MILA Optimisation Crash Course **Winter 2024**

Lecture 2: February 21

Lecturer: Damien Scieur *Scribe(s): Mehran Shakerinava, 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 second 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.

2.1 Recapitulation

Last week we focused on minimising smooth functions and smooth convex functions. We saw that with simply an assumption of smoothness, we could use gradient descent to achieve convergence rate in the inequality 2.1. Further, when assuming a function was both smooth and convex, we were able to show that a much stronger convergence guarantee could be achieved, see the inequality 2.2. Note that the latter is stronger because we are guaranteed to be getting closer to the minima $x^\star$ on each iteration of the algorithm whereas the former just states that the norm of the gradient is bounded.

$$\min_{k\in[0,N]} \left\|\nabla f(x^k)\right\|^2 \leq \frac{2L}{N}(f(x^0) - f(x^\star)) \tag{2.1}$$

$$f(x^N) - f(x^*) \leq \frac{1}{N}\frac{L}{2}\left\|(x^* - x^0)\right\|^2 \tag{2.2}$$

2.2 Introduction

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. However, we will assume our function f is non-smooth, and as such, may not be upper-bounded by a parabola [1]. In this setting, we still want to show some guarantees on the rate of convergence (e.g., $\nabla f(x^k) \to 0$, or $f(x^k) - f(x^\star) \to 0$, etc.). Figure 2.2 illustrates the difficulties of optimising a non-smooth function, namely, oscillation in the iterates.

This lecture starts by introducing non-smooth optimisation and subgradients. We then see how to derive convergence guarantees based on the definitions of subgradient descent and convexity. Then, to bypass the slow convergence rate of subgradient descent, we introduce dual averaging and show how to derive an optimisation algorithm for the method [2].

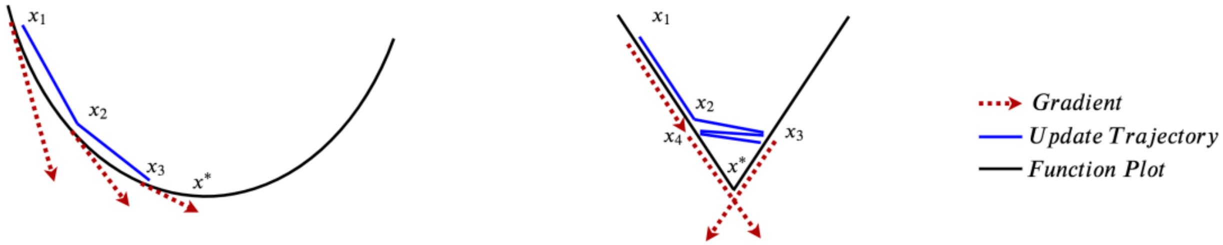

Figure 2.1: **Left**: Smooth function optimised via gradient descent. **Right**: Non-smooth function optimised via gradient descent showing the oscillatory nature of the sequence of points. Stolen from [1].

2.3 Preliminaries

Gradient descent is characterized by the following equation.

$$x^{k+1} = x^k - \alpha \nabla_x f(x^k) \tag{2.3}$$

Lemma 2.1 *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 2.2 *A continuously differentiable function $f : \mathbb{R}^p \to \mathbb{R}$ with $x, y \in \mathbb{R}^p$ is L-smooth if*

$$\|\nabla f(y) - \nabla f(x)\| \leq L\|y - x\|. \tag{2.4}$$

From [1] and [3].

Definition 2.3 *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*

$$f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha)f(y). \tag{2.5}$$

Lemma 2.4 *A convex function f is lower-bounded by a line, i.e.,*

$$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle. \tag{2.6}$$

2.4 Non-smooth Functions and their Minima

Non-smooth functions appear all over the place in optimisation problems. Below are some examples.

(a) Let $f(x) = |x|$. We have

$$\text{“}\nabla f(x)\text{”} = \begin{cases} 1 & \text{if } x \geq 0 \\ -1 & \text{if } x \leq 0 \end{cases}.$$

Then $\|\nabla f(-\varepsilon) - \nabla f(\varepsilon)\| = |1 - (-1)| = 2 \nleq L \cdot 2\varepsilon$ and f is non-smooth.

