1 Outline

In this lecture, we study

- Properties of smooth functions,
- Properties of strongly convex functions,
- Convergence rate of gradient descent for functions that are smooth and strongly convex.
- Projected gradient descent for constrained minimization.

2 Smooth functions

2.1 Quadratic upper bounds on smooth functions

In the last lecture, we discussed the convergence rate of gradient descent on smooth functions. The key property of smooth functions based on which we developed the analysis is the following quadratic upper bound inequality for smooth functions. Namely, for a β-smooth convex function $f : \mathbb{R}^d \to \mathbb{R}$, we have

$$f(y) \leq f(x) + \nabla f(x)^\top (y - x) + \frac{\beta}{2}\|y - x\|_2^2$$

for any $x, y \in \mathbb{R}^d$. Let us verify this while we discuss some important properties of smooth functions.

Theorem 11.1. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function. Then f is β-smooth in the ℓ_2 norm for some $\beta > 0$ if and only if $g(x) = (\beta/2)\|x\|_2^2 - f(x)$ is convex.*

Proof. ($\Rightarrow$) Recall that g is convex if and only if the monotonicity condition is satisfied. Note that $\nabla g(x) = \beta x - \nabla f(x)$ and

$$\langle \nabla g(x) - \nabla g(y), x - y \rangle = \beta\|x - y\|_2^2 - \langle \nabla f(x) - \nabla f(y), x - y \rangle.$$

Then by the Cauchy-Schwarz inequality,

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \leq \|\nabla f(x) - \nabla f(y)\|_2 \cdot \|x - y\|_2 \leq \beta\|x - y\|_2^2.$$

Therefore, we have $\langle \nabla g(x) - \nabla g(y), x - y \rangle \geq 0$, which implies that g is convex.

($\Leftarrow$) Since g is convex, the first-order characterization of convex functions states that

$$\frac{\beta}{2}\|z\|_2^2 - f(z) \geq \frac{\beta}{2}\|x\|_2^2 - f(x) + (\beta x - \nabla f(x))^\top (z - x) \quad \text{for any } z \in \mathbb{R}^d,$$

which is equivalent to

$$f(z) \leq f(x) + \nabla f(x)^\top (z - x) + \frac{\beta}{2}\|z - x\|_2^2 \quad \text{for any } z \in \mathbb{R}^d.$$

Then taking $z = y + (1/\beta)(\nabla f(x) - \nabla f(y))$,

$$\begin{aligned} f(x) - f(y) &= f(x) - f(z) + f(z) - f(y) \\ &\le -\nabla f(x)^\top (z - x) + \nabla f(y)^\top (z - y) + \frac{\beta}{2}\|z - y\|_2^2 \\ &= \nabla f(x)^\top (x - y) + (\nabla f(x) - \nabla f(y))^\top (y - z) + \frac{\beta}{2}\|z - y\|_2^2 \\ &= \nabla f(x)^\top (x - y) - \frac{1}{2\beta}\|\nabla f(x) - \nabla f(y)\|_2^2. \end{aligned}$$

Then it follows that

$$\frac{1}{2\beta}\|\nabla f(x) - \nabla f(y)\|_2^2 \le f(y) - f(x) + \nabla f(x)^\top (x - y).$$

Similarly, we obtain

$$\frac{1}{2\beta}\|\nabla f(y) - \nabla f(x)\|_2^2 \le f(x) - f(y) + \nabla f(y)^\top (y - x).$$

Adding these two inequalities, it follows that

$$\frac{1}{\beta}\|\nabla f(x) - \nabla f(y)\|_2^2 \le (\nabla f(x) - \nabla f(y))^\top (x - y).$$

Since $(\nabla f(x) - \nabla f(y))^\top (x - y) \le \|\nabla f(x) - \nabla f(y)\|_2 \cdot \|x - y\|_2$ by the Cauchy-Schwarz inequality, we obtain

$$\|\nabla f(x) - \nabla f(y)\|_2 \le \beta \|x - y\|_2.$$

Therefore, f is β-smooth. □

By the first-order characterization of convex functions, $g(x) = \frac{\beta}{2}\|x\|_2^2 - f(x)$ is convex if and only if $g(y) \ge g(x) + \nabla g(x)^\top (y - x)$ for any $x, y \in \mathbb{R}^d$. This condition is equivalent to

$$f(y) \le f(x) + \nabla f(x)^\top (y - x) + \frac{\beta}{2}\|y - x\|_2^2,$$

which verifies the quadratic upper bound property of smooth functions.

