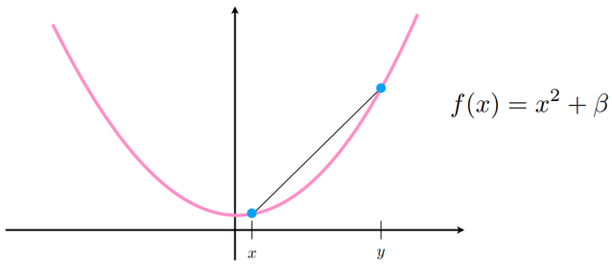

Fig. 15. $f(x) = x^2 + \beta$, for $\beta > 0$.

Chapter 3

We have introduced gradient descent, and we have studied its performance under Lipschitz gradient continuity. This lecture introduces the basic notions of convexity in optimization. We will discuss convex functions, convex constraints, and whether gradient descent is benefited by convexity. Apart from standard convexity, we will also introduce the notion of *strong convexity*, and discuss what is its effect in practice and theory.
This chapter continuous to evolve around convergence rates, and contains some discussion about lower bounds on such rates.

Convexity | Gradient Descent | Strong convexity | Other global assumptions | Projection onto convex sets

(*The discussion in this chapter will mostly focus on the unconstrained case:* $\min_x f(x)$*, unless otherwise stated.*)

Convexity. A key consequence of convexity is that any local solution is global in convex optimization. To understand convexity in functions, we will cover some definitions first.

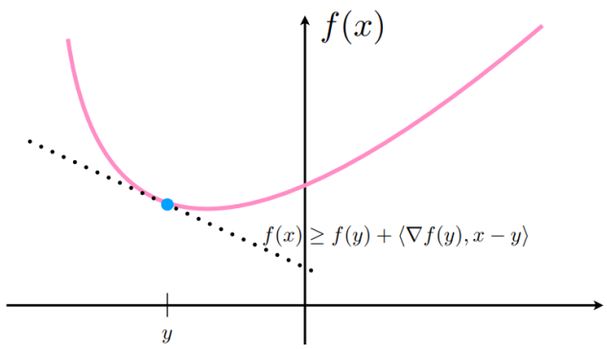

Fig. 16. Convex interpretation via gradients.

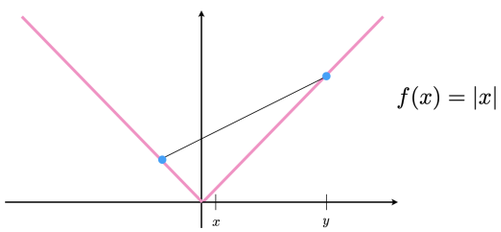

Fig. 17. $f(x) = |x|$.

Definition 16. (Convex Function) *$f : \mathbb{R} \to \mathbb{R}$ is a univariate convex function if, for $\forall \alpha \in [0, 1]$, the following holds:*

$$f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha) f(y), \quad \forall x, y.$$

This basically dictates that the function value of f at any point in a given interval $[x, y]$ is lower than any secant connecting two points within that interval; see Figure 15.

Alternatively, a convex function can be defined as a function that lies above any (hyper)plane that is tangential to f at any point. Using the gradient $\nabla f(x)$, this is interpreted as:

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle.$$

and depicted in low-dimensions as in Figure 16.

But what if f is not uniquely differentiable? I.e., what if f does not have a unique gradient at all points, but there are points where we can compute a set of gradients, *the subgradients* $\partial f(x)$? The same ideas apply in that case; see the Figure 17 for the case of $f(x) = |x|$. In this example, f has a set of subgradients $\partial f(x)$ at point $x = 0$. These subgradients could take any value in the interval $[-1, 1]$; i.e., the set of all lines that touch $(0, f(0))$ with slope between -1 and 1. In general, for convex f and any subgradient $g \in \partial f(y)$ we have:

$$f(x) \geq f(y) + \langle g, x - y \rangle.$$

The opposite (substitute $\leq$ with $\geq$) is a concave function.

A very useful inequality for convex functions is Jensen's inequality:

Lemma 2. *For a convex function f, Jensen's inequality states:*

$$f\left(\mathbb{E}\left[x\right]\right) \leq \mathbb{E}\left[f(x)\right].$$

The geometric interpretation of Jensen's inequality that relates to convex functions is that *the function value of the average of two points is less than the average of the function values of the two points*, i.e.,

$$f\left(\tfrac{x_1 + x_2}{2}\right) \leq \tfrac{1}{2} f(x_1) + \tfrac{1}{2} f(x_2).$$

What are some good examples of convex functions that we observe in practice? Examples are shown in Table 1.

Table 1. Convexity of common functions

Function	Example	Attributes
ℓ_p norms $p \geq 1$	$\|x\|_2$, $\|x\|_1$, $\|x\|_\infty$	convex
ℓ_p matrix norms $p \geq 1$	$\|X\|_2$	convex
Square root function	$\sqrt{x}$	concave
Maximum	$\max x_1, \cdots, x_n$	convex
Minimum	$\min x_1, \cdots, x_n$	concave
Sum of convex functions		convex
Logarithmic functions	log(det(X))	convex if $X \succeq 0$
Affine/linear functions	$\sum_{i=1}^{N} X_i i$	convex and concave
Eigenvalue functions	$\lambda_{\max}(X)$	convex if $X = X^\top$

Properties of convex functions. There are several alternative, and potentially more practical, definitions of a convex function:

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle, \quad \forall x, y$$
$$\langle \nabla f(y) - \nabla f(x), y - x \rangle \geq 0, \quad \forall x, y$$
$$\nabla^2 f(x) \succeq 0, \quad \forall x.$$

Key property of convex functions is the following lemma.

[3] Remember there might be multiple equivalent global minima, but we can assume we are converging to one of them.

