MILA Optimisation Crash Course **Winter 2024**

Lecture 3: February 28

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

These are the scribe notes for the third 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.

3.1 Recapitulation

Last week we saw subgradient descent and dual averaging for the optimisation of convex but non-smooth functions. With subgradient descent we saw how we could generalise the concept of a gradient to apply to non-smooth functions. Using this definition, we redefined gradient descent using subgradients. Since a non-smooth function doesn't necessarily have smaller gradient norms closer to the optimum, we must set an explicit learning rate schedule to ensure convergence. As we will see in this lecture, the choice of learning rate strongly affects the rate of convergence. With subgradient descent, we found the convergence guarantee of Inequality 3.1, where x^{BEST} is the best iterate found over k steps. Dual averaging is defined by averaging over the convexity lower bounds of the successive iterates x_i, where $i \in [0, k]$. By solving for the optimal iterates, we recover an averaging scheme where the subgradient descent iterates are averaged to produce the best iterate. Note that the 'dual' space here consists of the affine transformation $\nabla f(x)^\intercal(x_1 + x_2 + \cdots)$ that arises from averaging the convexity lower bounds. The 'primal space is simply the iterates themselves, e.g., x^{k+1}. Dual averaging recovers the convergence guarantee of Inequality 3.2. However, by comparing with the rates achieved for gradient descent on smooth and convex smooth functions, we notice that the rates of Inequalities 3.1 and 3.2 are worse (i.e., non-smooth functions are slower to optimise).

$$f(x^{\text{BEST}}) - f^\star \leq \mathcal{O}\left(\frac{1}{\sqrt{k}}\right) \tag{3.1}$$

$$f\left(\frac{\sum_{i=0}^{k-1} x_i}{k}\right) - f^\star \leq \mathcal{O}\left(\frac{1}{\sqrt{k}}\right) \tag{3.2}$$

3.2 Introduction

This lecture will finish off the content of the previous two lectures. First, we will discuss gradient descent with smooth and strongly convex functions. Then, we will discuss settings of the learning rate for dual averaging and how this affects the convergence rate that we derive.

We define our problem setting similarly to the last lecture. We have a function $f : \mathbb{R} \to \mathbb{R}$ that we would like to minimise. We would like to generate a sequence of inputs, $\{x^k\}_{k\in[0,N]}$, that decrease the function value until the minimum, $x^\star$, is reached. Of course, we want to show some guarantees on the rate of convergence.

3.3 Preliminaries

Lemma 3.1 *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., $\|a\| = |a|$.

Definition 3.2 *A continuously differentiable function $f : \mathbb{R}^p \to \mathbb{R}$ with $x, y \in \mathbb{R}^p$ is L-smooth if*

$$\|\nabla f(y) - \nabla f(x)\| \leq L\|y - x\|. \tag{3.3}$$

From [1] and [2].

Lemma 3.3 *A smooth function f is upper-bounded by a parabola, i.e.,*

$$f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|^2. \tag{3.4}$$

Definition 3.4 *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 all $\alpha \in [0, 1]$ it satisfies*

$$f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha)f(y). \tag{3.5}$$

Lemma 3.5 *A convex function f is lower-bounded by a line, i.e.,*

$$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle. \tag{3.6}$$

Definition 3.6 *Derivatives are defined by the following limit.*

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

Definition 3.7 *Gradient descent is characterized by the following equation.*

$$x^{k+1} = x^k - \alpha\nabla_x f(x^k) \tag{3.7}$$

Definition 3.8 *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.

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

[1] See here for more information

3.4 Strongly Convex Functions and their Minima

Definition 3.9 *A function $f : \mathbb{R}^p \to \mathbb{R}$ with $x, y \in \mathbb{R}^p$ is strongly 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*

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y) - \frac{\mu}{2}\alpha(1-\alpha)\left\|x-y\right\|^2 \tag{3.8}$$

for some positive μ.

Proposition 3.10 *A (differentiable) strongly convex function is lower-bounded by a parabola $\forall x, y$*

$$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{\mu}{2}\left\|y-x\right\|^2 \tag{3.9}$$

Proof: We use the definition of strong convexity to prove this lower bound. We take Equation (3.8) and subtract $f(y)$ from both sides to get

$$\begin{aligned} f(\alpha x + (1-\alpha)y) - f(y) &\leq \alpha f(x) + (1-\alpha)f(y) - \frac{\mu}{2}\alpha(1-\alpha)\left\|x-y\right\|^2 - f(y) \\ \frac{f(\alpha x + (1-\alpha)y) - f(y)}{\alpha} &\leq f(x) - f(y) - \frac{\mu}{2}(1-\alpha)\left\|x-y\right\|^2 \end{aligned} \tag{3.10}$$