(b) Maximum of linear functions:

$$\min_x \max_i a_i^T x$$

(c) Min-max bilinear games:

$$\min_{x \in \Delta_n} \max_{y \in \Delta_m} x^T A y$$

(d) Others:

- $\|x\|_1$
- $\|x\|_0$
- $\text{ReLU}(x)$
- $-\log(x)$
- $\sqrt{x}$
- $\min_x \max_i f_i(x)$
- More...

A non-smooth function cannot be bounded from above by a parabola in the way that we found in the last lecture. Instead, when optimising non-smooth functions, we must rely exclusively on the lower bound of Lemma 2.4 that results from the convexity of the function. Note, however, that for this lower bound to hold, we require $\nabla f(x)$, which may not exist in certain portions of the domain of a non-smooth function (i.e., a non-smooth function is not differentiable everywhere). To counteract this difficulty, we use subgradients.

Definition 2.5 *The subgradients of a non-smooth function f at a point $x \in \mathbb{R}^d$ is the set of vectors $\partial f(x)$ such that convexity holds. That is, $\partial f(x)$ is the set of subgradients of $f(x)$ if*

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

For example, if $f(x) = |x|$, then the subgradients of $f(x)$ would be the set

$$\partial f(x) = \begin{cases} \{1\} & \text{if } x > 0 \\ \{-1\} & \text{if } x < 0 \\ [-1, 1] & \text{if } x = 0 \end{cases}. \tag{2.7}$$

Notice how at $x = 0$, the subgradient can take any value between -1 and 1. This is because any of those values satisfy the convexity lower bound.

Assuming that f is convex, and that the domain of f is a convex set Q, we make two further assumptions on the subgradients of f.[1]

1. The subgradients must be bounded. This means we are dealing with finite subgradients, or equivalently, that $\nabla f(x)$ is Lipshitz. $G \in \mathbb{R}$.

$$\max_{x \in Q} \|\partial f(x)\| \leq G$$

[1] For more on convex sets see [3].

2. The distance between any two points in the domain of the function (what we have called the convex set Q), is bounded by the diameter of the set, $D \in \mathbb{R}$.

$$\max_{x \in Q} \frac{\left\|x - x^0\right\|^2}{2} \leq D$$

Subgradient descent is defined similarly to gradient descent except, as Damien said, "we don't have gradients so we replace them with subgradients and cross our fingers." Note that the learning rate α_k is no longer a constant. Since non-smooth functions have discontinuities, the norm of the gradient may not decrease as we approach the minimum. Thus, to encourage subgradient descent to converge, we must continually decrease the learning rate. Refer again to the absolute value example from Equation 2.7 and Figure 2.2 to understand this phenomenon better.

Definition 2.6 *Subgradient descent is characterised by the following, where $g_k \in \partial f(x)$*

$$x^{k+1} = x^k - \alpha_k g_k \tag{2.8}$$

Theorem 2.7 *Convex non-smooth functions optimised with subgradient descent converge with the following guarantee, where x^{BEST} is the best iterate found over k steps.*

$$f(x^{BEST}) - f(x^\star) \leq \frac{1}{2}\mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$$

Proof: We start by expanding the squared norm $\left\|x^{k+1} - x^\star\right\|^2$, where $x^{k+1} = x^k - \alpha_k g_k$ via Definition 2.6. Then we use a property of convexity via Lemma 2.4.

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

Notice that $\left\|x^k - x^\star\right\|^2$ can be expanded similarly to $\left\|x^{k+1} - x^\star\right\|^2$. This is a recurrence which can be expanded to yield the following

$$\begin{aligned}
\left\|x^{k+1} - x^\star\right\|^2 &\leq \left\|x^k - x^\star\right\|^2 + \alpha_k^2 \left\|g_k\right\|^2 - 2\alpha_k \left(f(x^k) - f(x^\star)\right) \\
&= \left\|x^{k-1} - \alpha_{k-1} g_{k-1} - x^\star\right\|^2 + \alpha_k^2 \left\|g_k\right\|^2 - 2\alpha_k \left(f(x^k) - f(x^\star)\right) \\
&= \left\|x^{k-1} - x^\star\right\|^2 + \alpha_{k-1}^2 \left\|g_{k-1}\right\|^2 + \alpha_k^2 \left\|g_k\right\|^2 - 2\alpha_k \left(f(x^k) - f(x^\star)\right) - 2\alpha_{k-1}\langle g_{k-1}, x^{k-1} - x^\star\rangle \\
&\leq \left\|x^{k-1} - x^\star\right\|^2 + \alpha_{k-1}^2 \left\|g_{k-1}\right\|^2 + \alpha_k^2 \left\|g_k\right\|^2 - 2\alpha_k \left(f(x^k) - f(x^\star)\right) - 2\alpha_{k-1} \left(f(x^{k-1}) - f(x^\star)\right) \\
&\ldots
\end{aligned}$$