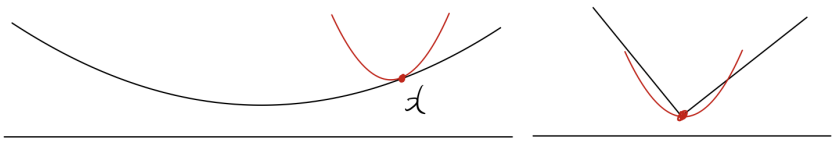

Figure 11.1: Quadratic upper bound on a smooth function

In fact, the inequality holds regardless of whether f is convex or not. We prove the following stronger statement.

Lemma 11.2. *If f is β-smooth in the ℓ_2 norm, then*

$$\left| f(y) - f(x) - \nabla f(x)^\top (y - x) \right| \le \frac{\beta}{2}\|y - x\|^2.$$

Proof. By the fundamental theorem of calculus and the Cauchy-Schwarz inequality, we obtain the following.

$$\begin{aligned}\left|f(y) - f(x) - \nabla f(x)^\top (y-x)\right| &= \left|\int_0^1 (y-x)^\top \left(\nabla f(x + t(y-x)) - \nabla f(x)\right) dt\right| \\ &\leq \int_0^1 \|y-x\|_2 \left\|\nabla f(x + t(y-x)) - \nabla f(x)\right\|_2 dt \\ &\leq \int_0^1 \beta t \|y-x\|_2^2 dt \\ &= \frac{\beta}{2}\|y-x\|_2^2\end{aligned}$$

where the equality is due to the fundamental theorem of calculus, the first inequality is by the Cauchy-Schwarz inequality, and the second inequality is from the β-smoothness of f. □

2.2 Optimality gap for smooth functions

Another interesting result is that when f is smooth, we can measure the gap between the optimal value and $f(x)$ for any given solution x. More precisely, we prove the following result.

Theorem 11.3. *If $f : \mathbb{R}^d \to \mathbb{R}$ is β-smooth in the ℓ_2 norm and convex, then*

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

where x^* is an optimal solution to $\min_{x\in\mathbb{R}^d} f(x)$.

Proof. Let us prove the upper bound on $f(x) - f(x^*)$ first. As f is β-smooth, we have

$$f(x) \leq f(x^*) + \nabla f(x^*)^\top (x - x^*) + \frac{\beta}{2}\|x - x^*\|_2^2,$$

which implies the upper bound as $\nabla f(x^*) = 0$. For the lower bound, note that for any $y \in \mathbb{R}^d$,

$$f(x^*) \leq f(y) \leq f(x) + \nabla f(x)^\top (y - x) + \frac{\beta}{2}\|y - x\|_2^2.$$

Here, we can take $y = x - (1/\beta)\nabla f(x)$, which makes the right-most side

$$f(x) - \frac{1}{2\beta}\|\nabla f(x)\|_2^2.$$

Then it follows that

$$f(x^*) \leq f(x) - \frac{1}{2\beta}\|\nabla f(x)\|_2^2,$$

as required. □

Based on Theorem 11.3, we can prove the following property of smooth functions.

Lemma 11.4. *If $f : \mathbb{R}^d \to \mathbb{R}$ is β-smooth in the ℓ_2 norm and convex, then*

$$(\nabla f(x) - \nabla f(y))^\top (x - y) \geq \frac{1}{\beta}\|\nabla f(x) - \nabla f(y)\|_2^2$$

for any $x, y \in \mathbb{R}^d$.