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

$$\frac{h(\alpha) - h(0)}{\alpha} \leq f(x) - f(y) - \frac{\mu}{2}(1-\alpha)\left\|x-y\right\|^2 \tag{3.11}$$

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 3.6.

$$\lim_{\alpha\to 0}\frac{h(\alpha) - h(0)}{\alpha} = h'(\alpha)|_{\alpha=0} \leq \lim_{\alpha\to 0} f(x) - f(y) - \frac{\mu}{2}(1-\alpha)\left\|x-y\right\|^2 \tag{3.12}$$

Notice that the factor $1-\alpha$ tends to 1 and hence, we can reduce the last term on the right-hand side. Following Definition 3.8 we can apply the chain rule for derivatives to $h'(\alpha)$ which gives the following equation

$$h'(\alpha) = \langle \nabla f(\alpha\, x + (1-\alpha)\, y), x - y\rangle \tag{3.13}$$

$$\to h'(0) = \langle \nabla f(y), x - y\rangle \tag{3.14}$$

Replacing this definition of $h'(0) := h'(\alpha)|_{\alpha=0}$ in Equation (3.12) we get

$$\langle \nabla f(y)\,,\, x - y\rangle \leq f(x) - f(y) - \frac{\mu}{2}\left\|x-y\right\|^2 \tag{3.15}$$

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

$$f(x) \geq f(y) + \langle \nabla f(y)\,,\, x - y\rangle + \frac{\mu}{2}\left\|x-y\right\|^2 \tag{3.16}$$

■

We now state a useful proposition for proving the convergence rate with strongly convex functions.

Proposition 3.11 *A strongly convex and differentiable function f satisfies the Polyak-Łojasiewicz inequality*

$$f(x) - f(x^*) \leq \frac{1}{2\mu}\left\|\nabla f(x)\right\|^2 \tag{3.17}$$

Proof: We start from Inequality 3.9, minimising both sides of the inequality.

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2} \|y - x\|^2$$
$$\min_y f(y) = f^\star \geq min_y \left(f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2} \|y - x\|^2 \right)$$

To minimise the right-hand side, we take the gradient with respect to y and set the equation to zero.

$$\begin{aligned} 0 &= \nabla_y \left(f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2} \|y - x\|^2 \right) \\ &= \nabla f(x) + \mu(y - x) \\ y &= x - \frac{\nabla f(x)}{\mu} \end{aligned}$$

Now, we can plug the optimal y into the previous result to recover the PL inequality.

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

Then, rearranging we recover $f(x) - f^\star \leq \frac{1}{2\mu} \|\nabla f(x)\|^2$. ■

Next, we show that for strongly convex functions we have a stronger (linear) convergence rate which is stated in the following theorem.

Theorem 3.12 *Let $f : \mathbb{R}^p \to \mathbb{R}$ be smooth and strongly convex and lower bounded, where $\exists f^\star$ s.t. $f(x) \geq f^\star, \forall x$[2], then the convergence rate of gradient descent on f is characterized as follows*

$$f(x^k) - f(x^*) \leq (1 - \frac{\mu}{L})^k (f(x^0) - f(x^*)) \tag{3.18}$$

Proof: We start from the strongly convex lower bound of Inequality 3.9 and minimise both sides.

$$\min_x f(x) \geq \min_x \left(f(y) + \langle \nabla f(y)\,,\, x - y \rangle + \frac{\mu}{2} \|x - y\|^2 \right)$$
$$f(x^\star) \geq \min_x \left(f(y) + \langle \nabla f(y)\,,\, x - y \rangle + \frac{\mu}{2} \|x - y\|^2 \right)$$

To find the minimiser of the right hand side, we take the gradient with respect to x and find $x - y$:

$$\begin{aligned} 0 &= \nabla_x \left(f(y) + \langle \nabla f(y)\,,\, x - y \rangle + \frac{\mu}{2} \|x - y\|^2 \right) \\ &= \nabla f(y) + \mu(x - y) \\ x - y &= -\frac{\nabla f(y)}{\mu} \end{aligned} \tag{3.19}$$

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

Replacing this value of $x - y$ to Inequality 3.9, we obtain the minimum value of the right-hand side of the inequality.

$$\begin{aligned}
\min_x f(x) &\geq \min_x \left(f(y) + \langle \nabla f(y)\,,\, x - y \rangle + \frac{\mu}{2} \|x - y\|^2 \right) \\
f(x^\star) &\geq f(y) + \langle \nabla f(y)\,,\, -\frac{\nabla f(y)}{\mu} \rangle + \frac{\mu}{2} \left\| \frac{\nabla f(y)}{\mu} \right\|^2 \\
&= f(y) - \frac{1}{2\mu} \|\nabla f(y)\|^2 \\
f(y) - f(x^\star) &\leq \frac{1}{2\mu} \|\nabla f(y)\|^2
\end{aligned}$$

