Chapter 2

Foundations

We outline here the foundations of the algorithms and theory discussed in later chapters. These foundations include a review of Taylor's theorem and its consequences that form the basis of much of smooth nonlinear optimization. We also provide a concise review of elements of convex analysis that will be used throughout the book.

2.1 A Taxonomy of Solutions to Optimization Problems

Before we can begin designing algorithms, we must determine what it means to *solve* an optimization problem. Suppose that f is a function mapping some domain $\mathcal{D} \subset \mathbb{R}^n$ to the real line $\mathbb{R}$. We have the following definitions.

- $x^* \in \mathcal{D}$ is a *local minimizer* of f if there is a neighborhood $\mathcal{N}$ of x^* such that $f(x) \geq f(x^*)$ for all $x \in \mathcal{N} \cap \mathcal{D}$.
- $x^* \in \mathcal{D}$ is a *global minimizer* of f if $f(x) \geq f(x^*)$ for all $x \in \mathcal{D}$.
- $x^* \in \mathcal{D}$ is a *strict local minimizer* if it is a local minimizer for some neighborhood $\mathcal{N}$ of x^*, and in addition $f(x) > f(x^*)$ for all $x \in \mathcal{N}$ with $x \neq x^*$.
- x^* is an *isolated local minimizer* if there is a neighborhood $\mathcal{N}$ of x^* such that $f(x) \geq f(x^*)$ for all $x \in \mathcal{N} \cap \mathcal{D}$ and in addition, $\mathcal{N}$ contains no local minimizers other than x^*.

For the constrained optimization problem

$$\min_{x \in \Omega} f(x), \tag{2.1}$$

where $\Omega \subset \mathcal{D} \subset \mathbb{R}^n$ is a closed set, we modify the terminology slightly to use the word "solution" rather than "minimizer." That is, we have the following definitions.

- $x^* \in \Omega$ is a *local solution* of (2.1) if there is a neighborhood $\mathcal{N}$ of x^* such that $f(x) \geq f(x^*)$ for all $x \in \mathcal{N} \cap \Omega$.
- $x^* \in \Omega$ is a *global solution* of (2.1) if $f(x) \geq f(x^*)$ for all $x \in \Omega$.

One of the immediate challenges is to provide a simple means of determining whether a particular point is a local or global solution. To do so, we introduce a powerful tool from calculus: Taylor's theorem. Taylor's theorem is the most important theorem in all of continuous optimization, and we review it next.

2.2 Taylor's Theorem

Taylor's theorem shows how smooth functions can be approximated locally by polynomials that depend on low-order derivatives of f.

Theorem 2.1. *Given a continuously differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, and given $x, p \in \mathbb{R}^n$, we have that*

$$f(x+p) = f(x) + \int_0^1 \nabla f(x+\gamma p)^T p \, d\gamma, \tag{2.2}$$

$$f(x+p) = f(x) + \nabla f(x+\gamma p)^T p, \quad \textit{some } \gamma \in (0,1). \tag{2.3}$$

If f is twice continuously differentiable, we have

$$\nabla f(x+p) = \nabla f(x) + \int_0^1 \nabla^2 f(x+\gamma p) p \, d\gamma, \tag{2.4}$$

$$f(x+p) = f(x) + \nabla f(x)^T p + \frac{1}{2} p^T \nabla^2 f(x+\gamma p) p, \quad \textit{some } \gamma \in (0,1). \tag{2.5}$$

(We sometimes call the relation (2.2) the "integral form" and (2.3) the "mean-value form" of Taylor's theorem.)

A consequence of (2.3) is that for f continuously differentiable at x, we have [1]

$$f(x+p) = f(x) + \nabla f(x)^T p + o(\|p\|). \tag{2.6}$$

We prove this claim by manipulating (2.3) as follows:

$$\begin{aligned} f(x+p) &= f(x) + \nabla f(x+\gamma p)^T p \\ &= f(x) + \nabla f(x)^T p + (\nabla f(x+\gamma p) - \nabla f(x))^T p \\ &= f(x) + \nabla f(x)^T p + O(\|\nabla f(x+\gamma p) - \nabla f(x)\| \|p\|) \\ &= f(x) + \nabla f(x)^T p + o(\|p\|), \end{aligned}$$

where the last step follows from continuity: $\nabla f(x+\gamma p) - \nabla f(x) \to 0$ as $p \to 0$, for all $\gamma \in (0,1)$.