Lemma 3. *Any stationary point of a convex function f is a global minimum.*

Proof: Assume that $f : \mathbb{R}^p \to \mathbb{R}$. Let $\widehat{x}$ denote a stationary point of f, where $\nabla f(\widehat{x}) = 0$. Since f is a convex function, we know that:

$$f(x) \geq f(\widehat{x}) + \langle \nabla f(\widehat{x}), x - \widehat{x} \rangle = f(\widehat{x}), \;\; \forall x,$$

where the last equality is due to $\nabla f(\widehat{x}) = 0$. However, the above holds for all x, and thus for $x^\star$ which is/are the global minimum/minima. Thus, we have:

$$f(x^\star) \geq f(\widehat{x}), \;\; \forall x^\star,$$

which is a contradiction. This implies that all stationary points $\widehat{x}$ are equivalent to global minimum. □

While this fact provides hope for finding the global minimum, it does not guarantee tractability and practicality of the proposed algorithms.

Does convexity help convergence rate. We will study whether convexity improves the convergence rate of gradient descent. We remind the reader that, for a differentiable function f with gradient $\nabla f(\cdot)$, gradient descent satisfies:

$$x_{t+1} = x_t - \eta_t \nabla f(x_t), \quad t = 0, 1, \dots$$

We will make the same baseline assumptions as before: we will assume that f has Lipschitz continuous gradients:

$$\|\nabla f(x_1) - \nabla f(x_2)\|_2 \leq L \cdot \|x_1 - x_2\|_2, \;\; \forall x_1, x_2.$$

The only additional assumption we make is that f is also convex.

We will follow a different perspective—let $x^\star$ denote a global minimum.[3] Further, assume we use a constant step size $\eta_t = \eta$. Then, the following equality holds:

$$\begin{aligned}\|x_{t+1} - x^\star\|_2^2 &= \|x_t - \eta \nabla f(x_t) - x^\star\|_2^2 \\ &= \|x_t - x^\star\|_2^2 + \eta^2 \|\nabla f(x_t)\|_2^2 \\ &\quad - 2\eta \langle \nabla f(x_t),\; x_t - x^\star \rangle\end{aligned}$$

We can show that Lipschitz gradient continuity, along with convexity, leads to the following inequality:

$$\tfrac{1}{L} \cdot \|\nabla f(x_1) - \nabla f(x_2)\|_2^2 \leq \langle \nabla f(x_1) - \nabla f(x_2), x_1 - x_2 \rangle .$$

Substituting $x_1 \equiv x^\star$, $x_2 \equiv x_t$, and assuming $\nabla f(x^\star) = 0$ in the above inequality, we get:

$$\begin{aligned}\tfrac{1}{L} \cdot \|\nabla f(x_t)\|_2^2 \leq \langle -\nabla f(x_t), x^\star - x_t \rangle \quad &\Rightarrow \\ \langle \nabla f(x_t), x_t - x^\star \rangle \geq \tfrac{1}{L} \cdot \|\nabla f(x_t)\|_2^2 \quad &\Rightarrow \\ -2\eta \langle \nabla f(x_t), x_t - x^\star \rangle \leq -\tfrac{2\eta}{L} \cdot \|\nabla f(x_t)\|_2^2 &\end{aligned}$$

Combining with the above, we obtain:

$$\|x_{t+1} - x^\star\|_2^2 \leq \|x_t - x^\star\|_2^2 - \eta \left(\tfrac{2}{L} - \eta\right) \cdot \|\nabla f(x_t)\|_2^2.$$

Assuming $0 < \eta < \frac{2}{L}$, the second term on the right hand side is negative. This implies that per iteration, we decrease the distance to optimum as in:

$$\|x_{t+1} - x^\star\|_2^2 \leq \|x_t - x^\star\|_2^2 \leq \dots \leq \|x_0 - x^\star\|_2^2.$$

(*Question: Would such a statement hold for non-convex scenarios? Under which conditions?*)

By the analysis of the previous—not necessarily convex—case in the previous chapter, we also know that:

$$f(x_{t+1}) \leq f(x_t) - \eta \left(1 - \tfrac{L}{2}\eta\right) \cdot \|\nabla f(x_t)\|_2^2$$

By convexity, we also have:

$$\begin{aligned}f(x^\star) \geq f(x_t) + \langle \nabla f(x_t), x^\star - x_t \rangle \quad &\Rightarrow \\ f(x_t) - f(x^\star) \leq \langle \nabla f(x_t), x_t - x^\star \rangle & \\ \leq \|x_t - x^\star\|_2 \cdot \|\nabla f(x_t)\|_2 & \\ \leq \|x_0 - x^\star\|_2 \cdot \|\nabla f(x_t)\|_2 &\end{aligned}$$

where the last inequality is based on the previous observation that $\|x_{t+1} - x^\star\|_2 \leq \|x_0 - x^\star\|_2$.

Then, we can combine the above into:

$$\begin{aligned}[f(x_{t+1}) - f(x^\star)] &\leq [f(x_t) - f(x^\star)] - \eta \left(1 - \tfrac{L}{2}\eta\right) \cdot \|\nabla f(x_t)\|_2^2 \\ &\leq [f(x_t) - f(x^\star)] - \eta \left(1 - \tfrac{L}{2}\eta\right) \cdot \tfrac{\left[f(x_t) - f(x^\star)\right]^2}{\|x_0 - x^\star\|_2^2}\end{aligned}$$

Define $\Delta_t := f(x_t) - f(x^\star)$. Then:

$$\begin{aligned}&\Delta_{t+1} \leq \Delta_t - \tfrac{\eta\left(1 - \frac{L}{2}\eta\right)}{\|x_0 - x^\star\|_2^2} \cdot \Delta_t^2 = \Delta_t \cdot \left(1 - \tfrac{\eta\left(1 - \frac{L}{2}\eta\right)}{\|x_0 - x^\star\|_2^2} \cdot \Delta_t\right) \quad \Rightarrow \\ &\tfrac{\Delta_{t+1}}{\Delta_t} \leq 1 - \tfrac{\eta\left(1 - \frac{L}{2}\eta\right)}{\|x_0 - x^\star\|_2^2} \cdot \Delta_t \Rightarrow \\ &\tfrac{1}{\Delta_{t+1}} \geq \tfrac{1}{\Delta_t} + \tfrac{\eta\left(1 - \frac{L}{2}\eta\right)}{\|x_0 - x^\star\|_2^2} \cdot \tfrac{\Delta_t}{\Delta_{t+1}} \geq \tfrac{1}{\Delta_t} + \tfrac{\eta\left(1 - \frac{L}{2}\eta\right)}{\|x_0 - x^\star\|_2^2},\end{aligned}$$

for a step size $\eta = \frac{1}{L}$.