$$\left\|x^{k+1} - x^\star\right\|^2 \leq \left\|x^0 - x^\star\right\|^2 + \sum_{i=0}^{k} \alpha_i^2 \left\|g_i\right\|^2 - 2\sum_{i=0}^{k} \alpha_i \left(f(x^i) - f(x^\star)\right)$$

Using the assumptions from above, we know that $\left\|x^0 - x^\star\right\|^2 \leq D$, $\left\|g_i\right\|^2 \leq G^2$, and $0 \leq \left\|x^{k+1} - x^\star\right\|^2$. In addition, note that $f(x^{\text{BEST}}) - f(x^\star) \leq f(x^i) - f(x^\star)$. Plugging these in and rearranging, we get the following.

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

To finish the proof, we need to define a learning rate schedule that takes the bracketed term on the right hand side to zero. Since we know that the learning rate must decrease, we define $\alpha_i \approx \mathcal{O}\left(\frac{1}{\sqrt{i+1}}\right)$. However, for the simplicity of analysis, we assume from the start that we'll take k steps and choose $\alpha_i = \frac{1}{\sqrt{k+1}}$. We get $\sum_{i=0}^{k} \alpha_i = \sqrt{k+1}$ and $\sum_{i=0}^{k} \alpha_i^2 = 1$. In the next lecture we will investigate what happens with different settings of the learning rate. For more information refer to [4]. Now, we can state the result from the theorem.

$$
\begin{aligned}
f(x^{\text{BEST}}) - f(x^\star) &\leq \frac{1}{2} \left(D + G^2 \right) \frac{1}{\sqrt{k+1}} \\
&\in \frac{1}{2} \mathcal{O}\left(\frac{1}{\sqrt{k+1}} \right)
\end{aligned}
$$

■

Note that subgradient descent doesn't ensure that the iterates are monotonically decreasing as we saw with gradient descent on smooth or convex functions. This is a direct consequence of not having a smoothness-based upper-bound on the function. Note as well that the learning rate α_i should decrease on each iteration, leading to the k$^{\text{th}}$ subgradient g_k contributing less to the optimisation than prior subgradients despite being closer to the minimum (and likely more informative). That is, the convergence rate is slow. To fix this issue, dual averaging uses the definition of convexity to construct an average lower-bound on the function.

2.5 Dual Averaging

Intuition. Gradient descent uses upper bounds while dual averaging uses lower bounds.

Convexity provides lower bounds at every visited point x_i.

$$
\begin{aligned}
f(x) &\geq f(x^0) + g_0^T (x - x^0) \\
f(x) &\geq f(x^1) + g_1^T (x - x^1) \\
&\vdots \\
f(x) &\geq f(x^k) + g_k^T (x - x^k)
\end{aligned}
$$

Taking a weighted average of these inequalities we arrive at

$$
f(x) \geq \frac{\sum_i \alpha_i (f(x^i) + g_i^T (x - x^i))}{\sum_i \alpha_i}. \tag{2.9}
$$

Moreover,

$$\min_{x\in\mathcal{X}} f(x) \geq \min_{x\in\mathcal{X}} \frac{\sum_i \alpha_i(f(x^i) + g_i^T(x - x^i))}{\sum_i \alpha_i}. \tag{2.10}$$

We will try to find a "dual" solution that lower bounds $f^\star$.

Problem: If $\mathcal{X}$ is unbounded (e.g., $\mathbb{R}^d$), then $\min_{x\in\mathcal{X}} \text{Linear}(x) = -\infty$. To mitigate this, we add a regularisation term to Equation 2.9: $\frac{\mu_k}{2}\left\|x - x^0\right\|^2$.