Proof. Given $x, y \in \mathbb{R}^d$, we take the following two functions.

$$g(z) = f(z) - \nabla f(x)^\top z,$$
$$h(z) = f(z) - \nabla f(y)^\top z.$$

As $\nabla g(z) = \nabla f(z) - \nabla f(x)$ and $\nabla h(z) = \nabla f(z) - \nabla f(y)$, it follows that x and y minimize g and h, respectively. Moreover, g and h are both β-smooth. Note that

$$\begin{aligned} f(y) - f(x) - \nabla f(x)^\top (y - x) &= g(y) - g(x) \\ &\geq \frac{1}{2\beta} \|\nabla g(y)\|_2^2 \\ &= \frac{1}{2\beta} \|\nabla f(y) - \nabla f(x)\|_2^2. \end{aligned}$$

Similarly, we have

$$\begin{aligned} f(x) - f(y) - \nabla f(y)^\top (x - y) &= h(x) - h(y) \\ &\geq \frac{1}{2\beta} \|\nabla h(x)\|_2^2 \\ &= \frac{1}{2\beta} \|\nabla f(x) - \nabla f(y)\|_2^2. \end{aligned}$$

Adding these two inequalities, we obtain

$$(\nabla f(x) - \nabla f(y))^\top (x - y) \geq \frac{1}{\beta} \|\nabla f(x) - \nabla f(y)\|_2^2,$$

as required. □

3 Strongly convex functions

3.1 Quadratic lower bounds on smooth functions

Recall that a function f is α-strongly convex in the norm $\|\cdot\|_2$ for some $\alpha > 0$ if $f(x) - \frac{\alpha}{2}\|x\|_2^2$ is convex. Why is strong-convexity useful? Let us first derive the following property of strongly convex functions.

Lemma 11.5. *If f is α-strongly convex in the ℓ_2 norm, then*

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{\alpha}{2} \|y - x\|_2^2.$$

Proof. By definition, we have that $g(x) = f(x) - \frac{\alpha}{2}\|x\|_2^2$ is convex. Then by the first-order characterization of convex functions, we have $g(y) \geq g(x) + \nabla g(x)^\top (y - x)$ for any $x, y \in \mathbb{R}^d$. This is equivalent to

$$\begin{aligned} f(y) &\geq f(x) + (\nabla f(x) - \alpha x)^\top (y - x) - \frac{\alpha}{2}\|x\|_2^2 + \frac{\alpha}{2}\|y\|_2^2 \\ &= f(x) + \nabla f(x)^\top (y - x) + \frac{\alpha}{2}\|y - x\|_2^2, \end{aligned}$$

as required. □

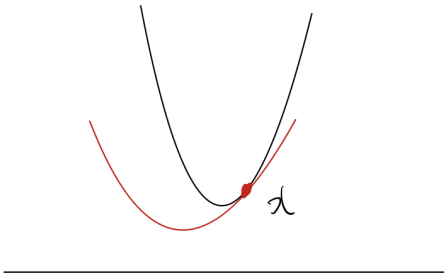

Figure 11.2: Quadratic lower bound on a strongly convex function

Lemma 11.5 implies that a strongly convex function is lower bounded by a quadratic function. This means that, when a point is far from an optimal solution, the gradient at this point has to be large. Hence, when applying gradient descent or the subgradient method, this leads to a faster convergence.

In Figure 11.3, we compare smoothness and strong convexity. The first figure shows a strongly convex function that is not smooth, obtained by taking the point-wise maximum of two smooth functions. The second figure illustrates a smooth function that is not strongly convex. In particular, the second figure shows a smooth curve around the minimum point while it exhibits a linear growth when being far away from the minimum point.

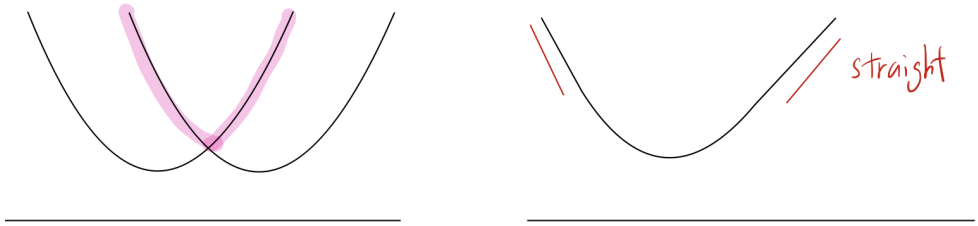

Figure 11.3: Strongly convex but non-smooth function vs smooth function that is not strongly convex

3.2 Optimality gap for strongly convex functions

Recall that Theorem 11.3 measures the optimality gap of any given solution x for a smooth function. We can provide a similar result for bounding the optimality gap for strongly convex functions.