Unfolding the recursion for T iterations:

$$\tfrac{1}{\Delta_T} \geq \tfrac{1}{\Delta_0} + \tfrac{\eta\left(1 - \frac{L}{2}\eta\right)}{\|x_0 - x^\star\|_2^2} \cdot T,$$

which leads to:

$$\begin{aligned}f(x_T) - f(x^\star) &\leq \frac{2L\left(f(x_0) - f(x^\star)\right) \cdot \|x_0 - x^\star\|_2^2}{2L\|x_0 - x^\star\|_2^2 + T \cdot \left(f(x_0) - f(x^\star)\right)} \\ &= O\left(\frac{1}{T}\right).\end{aligned}$$

The last expression is because all the other quantities are constant and depend on the initialization. Another way to interpret the above result is that, if we require $f(x_T) - f(x^\star) \leq \varepsilon$, we have to perform $O\left(\frac{1}{\varepsilon}\right)$ number of iterations.

How does this compare to the result we already know? Remember that assuming only Lipschitz gradient continuity, we have:

$$\min_t \|\nabla f(x_t)\|_2 = O\left(\tfrac{1}{\sqrt{T}}\right),$$

and we need $O\left(1/\varepsilon^2\right)$ iterations to achieve $\min_t \|\nabla f(x_t)\|_2 \leq \varepsilon$. This reveals that convexity gains are two-fold: *i*) gradient descent over convex functions lead to convergence to global minimum/minima, not just stationary points; *ii*) gradient descent over convex functions effectively shows improved performance, compared to gradient descent over only L-smooth functions.

Beyond boilerplate convexity: strong convexity. It is possible to achieve better convergence rates by assuming more than just convexity for f. *Strong convexity* is one such assumption that can lead to an improved result. In plain words, it implies that f should be steep enough so that gradient descent can make progress (more aggressively). To see this, let us first provide its definition:

Definition 17. (Strong Convexity) *A function $f : \mathbb{R}^p \to \mathbb{R}$ is a strongly convex function if it is convex and, for $\mu > 0$, satisfies:*

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

(*In the recent optimization literature, following a "machine learning" notation, L is usually substituted with β and μ with α. Here, we will follow the notation that Nesterov has used.*)

A visual illustration of strong convexity is provided in the next figure.

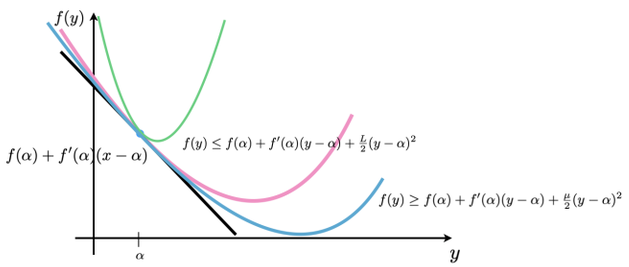

Fig. 18. Strong convexity interpretation and its relation to Lipschitz gradient continuity.

To interpret this further, while Lipschitz gradient continuity implies that, at any point of the domain of f, we can upper bound f with a quadratic (green curve), strong convexity implies that, at any point of the domain of f, we can lower bound f with a quadratic (blue curve).

A strongly convex function has a unique minimizer. Remember that a convex function has the nice property that every local minimum is a global minimum, but there is no guarantee that the set of global minima is a singleton.

There are several alternative and equivalent characterizations of strong convexity to know:

$$\begin{aligned}
\langle \nabla f(x) - \nabla f(y), x - y \rangle &\geq \mu \|x - y\|_2^2, \\
f(y) &\leq f(x) + \langle \nabla f(x), y - x \rangle \\
&\quad + \frac{1}{2\mu}\|\nabla f(x) - \nabla f(y)\|_2^2, \\
\langle \nabla f(x) - \nabla f(y), x - y \rangle &\leq \frac{1}{\mu}\|\nabla f(x) - \nabla f(y)\|_2^2, \\
\nabla^2 f(x) &\succeq \mu \cdot I.
\end{aligned}$$

(*The convergence rate proof of just a strongly convex function–not necessarily L-smooth–is left for exercise.*)

The L-smooth and μ-strongly convex functions. In convex optimization research, the two classes of convex functions that have attracted the most attention are the set of L-smooth functions (i.e., with Lipschitz continuous gradients), and the set of L-smooth AND μ-strongly convex functions.

To understand what strong convexity adds w.r.t. convergence rates, we study the performance of gradient descent under these assumptions, jointly.

