MILA Optimisation Crash Course **Winter 2024**

Lecture 1: February 14

Lecturer: Quentin Bertrand *Scribe(s): Behnoush Khavari, Danilo Vucetic*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this crash course only with the permission of the Instructor.*

These are the scribe notes for the first lecture of the optimisation crash course at MILA organised by Quentin Bertrand, Damien Scieur, Lucas Maes, and Danilo Vucetic. The purpose of this course is to provide proofs for standard optimisation techniques in order to help practitioners better understand why their algorithms learn. Here is the course web page.

1.1 Introduction

Say we're given a function $f : \mathbb{R} \to \mathbb{R}$ that we would like to minimise. Using gradient descent we can generate a sequence of inputs, $\{x^k\}_{k\in[0,N]}$, that progressively decrease the function value until the minimum, $x^\star$, is reached. See Figure 1.1 for an illustration. With our sequence of inputs we want to show some of the following.

- The existence of $x^\star$.
- Convergence rate of the sequence and whether it converges.
- Some conditions on the convergence of the sequence such as:
  - $\nabla f(x^k) \to 0$
  - $f(x^k) - f(x^\star) \to 0$
  - $x^k - x^\star \to 0$

This lecture starts by introducing the concepts of smoothness and convexity. We then see how to derive convergence guarantees based on the definitions and properties of smoothness and convexity when using gradient descent.

1.2 Preliminaries

Definition 1.1 *Derivatives are defined by the following limit.*

$$\frac{d}{dx} f(x) = \lim_{h\to 0} \frac{f(x+h) - f(x)}{h}$$

Definition 1.2 *Gradient descent is characterized by the following equation.*

$$x^{k+1} = x^k - \alpha \nabla_x f(x^k) \tag{1.1}$$

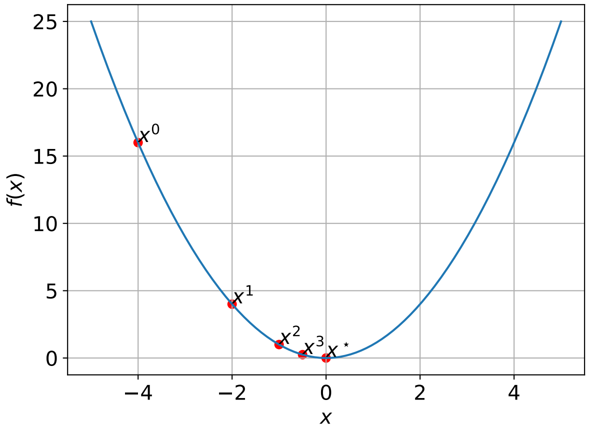

Figure 1.1: Graph of a parabola with some points of the sequence $\{x^k\}$ computed via gradient descent.

Definition 1.3 *The chain rule on a composition of functions* [1]. *Let* $f : \mathbb{R}^p \to \mathbb{R}$, $z : \mathbb{R} \to \mathbb{R}^p$, *and* $h(\lambda) := (f \circ z)(\lambda)$. *The derivative of the composition with respect to* $\lambda \in \mathbb{R}$ *is defined as follows.*

$$\frac{d}{d\lambda}(f \circ z)(\lambda) = \langle \nabla_z f(z(\lambda)), \frac{d}{d\lambda} z(\lambda) \rangle$$

Consequently, we can define the definite integral of the above by the following equation.

$$f(z(b)) - f(z(a)) = \int_a^b \langle \nabla_z f(z(\lambda)), \frac{d}{d\lambda} z(\lambda) \rangle \, d\lambda$$

Definition 1.4 *The Cauchy–Schwarz inequality is as follows. Let* $u, v \in \mathbb{R}^p$.

$$|\langle u, v \rangle| \leq \|u\| \|v\|$$

Note as well that the norm of a scalar is its absolute value, i.e., $\|a\| = |a|$.

1.3 Smooth Functions and their Minima

Definition 1.5 *A continuously differentiable function* $f : \mathbb{R}^p \to \mathbb{R}$ *with* $x, y \in \mathbb{R}^p$ *is L-smooth if*

$$\|\nabla f(y) - \nabla f(x)\| \leq L \|y - x\|$$

Our smoothness definition is pulled partly from [1] and [2]. Some examples of smooth and non-smooth functions are illustrated in Figure 1.2. A direct consequence of a function being L-smooth is that it can be bounded from above by a parabola.