Theorem 11.6. *If $f : \mathbb{R}^d \to \mathbb{R}$ is α-strongly convex in the ℓ_2 norm, then*

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

where x^* is an optimal solution to $\min_{x\in\mathbb{R}^d} f(x)$.

Proof. Let us prove the lower bound on $f(x) - f(x^*)$ first. As f is α-strongly convex, we have

$$f(x) \geq f(x^*) + \nabla f(x^*)^\top (x - x^*) + \frac{\alpha}{2}\|x - x^*\|_2^2,$$

which implies the lower bound as $\nabla f(x^*) = 0$. For the upper bound, note that

$$\begin{aligned} f(x^*) &\geq f(x) + \nabla f(x)^\top (x^* - x) + \frac{\alpha}{2}\|x^* - x\|_2^2 \\ &\geq \min_{y \in \mathbb{R}^d} f(x) + \nabla f(x)^\top (y - x) + \frac{\alpha}{2}\|y - x\|_2^2. \end{aligned}$$

The minimization term above is minimized when y satisfies $\nabla f(x) + \alpha(y - x) = 0$, which is equivalent to $y = x - (1/\alpha)\nabla f(x)$. Therefore,

$$f(x^*) \geq f(x) - \frac{1}{2\alpha}\|\nabla f(x)\|_2^2,$$

as required. □

Lastly, we show the following result for strongly convex functions, which holds because the monotonicity condition for $f(x) - (\alpha/2)\|x\|_2^2$.

Lemma 11.7. *If $f : \mathbb{R}^d \to \mathbb{R}$ is α-strongly convex in the ℓ_2 norm, then*

$$(\nabla f(x) - \nabla f(y))^\top (x - y) \geq \alpha\|x - y\|_2^2$$

for any $x, y \in \mathbb{R}^d$.

Proof. As $g(x) = f(x) - (\alpha/2)\|x\|_2^2$ is convex, the monotonicity of the gradient of g implies that

$$(\nabla g(x) - \nabla g(y))^\top (x - y) \geq 0$$

for any $x, y \in \mathbb{R}^d$. Note that $\nabla g(x) = \nabla f(x) - \alpha x$ and $\nabla g(y) = \nabla f(y) - \alpha y$, which implies that

$$(\nabla g(x) - \nabla g(y))^\top (x - y) = (\nabla f(x) - \nabla f(y))^\top (x - y) - \alpha\|x - y\|_2^2.$$

Then we obtain $(\nabla f(x) - \nabla f(y))^\top (x - y) \geq \alpha\|x - y\|_2^2$, as required. □

4 Convergence result for smooth and strongly convex functions

When f is both smooth and strongly convex, f satisfies the following property.

Lemma 11.8. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex. If f is β-smooth in the ℓ_2 norm and $\beta \geq \alpha$, then $f(x) - (\alpha/2)\|x\|_2^2$ is $(\beta - \alpha)$-smooth.*

Proof. By Theorem 11.1, $f(x) - (\alpha/2)\|x\|_2^2$ is $(\beta - \alpha)$-smooth if and only if

$$\frac{\beta - \alpha}{2}\|x\|_2^2 - \left(f(x) - \frac{\alpha}{2}\|x\|_2^2\right) = \frac{\beta}{2}\|x\|_2^2 - f(x)$$

is convex. Then, again, $(\beta/2)\|x\|_2^2 - f(x)$ is convex if and only if f is β-smooth. Since f is β-smooth, it follows that $f(x) - (\alpha/2)\|x\|_2^2$ is $(\beta - \alpha)$-smooth, as required. □

Based on Lemma 11.8, we can prove the following result on functions that are both smooth and strongly convex.

Lemma 11.9. *If $f : \mathbb{R}^d \to \mathbb{R}$ is β-smooth and α-strongly convex in the ℓ_2 norm, then*

$$(\nabla f(x) - \nabla f(y))^\top (x - y) \geq \frac{1}{\beta + \alpha}\|\nabla f(x) - \nabla f(y)\|_2^2 + \frac{\alpha\beta}{\beta + \alpha}\|x - y\|_2^2$$

for any $x, y \in \mathbb{R}^d$.