Proposition 2.8 *Dual averaging (first ingredient). The update rule for dual averaging is achieved my the following minimisation problem*

$$x^{k+1} = \arg\min_{x\in\mathbb{R}^d} \left\{ \frac{\sum_i \alpha_i(f(x^i) + g_i^T(x - x^i))}{\sum_i \alpha_i} + \frac{\mu_k}{2}\left\|x - x^0\right\|^2 \right\}$$

Proof: We begin by taking the gradient with respect to x, over the terms in the arg min. Let $\alpha_i = 1$, let $\mu_k = \frac{1}{\sqrt{1+k}}$, and let $S^{k+1} = \sum_{j=0}^k g_j$

$$\begin{aligned}
0 &= \nabla_x \left(\frac{\sum_{i=0}^k \alpha_i(f(x^i) + g_i^T(x - x^i))}{\sum_{i=0}^k \alpha_i} + \frac{\mu_k}{2}\left\|x - x^0\right\|^2 \right) \\
&= \frac{1}{1+k}\left(\sum_{i=0}^k (\nabla_x f(x^i) + \nabla_x g_i^T(x - x^i)) \right) + \frac{1}{2\sqrt{1+k}} \nabla_x \left\|x - x^0\right\|^2 \\
&= \frac{1}{1+k}\left(\sum_{i=0}^k g_i \right) + \frac{1}{\sqrt{1+k}}(x - x^0) \\
x &= x^0 - \frac{1}{\sqrt{k+1}}\left(\sum_{i=0}^k g_i \right) \\
x &= x^0 - \frac{1}{\sqrt{k+1}} S^{k+1}
\end{aligned}$$

Finally, with $x^{k+1} \leftarrow x$, we find the update rule for dual averaging.

$$x^{k+1} = x^0 - \frac{1}{\sqrt{k+1}} S^{k+1} \tag{2.11}$$

■

Typically $\alpha_i = 1$ and $\mu_k = \frac{1}{\sqrt{k+1}}$. The algorithm outputs $\hat{x}^k = \frac{\sum_{i=0}^k x^i}{k+1}$.

What if the domain is not bounded? The following theorem shows that we can get by without the bounded domain assumption.

Theorem 2.9 *If $\alpha_i = 1$ and $\mu_k = O\left(\frac{1}{\sqrt{k}}\right)$, then*

$$\frac{1}{2}\left\|x^k - x^\star\right\|^2 \leq \left\|x^0 - x^\star\right\|^2 + O(G^2), \tag{2.12}$$

where G satisfies $\|\nabla f(x)\| \leq G$ for all x.

Proof: Starting from the dual averaging update rule of Equation 2.11 and working from $\frac{1}{2}\left\|x^k - x^\star\right\|^2$ we find the norm is bounded.

$$\begin{aligned}
\frac{1}{2}\left\|x^k - x^\star\right\|^2 &= \frac{1}{2}\left\|x^0 - S^k - x^\star\right\|^2 \\
&= \frac{1}{2}\left\|x^0 - x^\star\right\|^2 + \frac{1}{2}\left\|S^{k+1}\right\|^2 - \langle S^{k+1}, x^0 - x^\star\rangle \\
&\leq \frac{1}{2}\left\|x^0 - x^\star\right\|^2 + \frac{1}{2}\left\|S^{k+1}\right\|^2 - \langle S^{k+1}, x^0 - x^\star\rangle + \frac{1}{2}\left\|x^0 + S^{k+1} - x^\star\right\|^2 \\
&= \left\|x^0 - x^\star\right\|^2 + \left\|S^{k+1}\right\|^2 \\
&= \left\|x^0 - x^\star\right\|^2 + \left\|\sum_{j=0}^{k} \frac{1}{\sqrt{k+1}} g_j\right\|^2 \\
&\leq \left\|x^0 - x^\star\right\|^2 + \sum_{j=0}^{k} \left\|\frac{1}{\sqrt{k+1}} g_j\right\|^2 \\
&\leq \left\|x^0 - x^\star\right\|^2 + \frac{1}{k+1}\sum_{j=0}^{k} G^2 \\
&= \left\|x^0 - x^\star\right\|^2 + G^2
\end{aligned}$$

Therefore, $\left\|x_k - x^\star\right\|^2$ is always bounded. ∎

The next theorem bounds the error of the dual averaging method.

