MILA Optimisation Crash Course **Fall 2024**

Lecture 1: September 4

Lecturer: Ryan D'Orazio *Scribe(s): Danilo Vucetic*

These are the scribe notes for the second edition of the optimisation crash course at MILA organised by Lucas Maes, Helen Zhang, 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. The course webpage can be accessed here.

1.1 Introduction

We start the second edition of the optimisation crash course with an introduction to some fundamental topics in convex analysis. Convex analysis studies convex functions and convex sets. It is often used to underpin proofs and provide intuition towards algorithms in convex optimisation. For instance, using convex analysis we can often simplify the arguments of a proof, or change the way we view a problem so that it is an "application, not a complication." That is to say, we can view optimisation problems as applications of ideas in convex analysis, and not complications of the problem itself. A great example of this is the application of constraints to an optimisation problem. We can often rewrite the constraints as part of the objective, rather than a constraint on the optimisation process, thus simplifying the process of finding a solution.

This lecture will first introduce convex sets and functions, defining some fundamental properties that can be used to derive the convergence bounds of gradient descent, among others. Then, subgradients are introduced and an example is provided to show how the combination of subgradients with constrained optimisation problems can allow us to simplify these problems.

1.2 Convex sets and functions

We start by defining various sets that will be of interest to us in this lecture. Then, convex functions are defined, as well as some properties of convex functions. We follow Beck [2017, pg. 3, 13-25].

Definition 1.1 (Affine set) *An affine set, S contains all lines going through each pair of points in the set:* $\forall x, y \in S, \lambda \in \mathbb{R},\ \lambda x + (1-\lambda)y \in S$

For example, if we had an affine S and the two unit vectors of $\mathbb{R}^2 : (0,1), (1,0) \in S$, we could then infer that the set S contained (at least) all points on the line $y = -x + 1$. In fact, this inference is the exact definition of the affine hull between the two points!

Definition 1.2 (Affine hull) *An affine hull,* $\text{aff}(S)$ *contains all the affine combinations of elements in S (not necessarily an affine set).*

$$\text{aff}(S) = \left\{ \sum_{i=1}^{K} \lambda_i x_i : K > 0, x_i \in S, \lambda_i \in \mathbb{R}, \sum_{i=1}^{K} \lambda_i = 1 \right\}$$

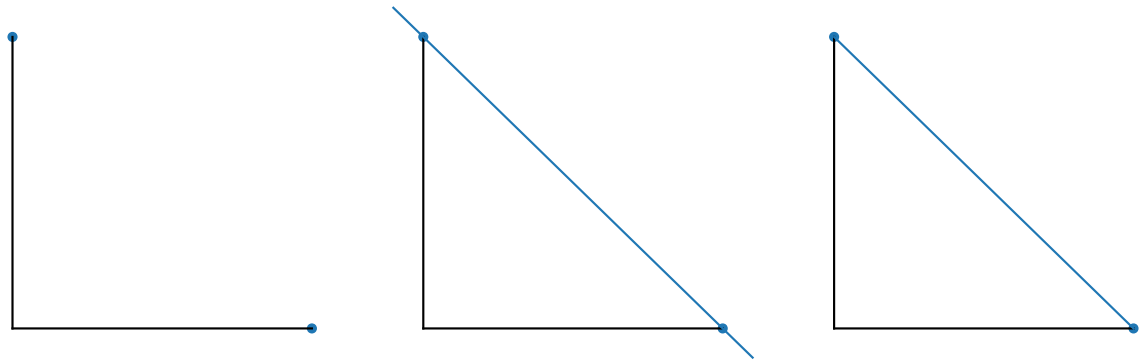

Figure 1.1: **(Left)** The points $(0,1),(1,0)$ in a set, presented on the x-y axis (black), **(middle)** The affine hull of the set, **(right)** the convex hull of the set.

Convex sets and hulls can be defined similarly to affine sets and hulls just with the added constraint that $\lambda \in [0,1]$.

Definition 1.3 (Convex set) *A convex set, C contains all lines between each pair of points in the set:* $\forall x, y \in S, \lambda \in [0,1],\ \lambda x + (1-\lambda)y \in C$

Definition 1.4 (Convex hull) *A convex hull,* $\text{conv}(S)$ *contains all the convex combinations of elements in S (not necessarily a convex set).*

$$\text{conv}(S) = \left\{ \sum_{i=1}^{K} \lambda_i x_i : K > 0, x_i \in S, \lambda_i \in [0,1], \sum_{i=1}^{K} \lambda_i = 1 \right\}$$

See Figure 1.1 for images corresponding to each of these concepts. Note that an affine set is convex, but a convex set cannot be affine. For more information on these sets, such as their properties, see Beck [2017, pg. 3].

To define a convex function, we first need to understand epigraphs and domains.

Definition 1.5 (Domain of a function) *The domain of a function $f : \mathbb{R}^n \to (-\infty, +\infty]$ is the set of all inputs x to the function that evaluate to finite outputs:* $\text{dom}(f) = \{x : f(x) < \infty\}$.

Definition 1.6 (Epigraph of a function) *The epigraph of a function $f : \mathbb{R}^n \to (-\infty, +\infty]$ is the set of all points that lie above the function:* $\text{epi}(f) = \{(x, y) : x \in \mathbb{R}^n, y \in \mathbb{R}, f(x) \leq y\}$.