Proof. Since f is α-strongly convex, $f(x) - (\alpha/2)\|x\|_2^2$ is convex. Moreover, $f(x) - (\alpha/2)\|x\|_2^2$ is $(\beta - \alpha)$-smooth by Lemma 11.8. Applying Lemma 11.4 to $f(x) - (\alpha/2)\|x\|_2^2$, it follows that

$$\begin{aligned}
&\left(\nabla f(x) - \nabla f(y)\right)^\top (x - y) - \alpha\|x - y\|_2^2 \\
&\geq \frac{1}{\beta - \alpha}\|\nabla f(x) - \nabla f(y)\|_2^2 - \frac{2\alpha}{\beta - \alpha}\left(\nabla f(x) - \nabla f(y)\right)^\top (x - y) + \frac{\alpha^2}{\beta - \alpha}\|x - y\|_2^2.
\end{aligned}$$

This implies that

$$\left(\nabla f(x) - \nabla f(y)\right)^\top (x - y) \geq \frac{1}{\beta + \alpha}\|\nabla f(x) - \nabla f(y)\|_2^2 + \frac{\alpha\beta}{\beta + \alpha}\|x - y\|_2^2,$$

as required. □

Observe that Lemma 11.9 is a combination of Lemma 11.4 for smooth functions and Lemma 11.7 for strongly convex functions. This strong peroperty of functions that are both smooth and strongly convex leads to a linear convergence of gradient descent.

Theorem 11.10. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be β-smooth and α-strongly convex in the ℓ_2-norm, and let $\{x_t : t = 1, \ldots, T+1\}$ be the sequence of iterates generated by gradient descent with sep size $\eta_t = 2/(\alpha + \beta)$ for each t. Then*

$$f(x_{T+1}) - f(x^*) \leq \frac{\beta}{2}\exp\left(-\frac{4T}{\kappa + 1}\right)\|x_1 - x^*\|_2^2$$

where x^* is an optimal solution to $\min_{x \in \mathbb{R}^d} f(x)$.

Proof. Let $\eta_t = \eta$ for each $t \geq 1$. Note that

$$\begin{aligned}
\|x_{t+1} - x^*\|_2^2 &= \|x_t - \eta\nabla f(x_t) - x^*\|_2^2 \\
&= \|x_t - x^*\|_2^2 - 2\eta\nabla f(x_t)^\top (x_t - x^*) + \eta^2\|\nabla f(x_t)\|_2^2 \\
&\leq \|x_t - x^*\|_2^2 - \frac{2\eta}{\alpha + \beta}\|\nabla f(x_t)\|_2^2 - \frac{2\eta\alpha\beta}{\alpha + \beta}\|x_t - x^*\|_2^2 + \eta^2\|\nabla f(x_t)\|_2^2 \\
&= \left(1 - \frac{2\eta\alpha\beta}{\alpha + \beta}\right)\|x_t - x^*\|_2^2 + \left(\eta^2 - \frac{2\eta}{\alpha + \beta}\right)\|\nabla f(x_t)\|_2^2
\end{aligned}$$

where the inequality follows from Lemma 11.9. Setting $\eta = 2/(\alpha + \beta)$, we obtain

$$\|x_{t+1} - x^*\|_2^2 \leq \left(\frac{\beta - \alpha}{\beta + \alpha}\right)^2\|x_t - x^*\|_2^2 = \left(\frac{\kappa - 1}{\kappa + 1}\right)^2\|x_t - x^*\|_2^2,$$

which implies that

$$\|x_{t+1} - x^*\|_2 \leq \left(\frac{\kappa - 1}{\kappa + 1}\right)\|x_t - x^*\|_2 \leq \left(\frac{\kappa - 1}{\kappa + 1}\right)^t\|x_1 - x^*\|_2.$$

Since f is β-smooth, we have

$$f(x_{t+1}) - f(x^*) \leq \frac{\beta}{2}\|x_{t+1} - x^*\|_2^2 \leq \frac{\beta}{2}\left(\frac{\kappa - 1}{\kappa + 1}\right)^{2t}\|x_1 - x^*\|_2^2.$$

Lastly,

$$\left(\frac{\kappa - 1}{\kappa + 1}\right)^{2t} = \left(1 - \frac{2}{\kappa + 1}\right)^{2t} \leq \exp\left(-\frac{4t}{\kappa + 1}\right),$$

as required. □

5 Projected gradient descent

So far, we considered gradient descent for unconstrained convex minimization under various settings. Gradient descent proceeds with the update rule

$$x_{t+1} = x_t - \eta_t \nabla f(x_t).$$

If f is not differentiable, we may take a subgradient $g \in \partial f(x_t)$ at x_t instead of the gradient.

