(x) + \nabla f(x)^\top (y - x) + \frac{L}{2}\|y - x\|^2. \tag{2.5}$$

Proof. Let $g(t) := f(x + t(y - x))$ with chain rule, by Fundamental Theorem of Calculus and Cauchy–Schwarz inequality, Lipschitz smooth (in the order of inequalities), we have

$$\begin{aligned}
f(y) = g(1) &= g(0) + \int_0^1 \frac{\partial}{\partial t} g(t) dt \\
&= f(x) + \int_0^1 \nabla f(x + t(y - x))^\top (y - x) dt \\
&= f(x) + \nabla f(x)^\top (y - x) + \int_0^1 (\nabla f(x + t(y - x)) - \nabla f(x))^\top (y - x) \\
&\leq f(x) + \nabla f(x)^\top (y - x) + \int_0^1 \|\nabla f(x + t(y - x)) - \nabla f(x)\| \|y - x\| \\
&\leq f(x) + \nabla f(x)^\top (y - x) + \int_0^1 Lt\|y - x\|^2 \\
&= f(x) + \nabla f(x)^\top (y - x) + \frac{L}{2}\|y - x\|^2.
\end{aligned}$$

□

If we insert gradient descent $y = x - \alpha \nabla f(x)$ into inequality (2.5), we have

$$f(x - \alpha \nabla f(x)) - f(x) \leq -\alpha \left(1 - \frac{\alpha L}{2}\right) \|\nabla f(x)\|^2. \tag{2.6}$$

Note that the Lemma 2.4 allows us to construct a convex quadratic upper bound [5] at every point if f is L-smooth. Therefore, Equation (2.5) is extremely useful in terms of majorization minimization. For instance, [1, Proposition 1.2.2, Figure 1.2.9] provides an insight of the next proposition.

Proposition 2.5

Suppose that f is L-smooth. Let $\{x^k\}$ be a sequence generated by gradient descent and α^k is chosen such that

$$\epsilon \leq \alpha^k \leq \frac{2 - \epsilon}{L}, \tag{2.7}$$

where $\epsilon \in (0, 1]$. Then every limit point of $\{x^k\}$ is a stationary point.

Proof. By L-smooth (2.6), we have

$$f(x^k) - f(x^k - \alpha^k \nabla f(x^k)) \geq \left(\underbrace{\alpha^k}_{\geq \epsilon} \left(1 - \frac{L}{2} \underbrace{\alpha^k}_{\leq (2-\epsilon)/L}\right)\right) \|\nabla f(x^k)\|^2 \geq \frac{\epsilon^2}{2} \|\nabla f(x^k)\|^2.$$

By the proof of Proposition 2.2, we have $f(x^k) - f(x^{k+1}) \to 0$, which implies $\|\nabla f(x^k)\|^2 \to 0$, that is, limit points are stationary points. □

Proposition 2.5 is also the convergence result of constant step size if we choose the constant roughly in $(0, 2/L)$, as well as diminishing step size when the step size is sufficiently small after some iterations.

3 Convergence Analysis

Another question is how fast does $\{x^k\}$ converge to $\bar{x}$? For evaluation, we use an *error function* $e : \mathbb{R}^n \to \mathbb{R}$ satisfying $e(x) \geq 0$ for all x and $e(\bar{x}) = 0$. Typical choices are $e(x) = |f(x) - f(\bar{x})|$, $e(x) = \|x - \bar{x}\|$, or $e(x) = \|x - \bar{x}\|^2$. We study the following concepts:

- *Error bound*: how far current iterate is to the optimal solution, i.e., upper bound of $e(x^k)$.
- *Iteration complexity*: the number of iterations needed to reach an error tolerance ϵ.
- *Rate of convergence*: A formal way to compare convergence speed.

Let $\{x^k\}$ be a sequence that converges to $\bar{x}$. Then the sequence has a *rate of convergence* c and *order of convergence* p if it satisfies

$$\lim_{k\to\infty} \frac{e(x^{k+1})}{e(x^k)^q} = c,$$

where $e(\cdot)$ is some error function. We say a sequence is (from slowest to fastest)

- *sublinear convergence* if $q = 1$ and $c = 1$,
- *linear convergence* if $q = 1$ and $c \in (0, 1)$,
- *superlinear convergence* if $q = 1$ and $c = 0$,
- *quadratic convergence* if $q = 2$ and any c.

The rate of convergence is, in general, harder to obtain because it involves a more formal definition and often requires stronger assumptions.