As we will see throughout this text, a crucial quantity in optimization is the Lipschitz constant L for the gradient of f, which is defined to satisfy

$$\|\nabla f(x) - \nabla f(y)\| \le L \|x-y\|, \quad \text{for all } x, y \in \text{dom}\,(f). \tag{2.7}$$

We say that a continuously differentiable function f with this property is *L-smooth* or has *L-Lipschitz gradients*. We say that f is *L_0-Lipschitz* if

$$|f(x) - f(y)| \le L_0 \|x-y\|, \quad \text{for all } x, y \in \text{dom}\,(f). \tag{2.8}$$

[1] See the Appendix for a description of the order notation $O(\cdot)$ and $o(\cdot)$.

From (2.2), we have

$$f(y) - f(x) - \nabla f(x)^T(y-x) = \int_0^1 [\nabla f(x+\gamma(y-x)) - \nabla f(x)]^T(y-x)\,d\gamma.$$

By using (2.7), we have

$$[\nabla f(x+\gamma(y-x)) - \nabla f(x)]^T(y-x) \le \|\nabla f(x+\gamma(y-x)) - \nabla f(x)\|\|y-x\| \le L\gamma\|y-x\|^2.$$

By substituting this bound into the previous integral, we obtain the following.

Lemma 2.2. *Given an L-smooth function f, we have for any $x, y \in \mathrm{dom}\,(f)$ that*

$$f(y) \le f(x) + \nabla f(x)^T(y-x) + \frac{L}{2}\|y-x\|^2. \tag{2.9}$$

Lemma 2.2 asserts that f can be upper bounded by a quadratic function whose value at x is equal to $f(x)$.

When f is *twice* continuously differentiable, we can characterize the constant L in terms of the eigenvalues of the Hessian $\nabla^2 f(x)$. Specifically, we have

$$-LI \preceq \nabla^2 f(x) \preceq LI, \quad \text{for all } x \tag{2.10}$$

as the following result proves.

Lemma 2.3. *Suppose f is twice continuously differentiable on $\mathbb{R}^n$. Then if f is L-smooth, we have $\nabla^2 f(x) \preceq LI$ for all x. Conversely, if $-LI \preceq \nabla^2 f(x) \preceq LI$, then f is L-smooth.*

Proof. From (2.9), we have by setting $y = x + \alpha p$ for some $\alpha > 0$ that

$$f(x+\alpha p) - f(x) - \alpha \nabla f(x)^T p \le \frac{L}{2}\alpha^2\|p\|^2.$$

From formula (2.5) from Taylor's theorem, we have

$$f(x+\alpha p) - f(x) - \alpha \nabla f(x)^T p = \frac{1}{2}\alpha^2 p^T \nabla^2 f(x+\gamma\alpha p)p.$$

By comparing these two expressions, we obtain

$$p^T \nabla^2 f(x+\gamma\alpha p)p \le L\|p\|^2.$$

By letting $\alpha \downarrow 0$, we have that all eigenvalues of $\nabla^2 f(x)$ are bounded by L, so that $\nabla^2 f(x) \preceq LI$, as claimed.

Suppose now that $-LI \preceq \nabla^2 f(x) \preceq LI$ for all x, so that $\|\nabla^2 f(x)\| \le L$ for all x. We have from (2.4) that

$$\begin{aligned}
\|\nabla f(y) - \nabla f(x)\| &= \left\| \int_{t=0}^1 \nabla^2 f(x+t(y-x))(y-x)\,dt \right\| \\
&\le \int_{t=0}^1 \|\nabla^2 f(x+t(y-x))\|\|y-x\|\,dt \\
&\le \int_{t=0}^1 L\|y-x\|\,dt = L\|y-x\|,
\end{aligned}$$

as required. This completes the proof. $\square$

2.3 Characterizing Minima of Smooth Functions

The results of Section 2.2 give us the tools needed to characterize solutions of the unconstrained optimization problem

$$\min_{x \in \mathbb{R}^n} f(x), \tag{2.11}$$

where f is a smooth function.