For the constrained case, however, the update rule does not necessarily generate a feasible solution. A natural fix for this is that we take the projection of the point $x_t - \eta_t \nabla f(x_t)$ onto the feasible set

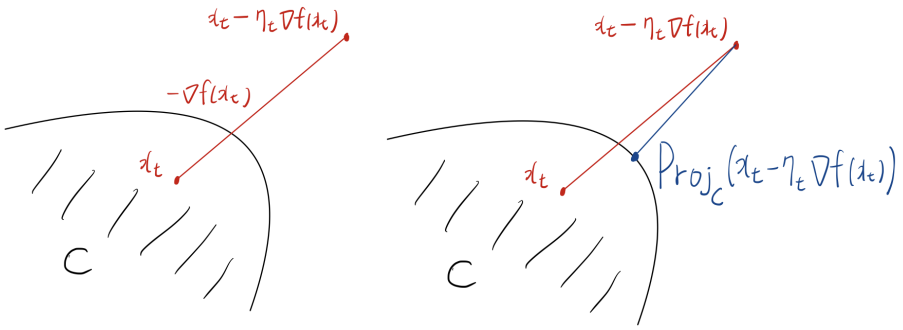

Figure 11.4: Infeasible point after a gradient descent update and projection

C. What we have just discussed is basically the projected gradient descent method! It is basically gradient descent with projection. To formalize, let us give a pseudo-code of the projected gradient descent method. In Algorithm 1, we use the operator $\mathrm{Proj}_C(\cdot)$, which is formally defined as

Algorithm 1 Projected gradient descent method

Initialize $x_1 \in C$.
for $t = 1, \ldots, T$ **do**
  $x_{t+1} = \mathrm{Proj}_C \{x_t - \eta_t \nabla f(x_t)\}$ for a step size $\eta_t > 0$.
end for
Return x_{T+1}.

$$\mathrm{Proj}_C(z) = \operatorname*{argmin}_{x \in C} \frac{1}{2}\|x - z\|_2^2 \quad \text{for } z \in \mathbb{R}^d.$$

Then it is straightforward that

$$\mathrm{Proj}_C(z) = \operatorname*{argmin}_{x \in C} \|x - z\|_2,$$

and in words, $\mathrm{Proj}_C(z)$ is a point C that is closest to point z with respect to the ℓ_2 norm distance. Although we have discussed the following lemma in a previous lecture, we include it again to make this note self-contained.

Lemma 11.11. *Let $x \in C$ and $z \in \mathbb{R}^d$. Then*

$$(Proj_C(z) - z)^\top (Proj_C(z) - x) \leq 0 \quad \textit{for all } x \in C.$$

Proof. We can apply the optimality condition to the definition $\mathrm{Proj}_C(z) = \mathrm{argmin}_{x\in C} \frac{1}{2}\|x - z\|_2^2$ for $z \in \mathbb{R}^d$. The gradient of $\frac{1}{2}\|x-z\|_2^2$ at $x = \mathrm{Proj}_C(z)$ is $(\mathrm{Proj}_C(z) - z)$. Then the statement is precisely the optimality condition for $\mathrm{Proj}_C(z)$. □

By definition, x_{t+1} is the point in C that is closest to $x_t - \eta_t \nabla f(x_t)$ with respect to the ℓ_2 distance. Moreover, we have another interpretation of the update rule based on the following.

$$\begin{aligned} x_{t+1} &= \operatorname*{argmin}_{x\in C} \left\{ \frac{1}{2}\|x - x_t + \mu_t \nabla f(x_t)\|_2^2 \right\} \\ &= \operatorname*{argmin}_{x\in C} \left\{ f(x_t) + \nabla f(x_t)^\top (x - x_t) + \frac{1}{2\eta_t}\|x - x_t\|_2^2 \right\}, \end{aligned}$$

which means that x_{t+1} is the solution in C minimizing the quadratic approximation of f at x_t.

