MILA Optimisation Crash Course **Winter 2024**

Lecture 4: March 6th

Lecturer: Damien Scieur *Scribe(s): Marco Jiralerspong, 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 fourth 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.

4.1 Recapitulation

As we saw in the two preceding lectures there is a major difference between gradient descent (GD) and sub-gradient descent/dual averaging (DA) algorithms; GD leverages the smoothness assumption on the function to be minimized and uses the upper-bounding parabola, while DA works based on minimizing the line (plus a regularization term) that lower-bounds the function. One uses overly optimistic iterates while the other is too cautions.

For the GD algorithm we proved a convergence rate over the value of the function w.r.t. the minimum value, of order $O(\frac{1}{K})$ after K gradient steps for the convex case and later we saw that this convergence rate automatically increases to $O(1-\frac{\mu}{L})^K$ for strongly convex functions (by automatically, we mean with the same algorithm we get a better convergence rate for a better-behaved function).

For the DA algorithm we proved that even if the function f is non-smooth, by leveraging the convexity of the function we only need to adjust the learning rate appropriately to get a convergence rate of $O(\frac{1}{\sqrt{K}})$.

4.2 Introduction

In this lecture, we are again interested in minimising some function $f(x)$, analysing the iterates of the optimisation algorithm to derive convergence bounds. This lecture will focus on Nesterov's acceleration algorithm, which produces a substantially better convergence rate than gradient descent. The algorithm will be introduced via the combination of dual averaging (this time on smooth functions) and gradient descent, which in itself is a novel, but surprisingly intuitive, derivation. Finally, while the convergence rate of Nesterov's acceleration is better than raw gradient descent, the function evaluated at the iterates is not guaranteed to be monotonically decreasing.

4.3 Preliminaries

Lemma 4.1 *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 4.2 *A continuously differentiable function $f : \mathbb{R}^p \to \mathbb{R}$ with $x, y \in \mathbb{R}^p$ is L-smooth if*

$$\|\nabla f(y) - \nabla f(x)\| \leq L\|y - x\|. \tag{4.1}$$

From [1] and [2].

Lemma 4.3 *A smooth function f is upper-bounded by a parabola, i.e.,*

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|y - x\|^2 . \tag{4.2}$$

Definition 4.4 *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*

$$f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha)f(y). \tag{4.3}$$

Lemma 4.5 *A convex function f is lower-bounded by a line, i.e.,*

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle. \tag{4.4}$$

Definition 4.6 *Derivatives are defined by the following limit.*

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

Definition 4.7 *Gradient descent is characterized by the following equation.*

$$x^{k+1} = x^k - \alpha \nabla_x f(x^k) \tag{4.5}$$

Definition 4.8 *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

Aside

For the class of smooth and convex functions, for any sequence of iterates $x_k : x_0 + \sum_{i=0}^{k} \alpha_i \nabla f(x_i)$, it can be shown that:

$$\exists f(x) : f(x_k) - f^* \geq O\left(\frac{1}{k^2}\right)(f(x_0) - f^*),$$

i.e. any gradient based method will at least have a convergence rate of $O\left(\frac{1}{k^2}\right)$. Meanwhile, for gradient descent, we get a rate of $f(x_k) - f^* \leq O(\frac{1}{k})(f(x_0) - f^*)$.

Given the gap between the two rates ($O\left(\frac{1}{k^2}\right) \leq ? \leq O\left(\frac{1}{k}\right)$), would it be possible construct a method that improves upon the GD rate using both smoothness and convexity? This was an open question for twenty years until Nesterov's work!

4.4 Nesterov's Acceleration

The idea of Nesterov acceleration is to leverage both the smoothness and convexity properties of the function by using both the upper-bounding parabola and the lower-bounding line to find a more effective update at each step. That is, at each update step k we have two choices for calculating the next step from the current value x_k:

$$y_{k+1} = x_k - \frac{1}{L}\nabla f(x_k) \qquad \text{(GD step which only exploits the upper bound).} \tag{4.6}$$

$$z_{k+1} = x_0 - \frac{1}{\mu_k}\left(\frac{\sum \alpha_i \nabla f(x_i)}{\sum \alpha_i}\right) \qquad \text{(DA step which only exploits the lower bound).} \tag{4.7}$$

Now, Nesterov suggests to combine the above updates as follows:

$$x_{k+1} = \beta_k y_{k+1} + (1 - \beta_k) z_{k+1} \tag{4.8}$$

4.4.1 Notation

Before turning to the proof/justification of the third formula, we would like to clarify the above notation that we will continue to use in the rest of this draft.