Definition 3.1: Convexity

A function $f : \mathbb{R}^n \to \mathbb{R}$ is *convex* if for all $x, y \in \mathbb{R}^n$ and $0 \leq t \leq 1$,

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y). \tag{3.1}$$

If f is differentiable, we have the *first-order convexity*: $f(y) \geq f(x) + \nabla f(x)^\top (y - x)$.

Theorem 3.2

Suppose that f is L-smooth and convex. Let $\{x^k\}$ be a sequence generated by gradient descent and α^k is chosen with constant step size satisfying $0 < \alpha \leq 1/L$. Then for all $x^* \in \arg\min f$, we have

$$f(x^k) - f^* \leq \frac{\|x^0 - x^*\|}{2\alpha k}. \tag{3.2}$$

Furthermore, for a given $\epsilon > 0$, the iteration complexity is $\mathcal{O}(1/\epsilon)$. The iterates enjoy a sublinear convergence.

Proof. By L-smooth (2.6) with $\alpha \leq 1/L$, we have

$$f(x^{k+1}) = f(x^k - \alpha\nabla f(x^k)) \leq f(x^k) - \alpha\left(1 - \frac{\alpha L}{2}\right)\|\nabla f(x^k)\|^2 \leq f(x^k) - \frac{\alpha}{2}\|\nabla f(x^k)\|^2,$$

which also implies $f(x^{k+1}) - f(x^k) \leq 0$, i.e., f is non-increasing. Since f is convex, by rearranging first-order convexity, we have

$$f^* = f(x^*) \geq f(x^k) - \nabla f(x^k)^\top (x^k - x^*).$$

Together, we obtain

$$\begin{aligned} f(x^{k+1}) - f^* &\leq \nabla f(x^k)^\top (x^k - x^*) - \frac{\alpha}{2}\|\nabla f(x^k)\|^2 \\ &= \frac{1}{2\alpha}\big(2\alpha \nabla f(x^k)^\top (x^k - x^*) - \alpha^2 \|\nabla f(x^k)\|^2 + \|x^k - x^*\|^2 - \|x^k - x^*\|^2\big) \\ &= \frac{1}{2\alpha}\big(\|x^k - x^*\|^2 - \|x^k - \alpha \nabla f(x^k) - x^*\|^2\big) \\ &= \frac{1}{2\alpha}\big(\|x^k - x^*\|^2 - \|x^{k+1} - x^*\|^2\big). \end{aligned}$$

Now take the summation over iterations, with the inequality we obtained, we have

$$\begin{aligned} \sum_{i=1}^{k} (f(x^i) - f^*) &\leq \frac{1}{2\alpha} \sum_{i=1}^{k} \big(\|x^{i-1} - x^*\|^2 - \|x^i - x^*\|^2\big) \\ &= \frac{1}{2\alpha}\|x^0 - x^*\|^2 - \|x^k - x^*\|^2 \\ &\leq \frac{1}{2\alpha}\|x^0 - x^*\|^2. \end{aligned}$$

Since f is non-increasing, we have

$$f(x^k) - f^* \leq \frac{1}{k}\sum_{i=1}^{k} (f(x^i) - f^*) \leq \frac{1}{2\alpha k}\|x^0 - x^*\|^2.$$

For a given $\epsilon > 0$, we have

$$k \geq \frac{\|x^0 - x^*\|^2}{2\alpha\epsilon} \implies \epsilon \geq \frac{\|x^0 - x^*\|^2}{2\alpha k} \geq f(x^k) - f^*.$$

It follows that the iteration complexity is

$$\frac{L\|x^0 - x^*\|^2}{2\epsilon} = \mathcal{O}(1/\epsilon).$$

Note that the error $e(x) = |f(x) - f^*|$ decreases at a rate of $\mathcal{O}(1/k)$, we have

$$\lim_{k\to\infty} \frac{|f(x^{k+1}) - f^*|}{|f(x^k) - f^*|} = \lim_{k\to\infty} \frac{\frac{1}{k+1}}{\frac{1}{k}} = 1,$$

which implies sublinear convergence[2]. □

Definition 3.3: Strong convexity

A function $f : \mathbb{R}^n \to \mathbb{R}$ is *μ-strongly convex* if there exists $\mu > 0$ such that for all $x, y \in \mathbb{R}^n$ and $0 \leq t \leq 1$,

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y) - \mu\frac{t(t-1)}{2}\|x - y\|^2. \tag{3.3}$$

One can craft a strongly convex function by adding $\|\cdot\|^2$ multiplied by some scalar to a convex function. An equivalent definition for μ-strongly convex function is there exists a convex function $g : \mathbb{R}^n \to \mathbb{R}$ such that

$$f(x) = g(x) + \frac{\mu}{2}\|x\|^2.$$

If f is continuous strongly convex, then it admits a *unique* minimizer. We could also extend the first-order convexity for μ-strongly convex functions:

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{\mu}{2}\|y - x\|^2. \tag{3.4}$$

[2] A similar result for backtracking line search can be found in [6, Theorem 6.2].