Hereinafter, we introduce notations y_{t+1} to denote $x_t - \eta_t \nabla f(x_t)$ for simpler presentations. Then the update rule can be written as

$$\begin{aligned} y_{t+1} &= x_t - \eta_t \nabla f(x_t), \\ x_{t+1} &= \mathrm{Proj}_C(y_{t+1}) \end{aligned}$$

for $t = 1, \ldots, T$. The analysis of projected gradient descent is quite similar to that of gradient descent for unconstrained minimization. The following is useful to make the analysis for gradient descent go through for the case of projected gradient descent.

Lemma 11.12. *For any t, we have*

$$\|x_{t+1} - x^*\|_2 \le \|y_{t+1} - x^*\|_2$$

where x^* is an optimal solution to $\min_{x\in C} f(x)$.

Proof. We use Lemma 11.11 and the fact that $x_{t+1} = \mathrm{Proj}_C(y_{t+1})$. By Lemma 11.11,

$$(x_{t+1} - y_{t+1})^\top (x_{t+1} - x^*) \le 0.$$

Since $x_{t+1} - y_{t+1} = x_{t+1} - x^* + x^* - y_{t+1}$, the inequality implies that

$$\|x_{t+1} - x^*\|_2^2 \le (y_{t+1} - x^*)^\top (x_{t+1} - x^*) \le \|y_{t+1} - x^*\|_2 \|x_{t+1} - x^*\|_2$$

where the last inequality is due to the Cauchy-Schwarz inequality. Dividing each side by $\|x_{t+1} - x^*\|_2$, we obtain the result. □

By Lemma 11.12, we deduce that

$$\begin{aligned} \|x_{t+1} - x^*\|_2^2 &\le \|y_{t+1} - x^*\|_2 \\ &= \|x_t - x^*\|_2^2 - 2\eta_t \nabla f(x_t)^\top (x_t - x^*) + \eta_t^2 \nabla f(x_t)^2 \\ &\le \|x_t - x^*\|_2^2 - 2\eta_t (f(x_t) - f(x^*)) + \eta_t^2 \nabla f(x_t)^2, \end{aligned}$$

which appears in the convergence analysis of gradient descent for Lipschitz continuous functions. Note that the only difference from the unconstrained case is the first inequality, which used to be an equality for the unconstrained case where $y_{t+1} = x_{t+1}$. Based on this, we recover the same convergence theorem for projected gradient descent for the case of Lipschitz continuous functions. In fact, we can work over the projected subgradient method, which is as the name suggests the subgradient method with projection for the constrained minimization.

The following theorem shows the convergence of the projected subgradient method for functions that have bounded subgradients.

Algorithm 2 Projected subgradient method

Initialize $x_1 \in C$.
for $t = 1, \ldots, T$ **do**
  Obtain a subgradient $g_t \in \partial f(x_t)$.
  $x_{t+1} = \text{Proj}_C \{x_t - \eta_t g_t\}$ for a step size $\eta_t > 0$.
end for
Return $(\sum_{t=1}^T \eta_t)^{-1} \sum_{t=1}^T \eta_t x_t$.

Theorem 11.13. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function such that $\|g\|_2 \leq L$ for any $g \in \partial f(x)$ for every $x \in \mathbb{R}^d$. Let $\{x_t : t = 1, \ldots, T\}$ be the sequence of iterates generated by the projected subgradient method with step size $\eta_t = \|x_1 - x^*\|_2 / L\sqrt{T}$ for each t. Then*

$$f\left(\frac{1}{T}\sum_{t=1}^T x_t\right) - f(x^*) \leq \frac{L\|x_1 - x^*\|_2}{\sqrt{T}}$$

where x^* is an optimal solution to $\min_{x \in C} f(x)$.

Moreover, we also recover the same "asymptotic" convergence rate for strongly convex, smooth, and strongly convex & smooth functions. In particular,

Theorem 11.14. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a β-smooth convex function, and let $\{x_t : t = 1, \ldots, T\}$ be the sequence of iterates generated by gradient descent with step size $\eta_t = 1/\beta$ for each t. Then*

$$f(x_T) - f(x^*) \leq \frac{3\beta\|x_1 - x^*\|_2^2 + f(x_1) - f(x^*)}{T}$$

where x^* is an optimal solution to $\min_{x \in C} f(x)$.