We start with *necessary* conditions, which give properties of the derivatives of f that are satisfied when x^* is a local solution. We have the following result.

Theorem 2.4 (Necessary Conditions for Smooth Unconstrained Optimization)**.**

(a) Suppose that f is continuously differentiable. If x^* is a local minimizer of (2.11)*, then $\nabla f(x^*) = 0$.*

(b) Suppose that f is twice continuously differentiable. If x^* is a local minimizer of (2.11)*, then $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*)$ is positive semidefinite.*

Proof. We start by proving (a). Suppose for contradiction that $\nabla f(x^*) \neq 0$, and consider a step $-\alpha \nabla f(x^*)$ away from x^*, where α is a small positive number. By setting $p = -\alpha \nabla f(x^*)$ in formula (2.3) from Theorem 2.1, we have

$$f(x^* - \alpha \nabla f(x^*)) = f(x^*) - \alpha \nabla f\left(x^* - \gamma\alpha \nabla f(x^*)\right)^T \nabla f(x^*), \quad \text{for some } \gamma \in (0,1). \tag{2.12}$$

Since ∇f is continuous, we have that

$$\nabla f\left(x^* - \gamma\alpha \nabla f(x^*)\right)^T \nabla f(x^*) \geq \frac{1}{2}\|\nabla f(x^*)\|^2,$$

for all α sufficiently small, and any $\gamma \in (0,1)$. Thus by substituting into (2.12), we have that

$$f(x^* - \alpha \nabla f(x^*)) = f(x^*) - \frac{1}{2}\alpha \|\nabla f(x^*)\|^2 < f(x^*),$$

for all positive and sufficiently small α. No matter how we choose the neighborhood $\mathcal{N}$ in the definition of local minimizer, it will contain points of the form $x^* - \alpha \nabla f(x^*)$ for sufficiently small α. Thus, it is impossible to choose a neighborhood $\mathcal{N}$ of x^* such that $f(x) \geq f(x^*)$ for all $x \in \mathcal{N}$, so x^* is not a local minimizer.

We now prove (b). It follows immediately from (a) that $\nabla f(x^*) = 0$, so we need to prove only positive semidefiniteness of $\nabla^2 f(x^*)$. Suppose for contradiction that $\nabla^2 f(x^*)$ has a negative eigenvalue, so there exists a vector $v \in \mathbb{R}^n$ and a positive scalar λ such that $v^T \nabla^2 f(x^*) v \leq -\lambda$. We set $x = x^*$ and $p = \alpha v$ in formula (2.5) from Theorem 2.1, where α is a small positive constant, to obtain

$$f(x^+ \alpha v) = f(x^*) + \alpha \nabla f(x^*)^T v + \frac{1}{2}\alpha^2 v^T \nabla^2 f(x^* + \gamma\alpha v) v, \quad \text{for some } \gamma \in (0,1). \tag{2.13}$$

For all α sufficiently small, we have for λ defined above that $v^T \nabla^2 f(x^* + \gamma\alpha v) v \leq -\lambda/2$, for all $\gamma \in (0,1)$. By substituting this bound together with $\nabla f(x^*) = 0$ into (2.13), we obtain

$$f(x^* + \alpha v) = f(x^*) - \frac{1}{4}\alpha^2 \lambda < f(x^*),$$

for all sufficiently small, positive values of α. Thus there is no neighborhood $\mathcal{N}$ of x^* such that $f(x) \geq f(x^*)$ for all $x \in \mathcal{N}$, so x^* is not a local minimizer. Thus we have proved by contradiction that $\nabla^2 f(x^*)$ is positive semidefinite. □

Condition (a) in Theorem 2.4 is called the *first-order necessary condition*, because it involves the first-order derivatives of f. Similarly, condition (b) is called the *second-order necessary condition.*

We call any point x satisfying $\nabla f(x) = 0$ a *stationary point.*

We additionally have the following *second-order sufficient condition.*

Theorem 2.5 (Sufficient Conditions for Smooth Unconstrained Optimization)**.** *Suppose that f is twice continuously differentiable and that for some x^*, we have $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*)$ is positive definite. Then x^* is a strict local minimizer of* (2.11).