Similar to the proof of L-smooth functions:

$$\begin{aligned}
\|x_{t+1} - x^\star\|_2^2 = \|x_t - x^\star\|_2^2 &+ \eta^2 \|\nabla f(x_t)\|_2^2 \\
&- 2\eta \langle \nabla f(x_t), \ x_t - x^\star \rangle
\end{aligned}$$

A key property of L-smooth and μ-strongly convex functions is the following lemma:

Lemma 4. *Let f satisfy L-smoothness and μ-strongly convexity. Then:*

$$\begin{aligned}
\langle \nabla f(x) - \nabla f(y), \ x - y \rangle &\geq \tfrac{\mu L}{\mu + L}\|x - y\|_2^2 \\
&\quad + \tfrac{1}{\mu + L} \cdot \|\nabla f(x) - \nabla f(y)\|_2^2
\end{aligned}$$

(*We will see how this convex condition will "inspire" similar conditions for non-convex optimization.*)

We use the lemma above, with the substitution $x \equiv x^\star$, $y \equiv x_t$, and knowing that $\nabla f(x^\star) = 0$. This leads to:

$$-\langle \nabla f(x_t), \ x^\star - x_t \rangle \geq \tfrac{\mu L}{\mu + L}\|x_t - x^\star\|_2^2 + \tfrac{1}{\mu + L} \cdot \|\nabla f(x_t)\|_2^2.$$

Using this in the inequality above, we obtain:

$$\begin{aligned}
\|x_{t+1} - x^\star\|_2^2 &\leq \|x_t - x^\star\|_2^2 + \eta^2 \|\nabla f(x_t)\|_2^2 \\
&\quad - \tfrac{2\eta\mu L}{\mu + L}\|x_t - x^\star\|_2^2 - \tfrac{2\eta}{\mu + L} \cdot \|\nabla f(x_t)\|_2^2 \\
&= \left(1 - \tfrac{2\eta\mu L}{\mu + L}\right) \cdot \|x_t - x^\star\|_2^2 \\
&\quad + \eta \cdot \left(\eta - \tfrac{2}{\mu + L}\right) \cdot \|\nabla f(x_t)\|_2^2
\end{aligned}$$

Here, assuming that $\eta \leq \frac{2}{\mu + L}$, the second term on the right hand side is ≤ 0; i.e., we can guarantee that the distance to $x^\star$ decreases per iteration, since the first term has $\left(1 - \frac{2\eta\mu L}{\mu + L}\right) < 1$. However, this does not say anything about the convergence rate. For that, we observe that:

$$\|x_{t+1} - x^\star\|_2^2 \leq \left(1 - \tfrac{2\eta\mu L}{\mu + L}\right) \cdot \|x_t - x^\star\|_2^2$$

Assume we use $\eta = \frac{2}{\mu + L}$. Then, we observe:

- $\left(1 - \frac{2\eta\mu L}{\mu + L}\right) = \left(1 - \frac{2\frac{2}{\mu + L}\mu L}{\mu + L}\right) = \left(1 - \frac{4}{(\mu + L)^2}\right) \geq 0.$
- $\frac{2\frac{2}{\mu + L}\mu L}{\mu + L} = \frac{4\mu L}{(\mu + L)^2} = \frac{4}{\frac{\mu}{L} + 2 + \frac{L}{\mu}} \geq \frac{2}{\kappa + 1}.$

where $\kappa := \frac{L}{\mu} > 1$ is defined as the condition number of f. Then:

$$\begin{aligned}
\|x_T - x^\star\|_2^2 &\leq \left(1 - \tfrac{2}{\kappa + 1}\right) \cdot \|x_{T-1} - x^\star\|_2^2 \\
&\leq \left(1 - \tfrac{2}{\kappa + 1}\right)^T \cdot \|x_0 - x^\star\|_2^2 \\
&= \left(\tfrac{\kappa - 1}{\kappa + 1}\right)^T \cdot \|x_0 - x^\star\|_2^2 \\
&= O\left(c^T\right) \cdot \|x_0 - x^\star\|_2^2
\end{aligned}$$

for $c < 1$ constant. This is what we call *linear convergence rate.*

To compare the number of iterations required to get to an ε-close solution, we get:

$$\begin{aligned}
\|x_T - x^\star\|_2^2 \leq \varepsilon \overset{\text{Requires}}{\longrightarrow} \left(\tfrac{\kappa - 1}{\kappa + 1}\right)^T \cdot \|x_0 - x^\star\|_2^2 &\leq \varepsilon \\
T &\geq \frac{\log\left(\|x_0 - x^\star\|_2^2 / \varepsilon\right)}{\log \frac{\kappa + 1}{\kappa - 1}}.
\end{aligned}$$

Comparing to just L-smooth convex functions, we have:

$$O\left(\tfrac{1}{\varepsilon}\right) \quad \text{vs} \quad O\left(\log \tfrac{1}{\varepsilon}\right).$$

Thus, if we require a solution that is $\varepsilon = 10^{-3}$-close in some sense, for L-smooth functions we require $O(1000)$ iterations, while for strongly convex functions we require $O(3)$ iterations (hiding though a lot of constants). Moreover, observe that the premise in the strongly convex case is stronger: we are guaranteed to converge to the unique global solution, while L-smoothness convex itself cannot guarantee anything about which global solution we converge to.

Please, revisit figures in previous chapters for an illustration and comparison between different convergence rates.

What should our expectations be: Lower bounds. Let us summarize some of the results we got so far, especially under the convexity assumption. We know that:

- For L-smooth convex functions, we have:

$$f(x_T) - f(x^\star) \leq \frac{2L\left(f(x_0) - f(x^\star)\right) \cdot \|x_0 - x^\star\|_2^2}{2L\|x_0 - x^\star\|_2^2 + T \cdot \left(f(x_0) - f(x^\star)\right)}.$$

- When we also have strong convexity:

$$\|x_T - x^\star\|_2^2 \leq \left(\tfrac{\kappa-1}{\kappa+1}\right)^T \cdot \|x_0 - x^\star\|_2^2.$$

But is this the best we can achieve when dealt with only L-smooth convex functions? E.g., is there another analysis that leads to $f(x_T) - f(x^\star) \leq c^T$, for some $c < 1$, when only L-smoothness holds? Can we achieve an even better convergence rate under L-smooth and μ-strong convexity?