Proposition 1.6 *An L-smooth function is upper-bounded by a parabola* $\forall x, y$

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|y - x\|^2$$

[1] See here for more information

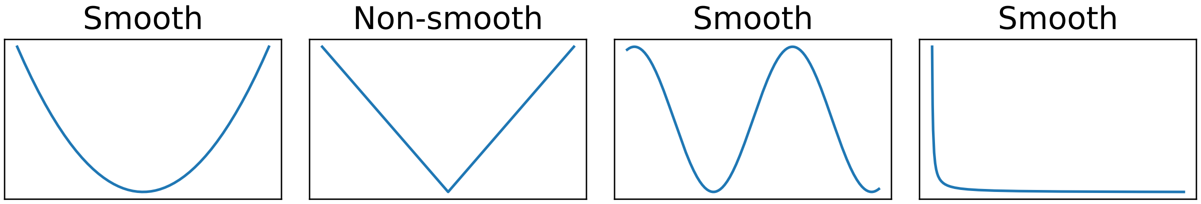

Figure 1.2: Smooth and non-smooth functions.

Proof: Starting from Definition 1.3, let $z(\lambda) := \lambda y + (1-\lambda)x$. Clearly, $\frac{d}{d\lambda} z(\lambda) = y - x$, $f(y) = f(z(1))$, and $f(x) = f(z(0))$. Thus, we have

$$\begin{aligned} f(y) - f(x) &= \int_0^1 \langle \nabla f(\lambda y + (1-\lambda)x), y - x \rangle \, d\lambda \\ &= \int_0^1 \langle \nabla f(\lambda y + (1-\lambda)x) - \nabla f(x) + \nabla f(x), y - x \rangle \, d\lambda \\ &= \int_0^1 \langle \nabla f(\lambda y + (1-\lambda)x) - \nabla f(x), y - x \rangle \, d\lambda + \int_0^1 \langle \nabla f(x), y - x \rangle \, d\lambda \end{aligned}$$

Moving the inner-product to the left hand side and applying a norm to both sides yields

$$\| f(y) - f(x) - \langle \nabla f(x), y - x \rangle \| = \left\| \int_0^1 \langle \nabla f(\lambda y + (1-\lambda)x) - \nabla f(x), y - x \rangle \, d\lambda \right\|$$

We also know that the norm of an integral is less than or equal to the integral of a norm.[2] Then, using Definition 1.4 and Definition 1.5, we solve the following

$$\begin{aligned} |f(y) - f(x) - \langle \nabla f(x), y - x \rangle| &\le \int_0^1 \| \langle \nabla f(\lambda y + (1-\lambda)x) - \nabla f(x), y - x \rangle \| \, d\lambda \\ &\le \int_0^1 \| \nabla f(\lambda y + (1-\lambda)x) - \nabla f(x) \| \, \| y - x \| \, d\lambda \\ &\le \| y - x \| \int_0^1 \| \nabla f(\lambda y + (1-\lambda)x) - \nabla f(x) \| \, d\lambda \\ &\le L \, \| y - x \| \int_0^1 \| \lambda y + (1-\lambda)x - x \| \, d\lambda \\ &= L \, \| y - x \|^2 \int_0^1 \lambda d\lambda \\ &= \frac{L}{2} \, \| y - x \|^2 \end{aligned}$$

Rearranging, we find the proposition.

$$f(y) \le f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \, \| y - x \|^2$$

[2] See here for more information.

■

We now state a theorem which provides the convergence rate of gradient descent on L-smooth functions that are lower bounded.