Proof. We use formula (2.5) from Taylor's theorem. Define a radius ρ sufficiently small and positive such that the eigenvalues of $\nabla^2 f(x^* + \gamma p)$ are bounded below by some positive number ϵ, for all $p \in \mathbb{R}^n$ with $\|p\| \leq \rho$, and all $\gamma \in (0, 1)$. (Because $\nabla^2 f$ is positive definite at x^* and continuous, and because the eigenvalues of a matrix are continuous functions of the elements of a matrix, it is possible to choose $\rho > 0$ and $\epsilon > 0$ with these properties.) By setting $x = x^*$ in (2.5), we have

$$f(x^* + p) = f(x^*) + \nabla f(x^*)^T p + \frac{1}{2} p^T \nabla^2 f(x^* + \gamma p) p \geq f(x^*) + \frac{1}{2}\epsilon \|p\|^2, \quad \text{for all } p \text{ with } \|p\| \leq \rho.$$

thus by setting $\mathcal{N} = \{x^* + p \,|\, \|p\| < \rho\}$, we have found a neighborhood of x^* such that $f(x) > f(x^*)$ for all $x \in \mathcal{N}$ with $x \neq x^*$, thus satisfying the conditions for a strict local minimizer. □

The sufficiency promised by Theorem 2.5 only guarantees a *locally* optimal solution. We now turn to a special but ubiquitous class of functions and sets for which we can provide necessary and sufficient guarantees for optimality, using only information from low order derivatives.

2.4 Convex Sets and Functions

Convex functions take a central role in optimization precisely because these are the instances for which it is easy to verify optimality and for which such optima are guaranteed to be discoverable within a reasonable amount of computation.

A convex set $\Omega \subset \mathbb{R}^n$ has the property that

$$x, y \in \Omega \;\; \Rightarrow \;\; (1 - \alpha)x + \alpha y \in \Omega \;\text{ for all } \alpha \in [0, 1]. \tag{2.14}$$

For all pairs of points (x, y) contained in Ω, the line segment between x and y is also contained in Ω. The convex sets that we consider in this book are usually *closed.*

The defining property of a convex function is the following inequality:

$$f((1 - \alpha)x + \alpha y) \leq (1 - \alpha) f(x) + \alpha f(y), \quad \text{for all } x, y \in \mathbb{R}^n \text{ and all } \alpha \in [0, 1]. \tag{2.15}$$

The line segment connecting $(x, f(x))$ and $(y, f(y))$ lies entirely above the graph of the function f. In other words, the *epigraph* of f, defined as

$$\operatorname{epi} f := \{(x, t) \in \mathbb{R}^n \times \mathbb{R} \,|\, t \geq f(x)\} \tag{2.16}$$

is a convex set.

The concepts of "minimizer" and "solution" for the case of convex objective function and constraint set become more elementary in the convex case than in the general case of Section 2.1. In particular, the distinction between "local" and "global" solutions goes away, as we show now.

Theorem 2.6. *Suppose that in the general constrained optimization problem* (2.1)*, the function f is convex and the set Ω is closed and convex. We have the following.*

(a) Any local solution of (2.1) *is also a global solution.*

(b) The set of global solutions of (2.1) *is a convex set.*

Proof. For (a), suppose for contradiction that $x^* \in \Omega$ is a local solution but not a global solution, so there exists a point $\bar{x} \in \Omega$ such that $f(\bar{x}) < f(x^*)$. Then by convexity we have for any $\alpha \in (0,1)$ that

$$f(x^* + \alpha(\bar{x} - x^*)) \le (1-\alpha) f(x^*) + \alpha f(\bar{x}) < f(x^*).$$

But for any neighborhood $\mathcal{N}$, we have for sufficiently small $\alpha > 0$ that $x^* + \alpha(\bar{x} - x^*)) \in \mathcal{N} \cap \Omega$ and $f(x^* + \alpha(\bar{x} - x^*)) < f(x^*)$, contradicting the definition of a local minimizer.