The above lead to the discussion on *lower bounds*: i.e., making the same assumptions—and no more—can we construct functions f that, under these assumptions, we cannot achieve something better than the above?[4]

The following summarize lower bounds on the types of objective functions we have previously discussed.

- For objective functions with Lipschitz continuous gradients, with constant L, we can prove that there are f instances such that we cannot achieve something better than:

$$f(x_T) - f(x^\star) \geq \frac{3L\|x_0 - x^\star\|_2^2}{32(T+1)^2} = O\left(\tfrac{1}{T^2}\right).$$

  Under this assumption, and only using gradients, we cannot achieve better than the above.
- For objectives functions with both Lipschitz continuous gradients and strong convexity, there are f instances with convergence rate lower bounds:

$$\|x_T - x^\star\|_2^2 \geq \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2T} \|x_0 - x^*\|_2^2,$$

  where $\kappa = L/\mu > 1$. Here we observe that, while we have achieved the same convergence rate with respect to the exponent—i.e.., in both cases we have c^T, for $c < 1$—in the lower bound case, we see $\sqrt{\kappa}$ instead of κ.

But how we obtain such lower bounds? By constructing special functions f that satisfy our assumptions and that provably show such lower bounds behavior in theory.[5]

Later in the course, we will see how we can achieve these lower bounds, under the same assumptions, and by relying only on a first-order oracle.

Other powerful global assumptions. Convexity is a strong assumption that dictates every local minimum be equivalent to a global minimum. In math, along with L-smoothness and strong convexity, we use the basic condition:

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle,$$

to obtain the results above. *But is there any other assumptions that we can use to prove similar convergence rates?*

In this subsection, we will focus on the notion of *Polyak-Lojasiewicz (PL) inequality*, use it in proof techniques, and conclude with other global assumptions, similar to PL.

The definition of PL is as follows:

Definition 18. *A function f satisfies the PL inequality, if the following holds for some $\xi > 0$:*

$$\tfrac{1}{2}\|\nabla f(x)\|_2^2 \geq \xi \cdot \left(f(x) - f(x^\star)\right), \quad \forall x.$$

(*Any thoughts what this inequality implies, with respect to stationary points?*)

Let us use this new definition to prove convergence. We will assume L-smoothness of f (*this does not imply anything about convexity*). Using step size $\eta = \frac{1}{L}$ in gradient descent leads to:

$$f(x_{t+1}) - f(x_t) \leq -\tfrac{1}{2L} \cdot \|\nabla f(x_t)\|_2^2, \forall t.$$

By PL, we know that:

$$-\tfrac{1}{2}\|\nabla f(x_t)\|_2^2 \leq -\xi \cdot \left(f(x_t) - f(x^\star)\right).$$

Then:

$$\begin{aligned}
&f(x_{t+1}) - f(x_t) \leq -\tfrac{\xi}{L} \cdot \left(f(x_t) - f(x^\star)\right) \quad \Rightarrow \\
&f(x_{t+1}) - f(x^\star) \leq f(x_t) - f(x^\star) - \tfrac{\xi}{L} \cdot \left(f(x_t) - f(x^\star)\right) \quad \Rightarrow \\
&f(x_{t+1}) - f(x^\star) \leq \left(1 - \tfrac{\xi}{L}\right) \cdot \left(f(x_t) - f(x^\star)\right).
\end{aligned}$$

Unfolding this recursion:

$$f(x_T) - f(x^\star) \leq \left(1 - \tfrac{\xi}{L}\right)^T \cdot \left(f(x_0) - f(x^\star)\right).$$

Under the assumption that $L \geq \xi$, this leads to linear convergence rate.

Some comments:

- Observe that we managed to prove linear convergence to the global optimum, without assuming strong convexity. This dictates that there might be different assumptions one can make that lead to favourable behavior.
- Further, PL inequality does not even imply convexity; i.e., we proved convergence with linear rate to the global optimum, even if the objective is not convex.
- PL assumption does not imply uniqueness of the global optimum; there might be several $x^\star$ (*one of the reason we do not have convergence guarantees in $\|x_t - x^\star\|_2$ terms*).

How does a function that satisfies PL inequality look like? Here is an example we have seen in the previous chapter.

[4] This also includes the assumption that we will only use first-order oracles; if we had the option to use more information—say Hessians—then we might be able to achieve more. We will defer this discussion in the chapters to follow.

[5] However, is this a pessimistic way of thinking convergence rates? For the careful reader, this is similar to characterizing a problem NP-hard by the time we find an instance that is NP-hard. Does this hold though for the most practical f cases? Food for thought.

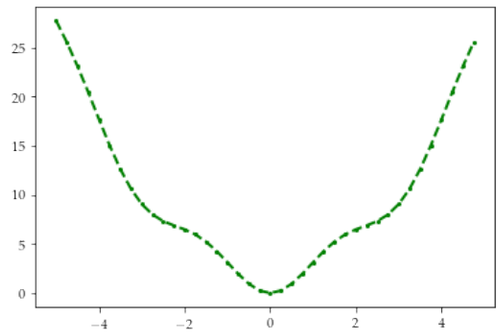

Fig. 19. $f(x) = x^2 + 3\sin^2(x)$

Some other conditions that have been used in convergence proofs, but we will not focus on this chapter, are:

Definition 19. *A function f satisfies the weak strong convexity (WSC) condition, if the following holds for some $\mu > 0$:*

$$f(x^\star) \geq f(x) + \langle \nabla f(x), x^\star - x \rangle + \tfrac{\mu}{2}\|x - x^\star\|_2^2, \quad \forall x.$$

Observe that this inequality holds for only $x^\star$ on the left hand side; this justifies the term "weak" in its name.

Definition 20. *A function f satisfies the restricted secant inequality (RSI), if the following holds for some $\mu > 0$:*

$$\langle \nabla f(x), x - x^\star \rangle \geq \mu\|x - x^\star\|_2^2, \quad \forall x.$$

If the function f is also convex, this is also called the restricted strong convexity.