- x_i: iterates of Nesterov's acceleration
- y_i: iterates of gradient descent
- z_i: iterates of dual averaging
- (α_i, μ_i): dual averaging parameters
- $A_k = \sum_{i=0}^{k} \alpha_i$
- β_i: coefficients for linear interpolation between GD and DA

4.5 Derivation of Nesterov's formula

The question now is what values should we assign to the parameters of the iterates mentioned above (i.e. α_k, μ_k, β_k) to get a better convergence rate than GD/DA? For this, we will use **2 inequalities** and **1 relation**. Combined with setting the appropriate parameter values, this will yield a convergence rate of $O\left(\frac{1}{k^2}\right)$!

4.5.1 Derivation ingredients

The first inequality is valid for the GD iterates:

$$f(y_{k+1}) \leq f(x_k) - \frac{1}{2L} \|\nabla f(x_k)\|^2 \tag{4.9}$$

This is what we already found while proving the convergence rate for smooth functions (Theorem 1.7 in the first lecture), and tells us that GD steps with correct learning rate (defined based on the smoothness coefficient L) always reduce the value of a smooth function.

The second inequality comes from the DA updates and is based on a lower bound valid for convex functions:

$$\frac{\alpha_k}{A_k} \left[(f(x^*) \geq f(x_k) + \nabla f(x_k)(x^* - x_k) \right] \tag{4.10}$$

Finally, we use a relation that comes from the dual averaging update:

$$z_{k+1} = z_0 - \frac{1}{\mu_k} \frac{\sum_{i=0}^{k} \alpha_i \nabla f(x_i)}{A_k}. \tag{4.11}$$

Notice that the above update rule comes from what we saw in lecture 2, Proposition 2.8. That is, by minimizing the key expression of dual averaging (weighted average of the convexity-based lower bounds at all K iteration steps plus the regularization term with the coefficient μ) in Proposition 2.8 we arrive to the above update equation for DA.

Let's now manipulate (4.11). We first rearrange the equality, multiply both sides by the factor $\mu_k A_k$ and also break the sum $\sum_{i=0}^{k}$ to $\sum_{i=0}^{k-1}$ and the single k-th term which is $\alpha_k \nabla f(x_k)$ as below:

$$\mu_k A_k (z_{k+1} - z_0) = -\sum_{i=0}^{k-1} \alpha_i \nabla f(x_i) - \alpha_k \nabla f(x_k) \tag{4.12}$$

Notice that the summation term $\sum_{i=0}^{k-1} \alpha_i \nabla f(x_i)$ can be again written in terms of the DA updates from the DA update equation (4.11), i.e.

$$\sum_{i=0}^{k-1} \alpha_i \nabla f(x_i) = \mu_{k-1} A_{k-1} (z_k - z_0).$$

Then, replacing this into (4.12) we get:

$$\mu_k A_k (z_{k+1} - z_0) - \mu_{k-1} A_{k-1} (z_k - z_0) = -\alpha_k \nabla f(x_k).$$

Now, if we make the simplifying assumption that for all k we set $\mu_k A_k = C$, the above equation can be rewritten as:

$$C(z_{k+1} - x^*) = C(z_k - x^*) - \alpha_k \nabla f(x_k).$$

In the above, $-Cz_0$ terms on the left/right side of the equation cancelled out and we replaced them with $-Cx^*$ as this will be useful later. By taking the squared norm of both sides we get:

$$C^2 \left\|z_{k+1} - x^*\right\|^2 = C^2 \left\|z_k - x^*\right\|^2 + \alpha_k^2 \left\|\nabla f(x_k)\right\|^2 - 2\alpha_k C \left\langle \nabla f(x_k), z_k - x^* \right\rangle$$

With some rearranging, the final relation we will use is:

$$\boxed{\alpha_k \langle \nabla f(x_k), z_k - x^* \rangle = \frac{C}{2} \left\|z_k - x^*\right\|^2 + \frac{\alpha_k^2}{2C} \left\|\nabla f(x_k)\right\|^2 - \frac{C}{2} \left\|z_{k+1} - x^*\right\|^2 . \qquad (4.13)}$$

4.5.2 Derivation

1. Combine inequalities

Until this point, we have two inequalities and one relation that we needed for our proof. Now, we start by combining the two first inequalities. This corresponds to the information we gain at a given step. It straightforwardly gives us the following inequality:

$$A_k \left[f(y_{k+1}) - f(x_k) + \frac{1}{2L} \left\|\nabla f(y_k)\right\|^2 \right] + \alpha_k \left[f(x_k) - f(x^*) + \nabla f(x_k)(x^* - x_k) \right] \leq 0.$$