Theorem 1.7 *Let $f : \mathbb{R}^p \to \mathbb{R}$ be L-smooth and lower bounded, $\exists f^\star$ s.t. $f(x) \geq f^\star, \forall x$, then the convergence rate of gradient descent on f is characterized as follows*

$$\min_{k \in [0,N]} \left\| \nabla f(x^k) \right\|^2 \leq \frac{C}{N}$$

Proof: Let $y \leftarrow x^{k+1}$ and $x \leftarrow x^k$ on the gradient descent step $x^{k+1} = x^k - \alpha \nabla_x f(x^k)$. Let $\alpha = \frac{1}{L}$, and $x^\star$ be the minimum for f. Starting from Proposition 1.6 and noting that $x^{k+1} - x^k = -\frac{1}{L}\nabla_x f(x^k)$

$$\begin{aligned}
f(x^{k+1}) &\leq f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{L}{2} \left\| x^{k+1} - x^k \right\|^2 \\
&\leq f(x^k) + \langle \nabla f(x^k), -\frac{1}{L} \nabla f(x^k) \rangle + \frac{L}{2} \left\| \frac{1}{L} \nabla f(x^k) \right\|^2 \\
&= f(x^k) - \frac{1}{L} \left\| \nabla f(x^k) \right\|^2 + \frac{1}{2L} \left\| \nabla f(x^k) \right\|^2 \\
&= f(x^k) - \frac{1}{2L} \left\| \nabla f(x^k) \right\|^2
\end{aligned}$$

Note that the above result shows that the function value will decrease at each step of gradient descent! Rearranging so that the norm is on the left and summing over all steps of gradient descent, we have the following

$$\begin{aligned}
\left\| \nabla f(x^k) \right\|^2 &\leq 2L(f(x^k) - f(x^{k+1})) \\
\sum_{k=0}^{N-1} \left\| \nabla f(x^k) \right\|^2 &\leq 2L \sum_{k=0}^{N-1} (f(x^k) - f(x^{k+1})) \\
N \min_{k \in [0,N]} \left\| \nabla f(x^k) \right\|^2 \leq \sum_{k=0}^{N-1} \left\| \nabla f(x^k) \right\|^2 &\leq 2L(f(x^0) - f(x^N)) \leq 2L(f(x^0) - f(x^\star)) \\
\min_{k \in [0,N]} \left\| \nabla f(x^k) \right\|^2 &\leq \frac{2L}{N}(f(x^0) - f(x^\star))
\end{aligned}$$

The above result indicates that for a lower bounded L-smooth function, the convergence rate of gradient descent depends on the initial state, the Lipschitz constant, and the number of steps taken. ■

1.4 Convex Functions and their Minima

In this section we study the convergence rate of gradient descent for convex functions. We start by defining convex functions and then show that at any point, a convex function can be lower-bounded by a line. Finally, we prove a convergence rate for gradient descent on convex functions which, as expected, is stronger than what we saw for smooth functions. Some examples of convex and non-convex functions are given in Figure 1.3.

Definition 1.8 *A function $f : \mathbb{R}^p \to \mathbb{R}$ with $x, y \in \mathbb{R}^p$ is convex if and only if its domain is a convex set (see [1], to be defined) and if for $\alpha \in [0, 1]$ it satisfies*

$$f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha) f(y) \tag{1.2}$$

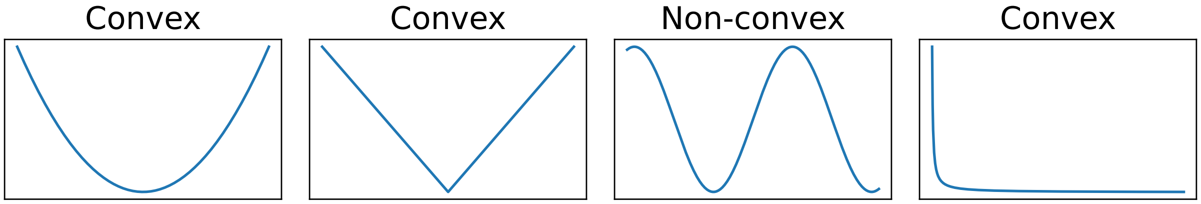

Figure 1.3: Convex and non-convex functions.