Definition 21. *A function f satisfies the error bound condition (EB), if the following holds for some $\mu > 0$:*

$$\|\nabla f(x)\|_2 \geq \mu\|x - x^\star\|_2, \quad \forall x.$$

Definition 22. *A function f satisfies the quadratic growth (QG) condition, if the following holds for some $\mu > 0$:*

$$f(x) - f(x^\star) \geq \tfrac{\mu}{2}\|x - x^\star\|_2, \quad \forall x.$$

In the above definitions, μ do not dictate the same value for all definitions; we use the same letter for clarity.

Finally, it turns out that there is a hierarchy on these conditions.

$$\texttt{(WSC)} \to \texttt{(RSI)} \to \texttt{(EB)} \equiv \texttt{(PL)} \to \texttt{(QG)}$$

The above indicate "implications": e.g., if we assume WSC holds for a function, then the rest of the conditions are also satisfied for some constants μ [15].

Constrained convex optimization and convex sets. In Chapter 1, we mentioned that the focus of this class will be a subset of problems of the form:

$$\begin{aligned} \min_{x \in \mathbb{R}^p} \quad & f(x) \\ \text{subject to} \quad & x \in \mathcal{C}. \end{aligned}$$

where $\mathcal{C}$ is the constraint set on x. The nature of $\mathcal{C}$ depends on the application; there are applications where $\mathcal{C}$ is simple enough and does not affect much how gradient descent behaves, and there are applications where $\mathcal{C}$ is not of a straightforward form (e.g., think of combinatorial constraints). One such example is the sparsity constraint, where we are looking for a sparse vector that minimizes f (i.e., there might be dense vectors that minimize f even further, but for some reason we are interested in sparse solutions).

Similarly to functions, we need to define what is the difference between *convex and non-convex sets*. And, in order to understand and appreciate the difficulty of including non-convex constraints, we need to understand how simple, convex constraints affect the performance of gradient descent.

When is our problem convex or non-convex? First, it is important to understand what convex optimization can solve and what it can not. When *both the objective and the constraints are convex*, then the problem (in most cases) can be solved by standard convex optimization tools (including gradient descent as a solver). When *either of the objective or the constraints are non-convex, or neither of them are convex*, then the problem is non-convex.

Just to provide a pictorial explanation, look at the following toy example curve.

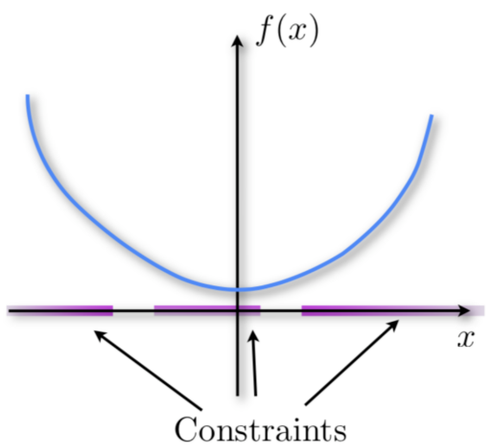

Fig. 20. Constrained optimization example, where the constraint set is non-convex. The purple parts do not belong in the feasibility set.

If we had no constraints, the function is smooth and convex; thus gradient descent would work as expected. However, including the purple parts on the feasibility set, the optimal is no longer the bottom of the "bowl". Also, solving the problem first without the constraints, and then applying the constraints, most of the times does not lead to a good solution.

It is natural to first study constrained convex optimization. The following are additional definitions related to convexity that will become important later in the course.

Definition 23. (Convex Set) *The set $\mathcal{C} \subset \mathbb{R}^p$ is a convex set if $\forall x_1, x_2 \in \mathcal{C}$, it holds that*

$$\forall \alpha \in [0,1] : \alpha x_1 + (1-\alpha)x_2 \in \mathcal{C}.$$

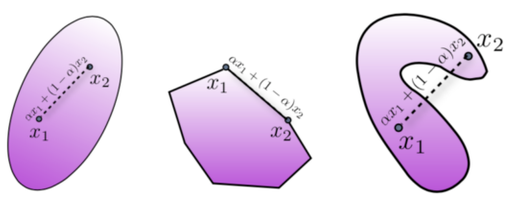

Fig. 21. Some convex set examples.

Definition 24. (Convex Hull) *The convex hull of a set of points in $\mathcal{Q}$ is the intersection of all convex sets containing $\mathcal{Q}$. For n points $\mathcal{Q} := \{x_1, \ldots, x_n\}$, the convex hull is*

$$\text{conv}(\mathcal{Q}) = \left\{ \sum_{j=1}^{n} \alpha_j x_j \ : \alpha_j \geq 0, \ \forall j, \ \sum_{j=1}^{n} \alpha_j = 1 \right\}.$$

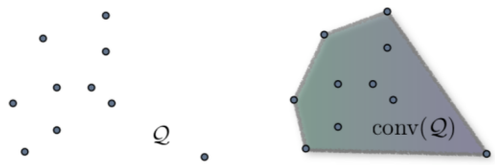

Fig. 22. Convex hull of a set of points.

Some notable convex sets:

- Linear spaces and halfspaces: e.g., $\{x \in \mathbb{R}^p \mid Ax = 0\}$ and $\{x \in \mathbb{R}^p \mid \langle z, x \rangle \geq 0\}$.
- Affine transformations of convex sets: e.g., if $\mathcal{C}$ is a convex set, then so it is the set $\{Ax + b \mid x \in \mathcal{C}\}$.
- Intersections of convex sets.
- Special cases that worth to be mentioned: Norm inequality constraints define convex sets (e.g., $\|x\|_2 \leq 1$, $\|x\|_1 \leq \lambda$, $\|X\|_F \leq c$, $\|y - Ax\|_2 \leq \varepsilon$—*however, the following set* $\|x\|_2 = 1$ *is not convex, why?*); linear constraints define convex sets, such as $Ax \leq b$; linear matrix inequalities define convex sets—special case the PSD constraint, $A \succeq 0$.