2. Reorganization

We then reorganize the inequality, adding and subtracting a $A_k[f(x_k) - f(x^*)]$ term:

$$A_k \big[f(y_{k+1}) - f(x^*) \big] - (A_k - \alpha_k)[f(x_k) - f(x^*)] + \frac{A_k}{2L} \left\|\nabla f(y_k)\right\|^2 + \alpha_k \nabla f(x_k)(x^* - x_k) \leq 0$$

3. Conversion

We first replace $A_k - \alpha_k$ by A_{k-1} (by definition of A_k). We can then convert the $f(x_k) - f(x^*)$ to y_k using convexity. Specifically, $f(y_k) \geq f(x_k) + \nabla f(x_k)(y_k - x_k)$. If we multiply the inequality by $(-A_{k-1})$, we have:

$$-A_{k-1} f(x_k) \leq -A_{k-1} f(y_k) + A_{k-1} \nabla f(x_k)(y_k - x_k).$$

We can then replace the $-A_{k-1} f(x_k)$ by the right side of the inequality and rearrange terms to get:

$$\begin{aligned} \big[f(y_{k+1}) - f(x^*) \big] A_k - A_{k-1}[f(y_k) - f(x^*)] + \frac{A_k}{2L} \left\|\nabla f(y_k)\right\|^2 \\ + \nabla f(x_k)[\alpha_k (x^* - x_k) + A_{k-1}(y_k - x_k)] \leq 0 \end{aligned}$$

4. Nesterov's magic

We now use a bit of magic, in the sense that we already know the correct iterates for the derivation:

$$A_k x_k := A_{k-1} y_k + \alpha_k z_k \tag{4.14}$$

We use them to simplify the last term:

$$\begin{aligned}
&\nabla f(x_k)[\alpha_k(x^* - x_k) + A_{k-1}(y_k - x^k)] \\
&= \nabla f(x_k)[\alpha_k x^* - (\alpha_k + A_{k-1})x_k + A_{k-1}y_k] \\
&= \nabla f(x_k)[\alpha_k x^* - A_k x_k + A_{k-1}y_k] && \text{(Using the definition of } A_k\text{)} \\
&= \nabla f(x_k)[\alpha_k(x^* - z_k)] && \text{(Using the Nesterov iterates)}
\end{aligned}$$

5. Bringing it all together

Finally, we use the DA relation 4.5.1 to replace the last term and get the following expression:

$$\begin{aligned}
A_k\big[f(y_{k+1}) - f(x^*)\big] - A_{k-1}[f(y_k) - f(x^*)] + \frac{C\left\|z_{k+1} - x^*\right\|^2}{2} \\
-\frac{C\left\|z_k - x^*\right\|^2}{2} + \left[\frac{A_k}{2L} - \frac{\alpha_k^2}{2C}\right]\left\|\nabla f(x_k)\right\|^2 \le 0
\end{aligned} \tag{4.15}$$

This is almost a recurrence except for the last term! To fix this, we want the following to be 0:

$$\frac{A_k}{2L} - \frac{\alpha_k^2}{2C} \implies C = L \text{ and } A_k = \alpha_k^2 \tag{4.16}$$

Solving the above yields:

$$A_{k-1} + \alpha_k = \alpha_k^2 \implies \alpha_k \sim k, A_k \sim k^2.$$

Finally, our recurrence is:

$$A_k\big[f(y_{k+1}) - f(x^*)\big] + \frac{C\left\|z_{k+1} - x^*\right\|^2}{2} \le A_{k-1}[f(y_k) - f(x^*)] + \frac{C\left\|z_k - x^*\right\|^2}{2}. \tag{4.17}$$

Thus, from the above, we have that the full set of Nesterov iterates is:

Nesterov Iterates

$$y_{k+1} := x_k - \frac{1}{L}\nabla f(x_k) \qquad \text{(GD step).} \tag{4.18}$$

$$z_{k+1} := z_k - \frac{\alpha_k}{L}\nabla f(x_k) \qquad \text{(DA step after using } \mu_k A_k = C = L\text{).} \tag{4.19}$$

$$x_{k+1} := \frac{A_k y_{k+1} + \alpha_{k+1} z_{k+1}}{A_{k+1}} \qquad \text{(Mixing used in 4.14).} \tag{4.20}$$

Theorem 4.9 (Convergence rate of Nesterov acceleration) *Given the iterates defined above, with:*

1. $\mu_k A_k = L$,
2. $A_k = \alpha_k^2 = (A_k - A_{k-1})^2$.