Proposition 1.9 *A (differentiable) convex function is lower-bounded by a line* $\forall x, y$

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

Proof: We use the definition of convexity to prove this lower bound. We take Equation (1.2) and subtract $f(y)$ from both sides to get

$$f(\alpha x + (1-\alpha)y) - f(y) \leq \alpha f(x) + (1-\alpha) f(y) - f(y)$$

$$\frac{f(\alpha x + (1-\alpha)y) - f(y)}{\alpha} \leq f(x) - f(y) \tag{1.3}$$

Consider the function $h : \alpha \to f(\alpha x + (1-\alpha)y)$. With this definition the above inequality can be written as

$$\frac{h(\alpha) - h(0)}{\alpha} \leq f(x) - f(y) \tag{1.4}$$

Since the convexity property implies the above inequality for arbitrary values of $\alpha \in [0,1]$, we can take the limit $\alpha \to 0$ in the above relation which gives the derivative of the function $h(\cdot)$ as per Definition 1.1.

$$\lim_{\alpha \to 0} \frac{h(\alpha) - h(0)}{\alpha} = h'(\alpha)|_{\alpha=0} \leq f(x) - f(y) \tag{1.5}$$

Following Definition 1.3 we can apply the chain rule for derivatives to $h'(\alpha)$ which gives the following equation

$$h'(\alpha) = \langle \nabla f(\alpha\, x + (1-\alpha)\, y), x - y \rangle \tag{1.6}$$

$$\to h'(0) = \langle \nabla f(y), x - y \rangle \tag{1.7}$$

Replacing this definition of $h'(0) := h'(\alpha)|_{\alpha=0}$ in Equation (1.5) we get

$$\langle \nabla f(y)\, ,\, x - y \rangle \leq f(x) - f(y) \tag{1.8}$$

Rearranging the terms completes the proof (up to swapping the names of variables x and y)

$$f(x) \geq f(y) + \langle \nabla f(y)\, ,\, x - y \rangle \tag{1.9}$$

■

Theorem 1.10 *Let* $f : \mathbb{R}^p \to \mathbb{R}$ *be smooth and convex and lower bounded, where* $\exists f^\star$ *s.t.* $f(x) \geq f^\star, \forall x$[3]*, then the convergence rate of gradient descent on* f *is characterized as follows*

$$f(x^k) - f(x^*) \leq \frac{L}{2} \frac{\|x_0 - x^*\|}{k} \tag{1.10}$$

[3] Note, this does not guarantee that the minimum is unique.

Proof: As before, let $y \leftarrow x^{k+1}$ and $x \leftarrow x^k$ on the gradient descent step $x^{k+1} = x^k - \alpha \nabla_x f(x^k)$. Let $\alpha = \frac{1}{L}$, and $x^\star$ be the minimum for f. Starting from Proposition 1.6 and noting that $x^{k+1} - x^k = -\frac{1}{L}\nabla_x f(x^k)$

$$\begin{aligned} f(x^{k+1}) &\le f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{L}{2}\left\|x^{k+1} - x^k\right\|^2 \\ &= f(x^k) + \langle \nabla f(x^k), -\frac{1}{L}\nabla f(x^k) \rangle + \frac{L}{2}\left\|\frac{1}{L}\nabla f(x^k)\right\|^2 \\ &= f(x^k) - \frac{1}{L}\left\|\nabla f(x^k)\right\|^2 + \frac{1}{2L}\left\|\nabla f(x^k)\right\|^2 \qquad (1.11) \\ &= f(x^k) - \frac{1}{2L}\left\|\nabla f(x^k)\right\|^2 \\ \rightarrow f(x^{k+1}) &\le f(x^k) - \frac{1}{2L}\left\|\nabla f(x^k)\right\|^2 \end{aligned}$$

Up to this point we only used the smoothness property of the function f. Now, we consider the convexity condition at the minimum point x^* which gives

$$f(x^*) \ge f(x^k) + \langle \nabla f(x^k), x^* - x^k \rangle \qquad (1.12)$$

$$f(x^*) - \langle \nabla f(x^k), x^* - x^k \rangle \ge f(x^k) \qquad (1.13)$$

We can read the above inequality as an upper bound on $f(x^k)$

$$f(x^k) \le f(x^*) - \langle \nabla f(x^k), x^* - x^k \rangle \qquad (1.14)$$

Using this result in (1.11) we get

$$\begin{aligned} f(x^{k+1}) &\le f(x^*) - \langle \nabla f(x^k), x^* - x^k \rangle - \frac{1}{2L}\left\|\nabla f(x^k)\right\|^2 \\ \rightarrow f(x^{k+1}) - f(x^*) &\le -\frac{1}{2L}\big(2L\langle \nabla f(x^k), x^* - x^k \rangle + \left\|\nabla f(x^k)\right\|^2\big) \qquad (1.15) \end{aligned}$$