Projections onto convex sets. The definition of a projection onto a set is as the following optimization problem:

$$\Pi_{\mathcal{C}}(x) = \underset{z \in \mathcal{C}}{\text{argmin}} \, \ell(x, z).$$

Here, $\mathcal{C}$ defines the set on which we want to project, x is a given point, and $\ell(x, z)$ defines a notion of distance between x and a point in $\mathcal{C}$, which we want to minimize. Classical examples for $\ell(x, z)$ are norms such as $\ell(x, z) = \|x - z\|_2^2$, which will be the focus here. Thus, a verbal description of

$$\Pi_{\mathcal{C}}(x) = \underset{z \in \mathcal{C}}{\text{argmin}} \, \|x - z\|_2^2$$

is *"Given a point x and a set $\mathcal{C}$, find a point in $\mathcal{C}$ that is closer to x with respect to the Euclidean distance"*. When the set $\mathcal{C}$ is convex, then this defines the Euclidean projection onto the convex set $\mathcal{C}$.

An illustration of the above is shown below.

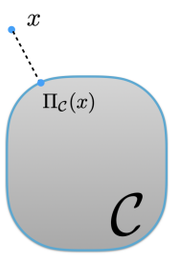

Fig. 23. Projection onto convex set $\mathcal{C}$.

Some useful properties for projections onto convex sets are:

- $\|x - \Pi_{\mathcal{C}}(x)\|_2^2 \leq \|x - y\|_2^2, \ \ \forall y \in \mathcal{C}, \ \forall x$, with the following illustration:

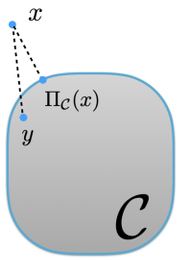

Fig. 24. The Euclidean distance from x to its projection onto $\mathcal{C}$ is the smallest among the points in $\mathcal{C}$.

This is actually the definition of the projection, as the minimum distance.

- $\langle \Pi_{\mathcal{C}}(x) - y, \ \Pi_{\mathcal{C}}(x) - x \rangle \leq 0, \ \ \forall y \in \mathcal{C}, \ \forall x$, with the following illustration:

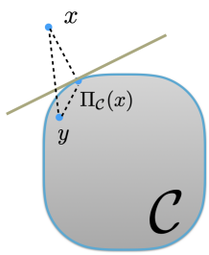

Fig. 25. The angle between the rays $\Pi_{\mathcal{C}}(x) - y$ and $\Pi_{\mathcal{C}}(x) - x$ are more than 90^o.

The interpretation is given in the caption above. This property does not hold for non-convex functions; a counterexample is given in the following figure.

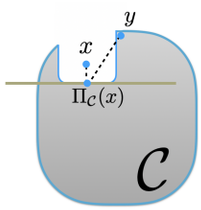

Fig. 26. The above property does not necessarily hold for non-convex sets.

- $\|\Pi_{\mathcal{C}}(x) - \Pi_{\mathcal{C}}(y)\|_2 \leq \|x - y\|_2, \ \ \forall x, y.$

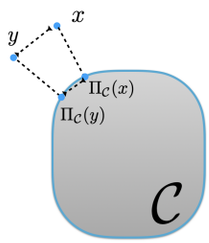

Fig. 27. The distance between any two points is greater than the distance of their projections onto a convex set.

Projected gradient descent. Given the notion of projections, we can define the projected version of gradient descent:

$$x_{t+1} = \Pi_{\mathcal{C}}\left(x_t - \eta_t \nabla f(x_t)\right), \quad t = 0, 1, \ldots,$$

which can be alternatively seen as a two-step procedure:

$$\widetilde{x}_{t+1} = x_t - \eta_t \nabla f(x_t), \quad t = 0, 1, \ldots,$$
$$x_{t+1} = \Pi_{\mathcal{C}}\left(\widetilde{x}_{t+1}\right), \quad t = 0, 1, \ldots,$$

But, do we lose anything by including the projection step? Can we preserve the same convergence guarantees?

Claim 5. *For a function $f : \mathbb{R}^p \to \mathbb{R}$ that is L-smooth and μ-strongly convex, projected gradient descent converges according to:*

$$\|x_{t+1} - x^\star\|_2^2 \leq \left(1 - \tfrac{2\eta\mu L}{\mu + L}\right) \|x_t - x^\star\|_2^2.$$

Proof: By definition, $x_{t+1} = \Pi_C(x_t - \eta \nabla f(x_t))$. So,

$$\begin{aligned}
\|x_{t+1} - x^\star\|_2^2 &= \|\Pi_{\mathcal{C}}\left(x_t - \eta \nabla f(x_t)\right) - x^\star\|_2^2 \\
&= \|\Pi_{\mathcal{C}}\left(x_t - \eta \nabla f(x_t)\right) - \Pi_{\mathcal{C}}(x^\star)\|_2^2 \\
&\leq \|x_t - \eta \nabla f(x_t) - x^\star\|_2^2 \\
&\leq \ldots \quad \text{(Similar analysis to GD)} \\
&\leq \left(1 - \tfrac{2\eta\mu L}{\mu + L}\right) \|x_t - x^\star\|_2^2.
\end{aligned}$$

□

Convergence of projected gradient descent for L-smooth functions. Similar analysis holds for the case of just L-smooth functions. However, we need a bit of care how to handle the analysis in f values. Remember that for just L-smooth functions, there might be multiple global solutions $x^\star$ that minimize the objective, and thus the notion of a distance $\|x_{t+1} - x^\star\|_2$ does not make sense in an recursion.