Theorem 2.10 *If $\alpha_i = 1$ and $\mu_k = O\left(\frac{1}{\sqrt{k}}\right)$, then*

$$f\left(\frac{\sum_{i=0}^{k} x_i}{k+1}\right) - f^\star \leq O\left(\frac{1}{\sqrt{k+1}}\right) \cdot (D + \frac{G^2}{2}) \tag{2.13}$$

Proof:

$$\begin{aligned}
0 = \frac{1}{2}\left\|x^{k+1} - x^{k+1}\right\|^2 &= \frac{1}{2}\left\|x^0 - \frac{S^{k+1}}{\sqrt{k+1}} - x^{k+1}\right\|^2 \\
&= \frac{1}{2}\left\|x^0 - x^{k+1}\right\|^2 + \frac{1}{2}\left\|\frac{S^{k+1}}{\sqrt{k+1}}\right\|^2 - \langle \frac{S^{k+1}}{\sqrt{k+1}}, x^0 - x^{k+1}\rangle \\
&\leq \frac{1}{2}\left(2D + G^2\right) + \frac{1}{\sqrt{k+1}}\sum_{j=0}^{k}\langle g_j, x^0 - x^{k+1}\rangle
\end{aligned}$$

Now, solving from below, we find an inequality for the inner product, since we know that the right-hand side is solved by x^{k+1}

$$\begin{aligned}
\min_{x\in\mathcal{X}} f(x) &\geq \min_{x\in\mathcal{X}} \frac{\sum_{i=0}^{k} \alpha_i (f(x^i) + \langle g_i, (x - x^i)\rangle)}{\sum_{i=0}^{k} \alpha_i} \\
f^\star &\geq \frac{\sum_{i=0}^{k} (f(x^i) + \langle g_i, (x^{k+1} - x^i)\rangle)}{k+1} \\
(k+1)f^\star &\geq \sum_{i=0}^{k} (f(x^i) + \langle g_i, (x^{k+1} - x^i)\rangle) \\
(k+1)f^\star - \sum_{i=0}^{k} f(x^i) &\geq \sum_{i=0}^{k} \langle g_i, (x^{k+1} - x^i)\rangle
\end{aligned}$$

Continuing from the previous inequalities using this result, and then applying the definition of convexity 2.3,

$$\begin{aligned}
0 &\leq \frac{1}{2}\left(2D + G^2\right) + \frac{1}{\sqrt{k+1}} \sum_{j=0}^{k} \langle g_j, x^0 - x^{k+1}\rangle \\
&\leq \frac{1}{2}\left(2D + G^2\right) + \frac{1}{\sqrt{k+1}} \left((k+1)f^\star - \sum_{i=0}^{k} f(x^i) \right) \\
\frac{k+1}{k+1} \sum_{i=0}^{k} f(x^i) - (k+1)f^\star &\leq \frac{\sqrt{k+1}}{2}\left(2D + G^2\right) \\
(k+1) f\left(\sum_{i=0}^{k} \frac{x^i}{k+1} \right) - (k+1)f^\star &\leq \frac{\sqrt{k+1}}{2}\left(2D + G^2\right) \\
f\left(\sum_{i=0}^{k} \frac{x^i}{k+1} \right) - f^\star &\leq \left(D + \frac{G^2}{2} \right) \frac{1}{\sqrt{k+1}}
\end{aligned}$$

This completes the proof and gives us the final step to understand Algorithm 1. ■

Algorithm 1 Dual Averaging Algorithm

```
Initialize $x^0$, $S_0 = \partial f(x^0)$.
for $i = 0$ to $k$ do
    $S_{i+1} = S_i + \partial f(x^i)$
    $x^{i+1} = x^0 - \frac{S_{i+1}}{\sqrt{i+1}}$
    $\hat{x}^{i+1} = \frac{i}{i+1}\hat{x}^i + \frac{1}{i+1}x^{i+1}$
end for
return $\hat{x}^{k+1}$
```

Exercises

1. Re-derive bound for subgradient method.

2. Prove formula for $x^{k+1} = \arg\min_{x\in\mathbb{R}^d} \left\{ \frac{\sum_i \alpha_i (f(x^i) + g_i^T(x - x^i))}{\sum_i \alpha_i} + \frac{\mu_k}{2} \left\| x - x^0 \right\|^2 \right\}$.

References

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

[2] Y. Nesterov, "Primal-dual subgradient methods for convex problems," *Math. Program.*, vol. 120, pp. 221–259, 2009.

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

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