Theorem 3.4

Suppose that f is L-smooth and μ-strongly convex, for some $L > \mu > 0$. Let $\{x^k\}$ be a sequence generated by gradient descent and α^k is chosen with constant step size satisfying $0 < \alpha \leq 1/L$. Then for all $x^* \in \arg\min f$, we have

$$\|x^k - x^*\|^2 \leq (1 - \alpha\mu)^k \|x^0 - x^*\|^2. \tag{3.5}$$

Furthermore, for a given $\epsilon > 0$, the iteration complexity is $\mathcal{O}(\log(1/\epsilon))$. The iterates enjoy a linear convergence.

Proof. By μ-strongly convex (3.4) and inequality (2.6) with $\alpha \leq 1/L$, we have

$$\begin{aligned}
\|x^{k+1} - x^*\|^2 &= \|x^k - \alpha\nabla f(x^k) - x^*\|^2 \\
&= \|x^k - x^*\|^2 - 2\alpha\nabla f(x^k)^\top (x^k - x^*) + \alpha^2 \|\nabla f(x^k)\|^2 \\
&\leq (1 - \alpha\mu)\|x^k - x^*\|^2 - 2\alpha(f(x^k) - f^*) + \alpha^2\|\nabla f(x^k)\|^2 \\
&\leq (1 - \alpha\mu)\|x^k - x^*\|^2 - 2\alpha(f(x^k) - f^*) + 2\alpha^2 L(f(x^k) - f(x^{k+1})) \\
&\leq (1 - \alpha\mu)\|x^k - x^*\|^2 - 2\alpha(f(x^k) - f^*) + 2\alpha^2 L(f(x^k) - f^*) \\
&= (1 - \alpha\mu)\|x^k - x^*\|^2 - 2\alpha(1 - \alpha L)(f(x^k) - f^*).
\end{aligned}$$

Since $\alpha \leq 1/L$, we can ensure $2\alpha(1 - \alpha L)$ is non-positive, we have

$$\|x^{k+1} - x^*\|^2 \leq (1 - \alpha\mu)\|x^k - x^*\|^2 \implies \|x^k - x^*\|^2 \leq (1 - \alpha\mu)^k \|x^0 - x^*\|^2.$$

Note that $L > \mu \implies 1 - \alpha\mu > 0$. To obtain iteration complexity, we use a fundamental inequality of logarithm: for $x > 0$,

$$1 - \frac{1}{x} \leq \log(x) \implies \frac{1}{1 - (1 - \alpha\mu)} \log\left(\frac{1}{1 - \alpha\mu}\right) \geq 1.$$

Apply the logarithm to both sides of (3.5) with some rearrangement, we have

$$\begin{aligned}
k \geq \frac{1}{\alpha\mu} \log\left(\frac{\|x^0 - x^*\|^2}{\epsilon}\right) \implies \log\left(\frac{\|x^0 - x^*\|^2}{\|x^k - x^*\|^2}\right) &\geq k \log\left(\frac{1}{1 - \alpha\mu}\right) \\
&\geq \frac{1}{\alpha\mu} \log\left(\frac{1}{1 - \alpha\mu}\right) \log\left(\frac{\|x^0 - x^*\|^2}{\epsilon}\right) \\
&\geq \log\left(\frac{\|x^0 - x^*\|^2}{\epsilon}\right).
\end{aligned}$$

By applying exponential to both sides, we obtain $\|x^0 - x^*\|^2 \leq \epsilon$. It follows that the iteration complexity is

$$\frac{1}{\alpha\mu} \log\left(\frac{\|x^0 - x^*\|^2}{\epsilon}\right) = \mathcal{O}(\log(1/\epsilon)).$$

From the previous arguments, we have the error $e(x) = \|x - x^*\|^2$ satisfying

$$\lim_{k \to \infty} \frac{\|x^{k+1} - x^*\|^2}{\|x^k - x^*\|^2} = (1 - \alpha\mu) \in (0, 1),$$

which implies linear convergence. □

We next introduce the Polyak-Łojasiewicz condition, which is weaker than strong convexity but still yields similar convergence results for non-convex functions.