Now, as we saw in the proof of the convergence rate for smooth functions (Theorem 1.7 in the first lecture), at each iteration step the value of the to-be-minimized function decreases as below.

$$f(x^{k+1}) \leq f(x^k) - \frac{1}{2L} \left\| \nabla f(x^k) \right\|^2 \tag{3.20}$$

By subtracting $f(x^*)$ from both sides of this inequality we get

$$f(x^{k+1}) - f(x^*) \leq f(x^k) - f(x^*) + \frac{1}{2L}(- \left\| \nabla f(x^k) \right\|^2) \tag{3.21}$$

Applying the negated Polyak-Łojasiewicz inequality, $- \left\| \nabla f(x^k) \right\|^2 \leq -2\mu(f(x^k) - f(x^*))$, on the last term on the right-hand side and expanding the resulting recurrence we get

$$\begin{aligned}
f(x^{k+1}) - f(x^*) &\leq f(x^k) - f(x^*) + \frac{1}{2L}(- \left\| \nabla f(x^k) \right\|^2) \\
&\leq f(x^k) - f(x^*) - \frac{\mu}{L}(f(x^k) - f(x^*)) \\
&= \left(1 - \frac{\mu}{L}\right)(f(x^k) - f(x^*)) \\
&\leq \left(1 - \frac{\mu}{L}\right)^2 (f(x^{k-1}) - f(x^*)) \\
&\cdots \\
&\leq \left(1 - \frac{\mu}{L}\right)^k (f(x^0) - f(x^*))
\end{aligned} \tag{3.22}$$

■

3.5 Learning Rates for Dual Averaging

How do we choose the learning rate (also called step size) for subgradient descent and dual averaging? In last week's lecture we assumed that the step size was $\alpha_i = \frac{1}{\sqrt{k}}$, where we were doing k optimisation steps. However, we also said in that lecture that we wanted to have a consistently decreasing step size so that we would encourage convergence. The reason we set this constant step size was to make the analysis simpler. As a reminder, for subgradient descent we had

$$f(x^{\text{BEST}}) - f(x^\star) \leq \frac{1}{2} \left(\frac{1}{\sum_{i=0}^k \alpha_i} D + \frac{\sum_{i=0}^k \alpha_i^2}{\sum_{i=0}^k \alpha_i} G^2 \right)$$

Notice that with $\alpha_i = \frac{1}{\sqrt{k}}$, $\sum_{i=0}^{k-1} \alpha_i = \sqrt{k}$ and $\sum_{i=0}^{k-1} \alpha_i^2 = 1$. Thus, the above reduces to

$$f(x^{\text{BEST}}) - f(x^\star) \leq \frac{1}{2\sqrt{k}} \left(D + G^2\right)$$

If we were to instead set the step size to $\alpha_i = \frac{1}{\sqrt{i+1}}$, we end up with the harmonic series $\sum_{i=0}^{k-1} \alpha_i^2 = \ln k + \gamma$ and the following

$$\begin{aligned}
\sum_{i=0}^{k-1} \alpha_i &= \sum_{i=0}^{k-1} \frac{1}{\sqrt{i+1}} \\
\sqrt{k} &\leq \sum_{i=0}^{k-1} \frac{1}{\sqrt{i+1}} \\
&\leq \int_0^k \frac{1}{\sqrt{x}} dx \\
&= 2\sqrt{x}\Big|_0^k = 2\sqrt{k}
\end{aligned}$$

If we accept that $\sum_{i=0}^{k-1} \alpha_i^2 \approx \ln k$ and $\sum_{i=0}^{k-1} \alpha_i \approx \sqrt{k}$, we get

$$f(x^{\text{BEST}}) - f(x^\star) \lesssim \mathcal{O}\left(\frac{\ln k}{\sqrt{k}}\right)$$

It would appear that our rate of convergence for the decreasing learning rate is significantly slower than for our constant learning rate for subgradient descent! However, this interpretation is incorrect. By instead following an analysis similar to [3], we can recover the same rate of $\frac{1}{\sqrt{k}}$. Hence, both settings of the learning rate work, but in reality we don't know how many steps we'll do in advance, so we should use the decreasing rate.

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.

[3] J. Jiao, A. AlAnqary, P. Canoza, C. Ikeokwu, and M. Zhang, "Ee c227c convex optimization and approximation: Subgradient descent method.".