We know from the analysis of unconstrained optimization that, for L-smooth functions and for step size $\eta = \frac{1}{L}$ in gradient descent:

$$f(x_{t+1}) - f(x_t) \leq -\tfrac{1}{2L} \|\nabla f(x_t)\|_2^2.$$

This is based on the application of L-smoothness, where:

$$\begin{aligned}
f(x_{t+1}) &\leq f(x_t) + \langle \nabla f(x_t), x_{t+1} - x_t \rangle + \tfrac{L}{2} \|x_{t+1} - x_t\|_2^2 \\
&\overset{(i)}{=} f(x_t) + \langle \nabla f(x_t), x_t - \eta_t \nabla f(x_t) - x_t \rangle \\
&\qquad + \tfrac{L}{2} \|x_t - \eta_t \nabla f(x_t) - x_t\|_2^2 \\
&= \cdots
\end{aligned}$$

Though, in a constrained case such as,

$$\begin{aligned}
\min_{x \in \mathbb{R}^p} \quad & f(x) \\
\text{subject to} \quad & x \in \mathcal{C}.
\end{aligned}$$

we have:

$$x_{t+1} = \Pi_{\mathcal{C}}\left(x_t - \eta \nabla f(x_t)\right),$$

which complicates things, so that equation (i) does not hold.

To overcome this difficulty, we will need the notion of *gradient mapping*. Without getting into many details (*gradient mapping is not going to be used for the rest of the course*), we can prove the following result:

Lemma 5. *Let $\mathcal{C} \subset \mathbb{R}^p$ be a convex set, and let $x, y \in \mathcal{C}$. Let $f : \mathcal{C} \to \mathbb{R}$ be a convex function that we want to minimize and that satisfies L-smoothness. Define: $x^+ = \Pi_{\mathcal{C}}\left(x - \frac{1}{L}\nabla f(x)\right)$. Define also the function $g_{\mathcal{C}}(x) = L \cdot (x - x^+)$. Then, the following inequality holds true:*

$$f(x^+) - f(y) \leq \langle g_{\mathcal{C}}(x), x - y \rangle - \tfrac{1}{2L} \|g_{\mathcal{C}}(x)\|_2^2.$$

Proof: Since $\mathcal{C}$ is a convex set, by the projection properties we know that:

$$\begin{aligned}
&\left\langle x^+ - \left(x - \tfrac{1}{L}\nabla f(x)\right),\ x^+ - y \right\rangle \leq 0 \Rightarrow \\
&\left\langle x^+ - x,\ x^+ - y \right\rangle + \tfrac{1}{L}\left\langle \nabla f(x),\ x^+ - y \right\rangle \leq 0 \Rightarrow \\
&\qquad \tfrac{1}{L}\left\langle \nabla f(x),\ x^+ - y \right\rangle \leq \left\langle x - x^+,\ x^+ - y \right\rangle \Rightarrow \\
&\qquad\qquad \left\langle \nabla f(x),\ x^+ - y \right\rangle \leq \left\langle g_{\mathcal{C}}(x),\ x^+ - y \right\rangle
\end{aligned}$$

Then, we observe the following series of (in)equalities:

$$\begin{aligned}
f(x^+) - f(y) &= f(x^+) - f(x) + f(x) - f(y) \\
&\leq \left\langle \nabla f(x),\ x^+ - x \right\rangle + \tfrac{L}{2}\|x^+ - x\|_2^2 + \langle \nabla f(x),\ x - y \rangle \\
&= \left\langle \nabla f(x),\ x^+ - y \right\rangle + \tfrac{1}{2L}\|g_{\mathcal{C}}(x)\|_2^2 \\
&\leq \left\langle g_{\mathcal{C}}(x),\ x^+ - y \right\rangle + \tfrac{1}{2L}\|g_{\mathcal{C}}(x)\|_2^2 \\
&= \langle g_{\mathcal{C}}(x),\ x - y \rangle - \tfrac{1}{2L}\|g_{\mathcal{C}}(x)\|_2^2
\end{aligned}$$

where the first inequality is due to L-smoothness and by convexity.

□

Given the above and the recursion of projected gradient descent:

$$x_{t+1} = \Pi_{\mathcal{C}}\left(x_t - \eta \nabla f(x_t)\right),$$

by the lemma above we can compute:

$$f(x_{t+1}) - f(x_t) \leq -\tfrac{1}{2L}\|g_{\mathcal{C}}(x_t)\|_2^2,$$

and

$$f(x_{t+1}) - f(x^\star) \leq \|g_{\mathcal{C}}(x_t)\|_2 \cdot \|x_t - x^\star\|_2$$

Using similar analysis with the unconstrained case, we can prove:

$$\Delta_T \leq \frac{3L\|x_1 - x^\star\|_2^2 + (f(x_1) - f(x^\star))}{T},$$

where showing $\|x_{t+1} - x^\star\|_2 \leq \|x_t - x^\star\|_2$ stems from the use of the above lemma.

Thus, overall, projections do not change the convergence rate. However, it definitely changes the per iteration complexity: e.g., consider a projection procedure that is as difficult to complete as the original problem.

Opinion: Convex optimization is a technology. Convex optimization has become one of the most well-studied and well-understood areas of optimization; another such area is that of linear programming. To this point, there are several off-the-shelf solvers that are available online

- `CVXOPT` - `https://cvxopt.org`
- `CVXPy` - `http://www.cvxpy.org/`
- `CVX` - `http://cvxr.com/cvx/`
- `JuliaOpt` - `https://www.juliaopt.org/`
- TensorFlow - `https://www.tensorflow.org/`
- PyTorch - `https://pytorch.org/`

Why do we still care about convex optimization?

- Several practical problems are actually convex.
- Many practical problems can be approximated by convex ones
- If one does not understand convex optimization, why even try understanding non-convex optimization? :)