3341.454: Mathematical and Numerical Optimization — **SNU, Spring 2026**

Lecture #10: Strongly Convex Functions and Gradient Descent — April 6, 2026

Lecturer: Dabeen Lee

1 Outline

In this lecture, we study

- strongly convex functions,
- convergence analysis of gradient descent for strongly convex functions, and
- smooth and strongly convex functions.

2 Strongly convex functions

2.1 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 10.1. *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. □

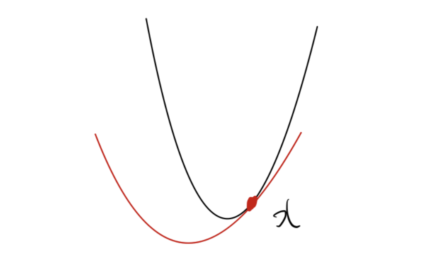

Figure 10.1: Quadratic lower bound on a strongly convex function

Lemma 10.1 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 10.2, 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.

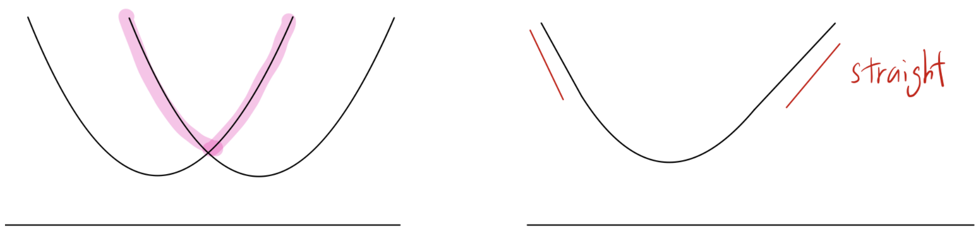

Figure 10.2: Strongly convex but non-smooth function vs smooth function that is not strongly convex

2.2 Optimality gap for strongly convex functions

We saw a theorem on 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 10.2. *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. $\square$

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 10.3. *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. □

3 Gradient descent for strongly convex functions

3.1 Strongly convex functions

Theorem 10.4. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be L-Lipschitz and α-strongly convex in the ℓ_2 norm for some $L, \alpha > 0$, and let $\{x_t : t = 1, \ldots, T+1\}$ be the sequence of iterates generated by gradient descent with step size $\eta_t = 2/(\alpha(t+1))$ for each t. Then*

$$f\left(\sum_{t=1}^{T} \frac{2t}{T(T+1)} x_t\right) - f(x^*) \leq \frac{2L^2}{\alpha(T+1)},$$

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

Proof. Let $\eta_t = 2/(\alpha(t+1))$ for each $t \geq 1$. Note that

$$\begin{aligned}
\|x_{t+1} - x^*\|_2^2 &= \|x_t - \eta_t \nabla f(x_t) - x^*\|_2^2 \\
&= \|x_t - x^*\|_2^2 - 2\eta_t \nabla f(x_t)^\top (x_t - x^*) + \eta_t^2 \|\nabla f(x_t)\|_2^2 \\
&\leq \|x_t - x^*\|_2^2 - 2\eta_t (f(x_t) - f(x^*)) - \alpha \eta_t \|x_t - x^*\|_2^2 + \eta_t^2 L^2 \\
&= (1 - \alpha \eta_t) \|x_t - x^*\|_2^2 + \eta_t^2 L^2 - 2\eta_t (f(x_t) - f(x^*)).
\end{aligned}$$

Rearranging the above inequality, we obtain

$$f(x_t) - f(x^*) \leq \frac{\eta_t}{2} L^2 + \left(\frac{1}{2\eta_t} - \frac{\alpha}{2}\right) \|x_t - x^*\|_2^2 - \frac{1}{2\eta_t} \|x_{t+1} - x^*\|_2^2.$$

Multiplying both sides of the above inequality by t, we have

$$t(f(x_t) - f(x^*)) \leq \frac{L^2}{\alpha} + \frac{(t-1)t}{4} \alpha \|x_t - x^*\|_2^2 - \frac{t(t+1)}{4\alpha} \|x_{t+1} - x^*\|_2^2.$$

Adding the above inequality for $t = 1, \ldots, T$, we obtain

$$\sum_{t=1}^{T} t(f(x_t) - f(x^*)) \leq \frac{L^2 T}{\alpha} - \frac{T(T+1)}{4\alpha} \|x_{T+1} - x^*\|_2^2 \leq \frac{L^2 T}{\alpha},$$

where the last step follows from the non-negativity of $\|x_{T+1} - x^*\|_2^2$. Therefore,

$$f\left(\sum_{t=1}^{T} \frac{2t}{T(T+1)} x_t\right) - f(x^*) \leq \sum_{t=1}^{T} \frac{2t}{T(T+1)} (f(x_t) - f(x^*)) \leq \frac{2L^2}{\alpha(T+1)},$$

as required. □

3.2 Smooth and strongly convex functions

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

Lemma 10.5. *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. We know that $f(x) - (\alpha/2)\|x\|_2^2$ 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 10.5, we can prove the following result on functions that are both smooth and strongly convex.

Lemma 10.6. *If $f : \mathbb{R}^d \to \mathbb{R}$ is β-smooth and α-strongly convex in the ℓ_2 norm, then*

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

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 10.5. By the co-coercivity of $f(x) - (\alpha/2)\|x\|_2^2$, 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 10.6 is a combination of the co-coercivity for smooth functions and Lemma 10.3 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 10.7. *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 step 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 10.6. 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. □

4 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 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

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

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.

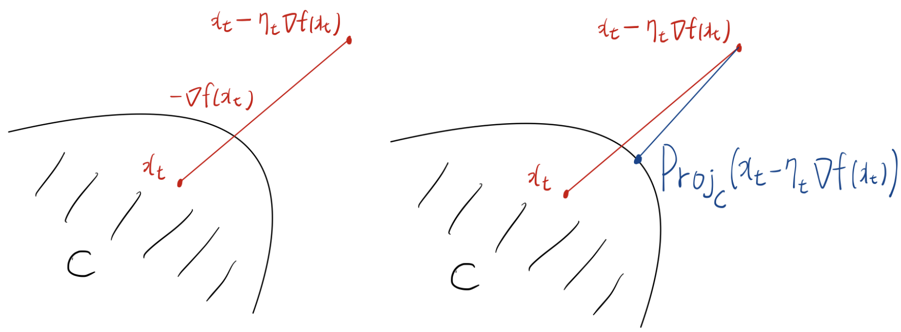

Figure 10.3: Infeasible point after a gradient descent update and projection

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}.

Lemma 10.8. *Let $x \in C$ and $z \in \mathbb{R}^d$. Then*

$$(\mathit{Proj}_C(z) - z)^\top (\mathit{Proj}_C(z) - x) \le 0 \quad \text{for all } x \in C.$$

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.

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

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 10.9. *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 10.8 and the fact that $x_{t+1} = \mathrm{Proj}_C(y_{t+1})$. By Lemma 10.8,

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

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

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

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 10.9, we deduce that

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

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.

Algorithm 2 Projected subgradient method

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

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

Theorem 10.10. *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 10.11. *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 $\min_{x \in C} f(x)$.