Definition 3.5: Polyak-Łojasiewicz condition

A differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is *μ-Polyak-Łojasiewicz* if it is bounded below, and if there exists $\mu > 0$ such that for all $x \in \mathbb{R}^n$,

$$f(x) - f^* \leq \frac{1}{2\mu}\|\nabla f(x)\|^2. \tag{3.6}$$

By inserting $x = x^*$ into inequality (3.6), we immediately have an important result of Polyak-Łojasiewicz functions: $x^* \in \arg\min f$ if and only if $\nabla f(x^*) = 0$.

Theorem 3.6

Suppose that f is L-smooth and μ-Polyak-Łojasiewicz, for some $L > \mu > 0$. Let $\{x^k\}$ be a sequence generated by gradient descent and α^k is chosen with constant step size satisfying $0 < \alpha \leq 1/L$. Then for all $x^* \in \arg\min f$, we have

$$f(x^k) - f^* \leq (1 - \alpha\mu)^k (f(x^0) - f^*). \tag{3.7}$$

Furthermore, for a given $\epsilon > 0$, the iteration complexity is $\mathcal{O}(\log(1/\epsilon))$. The iterates enjoy a linear convergence.

Proof. By L-smooth (2.6), $\alpha \leq 1/L$, and Polyak-Łojasiewicz condition (in the order of inequalities), we have

$$\begin{aligned} f(x^{k+1}) &\leq f(x^k) - \frac{\alpha}{2}(2 - \alpha L)\|\nabla f(x^k)\|^2 \\ &\leq f(x^k) - \frac{\alpha}{2}\|\nabla f(x^k)\|^2 \\ &\leq f(x^k) - \alpha\mu(f(x^k) - f^*). \end{aligned}$$

Subtract f^* on both sides, we obtain

$$f(x^{k+1}) - f^* \leq (1 - \alpha\mu)(f(x^k) - f^*) \implies f(x^k) - f^* \leq (1 - \alpha\mu)^k (f(x^0) - f^*).$$

Following the proof of Theorem 3.4 and apply the logarithm to both sides of (3.7), we have

$$\begin{aligned} k \geq \frac{1}{\alpha\mu}\log\left(\frac{f(x^0) - f^*}{\epsilon}\right) &\implies \log\left(\frac{f(x^0) - f^*}{f(x^k) - f^*}\right) \geq \log\left(\frac{f(x^0) - f^*}{\epsilon}\right) \\ &\implies f(x^k) \leq \epsilon. \end{aligned}$$

It follows that the iteration complexity is also $\mathcal{O}(\log(1/\epsilon))$. With the error $e(x) = |f(x) - f^*|$, we have

$$\lim_{k\to\infty} \frac{|f(x^{k+1}) - f^*|}{|f(x^k) - f^*|} = (1 - \alpha\mu) \in (0, 1),$$

which again implies linear convergence. □

Acknowledge

The proof of Proposition 2.1 is taken from [4, Proposition 17.6], Proposition 2.2 is taken from [1, Proposition 1.2.1] and [4, Proposition 17.7], Lemma 2.4 is taken from [3, Lemma 2.25], and Proposition 2.5 is taken from [1, Proposition 1.2.3]. Proofs of Theorem 3.2, Theorem 3.4, and Theorem 3.6 are taken from [6, Theorem 6.1], [3, Theorem 3.6], and [3, Theorem 3.9], respectively, with some modifications for consistency across proofs. I would like to thank the lecture notes/slides [4–7] for helping me understand the tedious convergence analysis, and I highly recommend that readers check them.

References

[1] D. P. Bertsekas. *Nonlinear programming.* Athena Scientific, 1999.

[2] S. Boyd and L. Vandenberghe. *Convex optimization.* Cambridge University Press, 2004.

[3] G. Garrigos and R. M. Gower. Handbook of convergence theorems for (stochastic) gradient methods. *arXiv:2301.11235*, 2023.

[4] M. Hardt. EE 227C: Convex optimization and approximation (lecture notes), 2018.

[5] M. Schmidt. CPSC 5XX: First-order optimization algorithms for machine learning (lecture slides), 2018.

[6] R. Tibshirani. 10-725: Convex optimization (lecture notes), 2013.

[7] S. Zhang. IE 8521: Optimization (lecture slides), 2020.