We can use the identity $\|a+b\|^2 = \|a\|^2 + 2\langle a, b\rangle + \|b\|^2$ or equivalently $\|a\|^2 + 2\langle a, b\rangle = \|a+b\|^2 - \|b\|^2$ to rewrite the expression inside the parenthesis on the right-hand side as below

$$2L\langle \nabla f(x^k), x^* - x^k \rangle + \left\|\nabla f(x^k)\right\|^2 = \big(\left\|\nabla f(x^k) + L(x^* - x^k)\right\|\big)^2 - L^2\left\|x^* - x^k\right\|^2 \qquad (1.16)$$

More explicitly, we have considered $a = \nabla f(x^k)$ and $b = L(x^* - x^k)$; then the left-hand side of the above equation is $2\langle a, b\rangle + \|a\|^2$ and the right-hand side is $\|a+b\|^2 - \|b\|^2$.

From the gradient descent equation, i.e., Equation (1.1), with the learning rate value set as $\alpha = \frac{1}{L}$, we have

$$\nabla_x f(x^k) = L(x^k - x^{k+1}) \qquad (1.17)$$

applying this to the right-hand side of Equation (1.16) we get

$$\begin{aligned} &2L\langle \nabla f(x^k), x^* - x^k \rangle + \left\|\nabla f(x^k)\right\|^2 = \big(\left\|L(x^k - x^{k+1}) + L(x^* - x^k)\right\|\big)^2 - L^2\left\|x^* - x^k\right\|^2 \\ &= L^2\left\|(x^* - x^{k+1})\right\|^2 - L^2\left\|x^* - x^k\right\|^2 \qquad (1.18) \end{aligned}$$

From this, Equation (1.15) simplifies to

$$\rightarrow f(x^{k+1}) - f(x^*) \le -\frac{L}{2}\big(\left\|(x^* - x^{k+1})\right\|^2 - \left\|x^* - x^k\right\|^2\big) \qquad (1.19)$$

The last step is to sum both sides over the iteration index k

$$\sum_k \left[f(x^{k+1}) - f(x^*)\right] \leq \sum_k -\frac{L}{2}\big(\left\|(x^* - x^{k+1})\right\|^2 - \left\|x^* - x^k\right\|^2\big) = -\frac{L}{2}\big(\left\|x^* - x^N\right\|^2 - \left\|(x^* - x^0)\right\|^2\big)$$
$$= \frac{L}{2}\big(\left\|(x^* - x^0)\right\|^2 - \left\|x^* - x^N\right\|^2\big) \leq \frac{L}{2}\left\|(x^* - x^0)\right\|^2 \tag{1.20}$$

where we have used the telescoping sum formula to write the equality relation in the first line of Equation (1.20).

Finally, we notice that at each gradient step on the convex function the value of f decreases, i.e., $f(x^{k+1}) < f(x^k)$. Therefore the left-hand side of the above inequality for N gradient steps is lower-bounded as below

$$N\big(f(x^N) - f(x^*)\big) \leq \sum_{k=0}^{n-1} f(x^{k+1}) - f(x^*) \leq \frac{L}{2}\left\|(x^* - x^0)\right\|^2 \tag{1.21}$$

Hence the result

$$f(x^N) - f(x^*) \leq \frac{1}{N}\frac{L}{2}\left\|(x^* - x^0)\right\|^2 \tag{1.22}$$

■

References

[1] I. Mitliagkas *et al.*, "Gradients for smooth and for strongly convex functions." Course notes for IFT 6085 at UdeM.

[2] S. Bubeck *et al.*, "Convex optimization: Algorithms and complexity," *Foundations and Trends® in Machine Learning*, vol. 8, no. 3-4, pp. 1–357, 2015.