For (b), we simply apply the definition of convexity for both sets and functions. Given any global solutions x^* and $\bar{x}$, we have $f(\bar{x}) = f(x^*)$, so for any $\alpha \in [0,1]$ we have

$$f(x^* + \alpha(\bar{x} - x^*)) \le (1-\alpha) f(x^*) + \alpha f(\bar{x}) = f(x^*).$$

We have also that $f(x^* + \alpha(\bar{x} - x^*)) \ge f(x^*)$, since $x^* + \alpha(\bar{x} - x^*) \in \Omega$ and x^* is a global minimizer. It follows from these two inequalities that $f(x^* + \alpha(\bar{x} - x^*)) = f(x^*)$, so that $x^* + \alpha(\bar{x} - x^*)$ is also a global minimizer. □

By applying Taylor's theorem (in particular, (2.6)) to the left-hand side of the definition of convexity (2.15), we obtain

$$f(x + \alpha(y-x)) = f(x) + \alpha \nabla f(x)^T (y-x) + o(\alpha) \le (1-\alpha) f(x) + \alpha f(y).$$

By canceling the $f(x)$ term, rearranging, and dividing by α, we obtain

$$f(y) \ge f(x) + \nabla f(x)^T (y-x) + o(1),$$

and when $\alpha \downarrow 0$, the $o(1)$ term vanishes, so we obtain

$$f(y) \ge f(x) + \nabla f(x)^T (y-x), \quad \text{for any } x, y \in \operatorname{dom}(f), \tag{2.17}$$

which is a fundamental characterization of convexity of a smooth function.

While Theorem 2.4 provides a necessary link between the vanishing of ∇f and the minimizing of f, the first-order necessary condition is actually a *sufficient* condition when f is convex.

Theorem 2.7. *Suppose that f is continuously differentiable and convex. Then if $\nabla f(x^*) = 0$, then x^* is a global minimizer of* (2.11).

Proof. The proof of the first part follows immediately from condition (2.17), if we set $x = x^*$. Using this inequality together with $\nabla f(x^*) = 0$, we have for any y that

$$f(y) \ge f(x^*) + \nabla f(x^*)^T (y - x^*) = f(x^*),$$

so that x^* is a global minimizer. □

2.5 Strongly Convex Functions

For the remainder of this section, we assume that f is continuously differentiable and also *convex*. If there exists a value $m > 0$ such that

$$f((1-\alpha)x + \alpha y) \leq (1-\alpha)f(x) + \alpha f(y) - \frac{1}{2}m\alpha(1-\alpha)\|x-y\|_2^2 \tag{2.18}$$

for all x and y in the domain of f, we say that f is *strongly convex with modulus of convexity* m. When f is differentiable, we have the following equivalent definition, obtained by working on (2.18) with a similar argument to the one above:

$$f(y) \geq f(x) + \nabla f(x)^T(y-x) + \frac{m}{2}\|y-x\|^2. \tag{2.19}$$

Note that this inequality complements the inequality satisfied by functions with smooth gradients. When the gradients are smooth, a function can be upper bounded by a quadratic which takes the value $f(x)$ at x. When the function is strongly convex, it can be *lower bounded* by a quadratic which takes the value $f(x)$ and x.

We have the following extension of Theorem 2.7, whose proof follows immediately by setting $x = x^*$ in (2.19).

Theorem 2.8. *Suppose that f is continuously differentiable and strongly convex. Then if $\nabla f(x^*) = 0$, then x^* is the* unique *global minimizer of f.*

This approximation of convex f by quadratic functions is a key theme in continuous optimization.

When f is strongly convex and twice continuously differentiable, (2.5) implies the following, when x^* is the minimizer:

$$f(x) - f(x^*) = \frac{1}{2}(x-x^*)^T\nabla^2 f(x^*)(x-x^*) + o(\|x-x^*\|^2). \tag{2.20}$$

Thus, f behaves like a strongly convex *quadratic* function in a neighborhood of x^*. It follows that we can learn a lot about local convergence properties of algorithms just by studying convex quadratic functions. We use quadratic functions as a guide for both intuition and algorithmic derivation throughout.

Just as we could characterize the Lipschitz constant of the gradient in terms of the eigenvalues of the Hessian, the strong convexity parameter provides a lower bound on the eigenvalues of the Hessian when f is twice continuously differentiable.