Then, $f(y_{k+1}) - f(x^*) \le \frac{L}{2}\frac{\|x_0 - x^*\|^2}{A_k}$ *and* $A_k \sim \frac{(k+1)^2}{4}$.

Proof:

$$\begin{aligned}
[f(y_{k+1}) - f(x^*)]A_k &\leq [f(y_{k+1}) - f(x^*)]A_k + \frac{L}{2}\|z_{k+1} - x^*\|^2 && \text{(Adding a nonnegative term)}\\
&\leq [f(y_k) - f(x^*)]A_{k-1} + \frac{L}{2}\|z_k - x^*\|^2 && \text{(From 4.17)}\\
&\leq \cdots \\
&\leq [f(y_0) - f(x^*)]\underbrace{A_0}_{=0} + \frac{L}{2}\left\|\underbrace{z_0}_{=x_0} - x^*\right\|^2\\
\implies f(y_{k+1}) - f(x^*) &\leq \frac{L}{2}\frac{\|x_0 - x^*\|^2}{A_k}
\end{aligned}$$

■

Aside

This is a proof by Lyapunov function. The goal of such proofs is to find a Lyapunov function denoted ϕ_k such that $\phi_k \leq \phi_{k-1}$. Applying this recursively implies that $\phi_k \leq \phi_0$.

Here, we let:

$$\phi_k := [f(y_k) - f(x^*)]A_k + \frac{L}{2}\|z_k - x^*\|^2 \tag{4.21}$$

Then, since it just corresponds to $[f(y_k) - f(x^*)]A_k$ + a nonnegative term:

$$[f(y_k) - f(x^*)]A_k \leq \phi_0$$
$$f(y_k) - f(x^*) \leq \frac{\phi_0}{A_k}$$

4.6 Nesterov Momentum

Idea: The goal is to combine the *dual-averaging* and *mixing* steps into one to only have 2 iterates.

We begin by setting the α_k, A_k such that:

$$A_k = \alpha_k^2, A_k = A_{k-1} + \alpha_k. \tag{4.22}$$

Then, the z_k iterates follow:

$$\begin{aligned}
z_{k+1} &= z_k - \frac{\alpha_k}{L}\nabla f(x_k)\\
&= (x_k - \frac{A_{k-1}y_k}{A_k})\frac{A_k}{\alpha_k} - \frac{\alpha_k}{L}\nabla f(x_k) && \text{(solving for } z_k \text{ in the mixing step 4.20).}\\
&= \underbrace{\frac{A_k}{\alpha_k}x_k - \frac{\alpha_k}{L}\nabla f(x_k)}_{=\alpha_k y_{k+1}} - \frac{A_{k-1}y_k}{\alpha_k} && \text{(using } \frac{A_k}{\alpha_k} = \alpha_k \text{ and the GD step 4.18).}\\
&= \alpha_k y_{k+1} + (1 - \alpha_k)y_k && (-\frac{A_{k-1}}{\alpha_k} \text{ can be rewritten as } (1 - \alpha_k) \text{ using 4.22).}
\end{aligned}$$

This can be rewritten as:

$$z_{k+1} - y_{k+1} = (1 - \alpha_k)[y_k - yk + 1] \tag{4.23}$$

Thus, the mixing term can be rewritten as:

$$\begin{aligned}
x_{k+1} &= \frac{A_k y_{k+1} + \alpha_{k+1} z_{k+1}}{A_{k+1}} \\
&= y_{k+1} + \frac{\alpha_{k+1}}{A_{k+1}}[z_{k+1} - y_{k+1}] && \text{(Using } \frac{A_k y_{k+1}}{A_{k+1}} = y_{k+1} - \frac{\alpha_{k+1}}{A_{k+1}} y_{k+1}\text{).} \\
&= y_{k+1} + \underbrace{\frac{\alpha_{k+1}}{A_{k+1}}}_{\frac{1}{\alpha_{k+1}}}(1 - \alpha_k)(y_k - y_{k+1}) && \text{(Using 4.23).} \\
&= y_{k+1} + \frac{1 - \alpha_k}{\alpha_{k+1}}(y_k - y_{k+1})
\end{aligned}$$

Thus, the mixing term can be written solely as a function of y_k, giving us the following updates:

Nesterov Momentum

$$y_{k+1} = x_k - \frac{1}{L}\nabla f(x_k) \tag{4.24}$$

$$x_{k+1} = y_{k+1} + \underbrace{\frac{1 - \alpha_k}{\alpha_{k+1}}}_{\frac{-k-1}{k+1}}[y_k - y_{k+1}] \tag{4.25}$$

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.