Lemma 2.9. *Suppose that f is twice continuously differentiable on $\mathbb{R}^n$. Then f has modulus of convexity m if and only if $\nabla^2 f(x) \succeq mI$ for all x.*

Proof. For any $x, u \in \mathbb{R}^n$ and $\alpha > 0$, we have from Taylor's theorem that

$$f(x+\alpha u) = f(x) + \alpha\nabla f(x)^T + \frac{1}{2}\alpha^2 u^T\nabla^2 f(x+t\alpha u)u, \quad \text{for some } t \in (0,1).$$

From the strong convexity property, we have

$$f(x+\alpha u) \geq f(x) + \alpha\nabla f(x)^T u + \frac{m}{2}\alpha^2\|u\|^2.$$

By comparing these two expressions, canceling terms, and dividing by α^2, we obtain

$$u^T \nabla^2 f(x + t\alpha u)u \geq m\|u\|^2.$$

By taking $\alpha \downarrow 0$, we obtain $u^T \nabla^2 f(x)u \geq m\|u\|^2$, thus proving that $\nabla^2 f(x) \succeq mI$.

For the converse, suppose that $\nabla^2 f(x) \succeq mI$ for all x. Using the same form of Taylor's theorem as above, we obtain

$$f(z) = f(x) + \nabla f(x)^T(z - x) + \frac{1}{2}(z - x)^T \nabla^2 f(x + t(z - x))(z - x), \quad \text{for some } t \in (0, 1).$$

We obtain the strong convexity expression when we bound the last term as follows:

$$(z - x)^T \nabla^2 f(x + t(z - x))(z - x) \geq m\|z - x\|^2,$$

completing the proof. □

The following corollary is a immediate consequence of Lemma 2.3.

Corollary 2.10. *Suppose that the conditions of Lemma 2.3 hold, and in addition that f is convex. Then $0 \preceq \nabla^2 f(x) \preceq LI$ if and only if f is L-smooth.*

Notation

We use $\|\cdot\|$ to denote the Euclidean norm $\|\cdot\|_2$ of a vector in $\mathbb{R}^n$. Other norms, such as $\|\cdot\|_1$ and $\|\cdot\|_\infty$, will be denoted explicitly.

Notes and References

The classic reference on convex analysis remains the 1970 text of Rockafellar [85], which is still remarkably fresh with many useful results. A more recent classic of Boyd and Vandenberghe [16] contains an enormous range of information about convex optimization, especially on formulations and applications.

Exercises

1. Prove that the effective domain of a convex function f (that is, the set of points $x \in \mathbb{R}^n$ such that $f(x) < \infty$) is a convex set.

2. Prove that $\operatorname{epi} f$ is a convex subset of $\mathbb{R}^n \times \mathbb{R}$ for any convex function f.

3. Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is convex and concave. Show that f must be an affine function.

4. Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is convex and bounded above. Show that f must be a constant function.

5. Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is strongly convex and Lipschitz. Show no such f exists.

6. Show rigorously how (2.19) is derived from (2.18) when f is continuously differentiable.

7. Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is a convex function with L-Lipschitz gradient and a minimizer x^* with function value $f^* = f(x^*)$.

   (a) Show (by minimizing both sides of (2.9) with respect to y) that for any $x \in R^n$ we have

   $$f(x) - f^* \geq \frac{1}{2L}\|\nabla f(x)\|^2.$$

   (b) Prove the following *co-coercivity* property: For any $x, y \in \mathbb{R}^n$, we have

   $$[\nabla f(x) - \nabla f(y)]^T(x - y) \geq \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|^2.$$

   Hint: Apply part (a) to the following two functions:

   $$h_x(z) := f(z) - \nabla f(x)^T z, \quad h_y(z) := f(z) - \nabla f(y)^T z.$$

8. Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is an m-strongly convex function with L-Lipschitz gradient and (unique) minimizer x^* with function value $f^* = f(x^*)$.

   (a) Show that the function $q(x) := f(x) - \frac{m}{2}\|x\|^2$ is convex with $L - m$-Lipschitz continuous gradients.

   (b) By applying the co-coercivity property of the previous question to this function q, show that the following property holds:

   $$[\nabla f(x) - \nabla f(y)]^T(x - y) \geq \frac{mL}{m + L}\|x - y\|^2 + \frac{1}{m + L}\|\nabla f(x) - \nabla f(y)\|^2. \tag{2.21}$$