Note that we are using functions $f : \mathbb{R}^n \to (-\infty, +\infty]$. These are called proper functions since they don't reach $-\infty$, and they are closed if the epigraph of the function is closed. The notion of closedness can be further explored under its equivalency with lower semi-continuous functions, again, see Beck [2017, pg. 15-16]. We can also show that proper closed functions attain a minimum over some set of inputs (must be compact); for more details on this topic see Theorem 2.12 of Beck [2017]. We are finally ready to define convex functions.

Definition 1.7 (Convex function) *A function $f : \mathbb{R}^n \to (-\infty, +\infty]$ is convex if its epigraph is a convex set.*

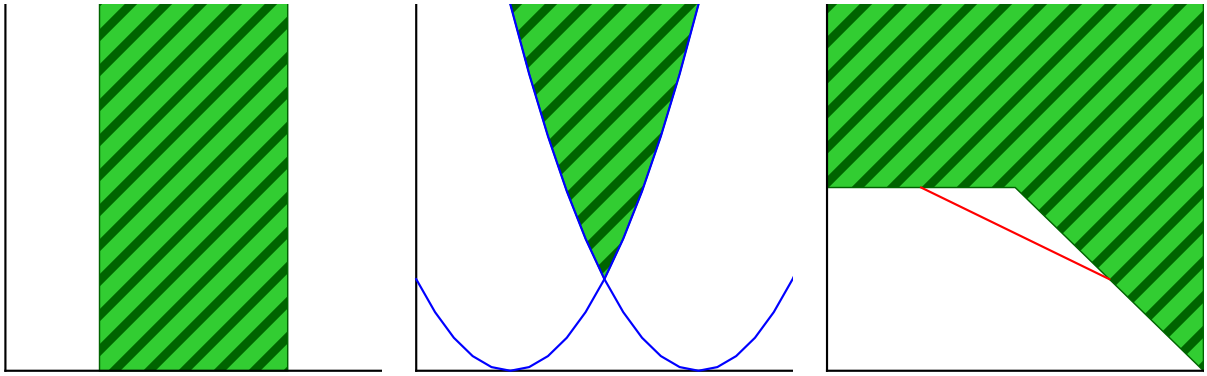

Figure 1.2: **(Left)** The epigraph of an indicator function on the set $[1, 2]$. **(Middle)** The epigraph of the maximum between two quadratic functions centered at 0.5 and 1.5. **(Right)** The epigraph of a piecewise (upside down ReLU) function, where the red line indicates that there exist convex combinations between points in the epigraph not in the epigraph, making this function non-convex.

Simple examples of convex functions include the indicator function δ_C, quadratic functions, and the maximum over convex functions. Examples of convex and non-convex functions are illustrated in Figure 1.2 By example of the indicator function, we see that its epigraph is convex so long as the set, C, is convex. Note, a set with "holes" would allow for convex combinations not in the epigraph, *e.g.*, the set of points including $[0, 1], [1.5, 2]$ would not have a convex indicator function.

$$\delta_C(x) = \begin{cases} 0 \text{ if } x \in C \\ \infty \text{ else} \end{cases}$$

$$\text{epi}(\delta_C) = \{(x, y) : x \in \mathbb{R}^n, y \in \mathbb{R}, \delta_C(x) \leq y\} = C \times \mathbb{R}^+$$

We now move on to a fundamental property of convex functions which is used extensively in convex optimisation for proofs of convergence. The following proposition states that any line drawn between two points in a convex curve will be above the curve.

Proposition 1.8 *f is convex if and only if* $\forall x, y \in \text{dom}(f), \forall \lambda \in [0, 1]$

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

Proof: $\Rightarrow$ We already know that the convex combination of any two points in the epigraph of a convex function is within the epigraph. Thus, we know that $\forall v_1 = (x_1, y_1), v_2 = (x_2, y_2) \in \text{epi}(f), \forall \lambda \in [0, 1]$

$$v_3 = \lambda v_1 + (1 - \lambda)v_2 \in \text{epi}(f)$$

Since v_1, v_2, v_3 are within the epigraph, and $v_3 = (x_3, y_3)$, we know that evaluating the function at the point x_3 will yield a value within the epigraph. Moreover, the convex combination of y_1, y_2 must lie above the function:

$$f(x_3) \leq y_3$$
$$f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda y_1 + (1 - \lambda)y_2$$

We are free to choose y_1, y_2, so we set them to $f(x_1), f(x_2)$ respectively, and the result is achieved.

$\Leftarrow$ We have that $f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda)f(x_2)$. Let $f(x_1) \leq y_1$ and $f(x_2) \leq y_2$. We trivially recover that

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda)f(x_2) \leq \lambda y_1 + (1-\lambda)y_2$$
$$f(x_3) \leq y_3$$

As such, we can recover the convex epigraph of f, meaning that f is a convex function. ■

There are other interesting properties of convex functions, such as convexity under a partial minimum, but we will not cover those topics in these notes. For more information, see Section 2.3 in Beck [2017].

1.3 Subgradients

Subgradients are particularly useful when optimising functions with discontinuous gradients. This is because they generalise the notion of a gradient to allow for these discontinuities. For more information on non-smooth optimisation, see the second lecture of the Winter 2024 crash course. We will use a slightly different definition of subgradients in these notes.

Definition 1.9 (Subgradient inequality) *Let $f : \mathbb{R}^n \to (-\infty, +\infty]$ be a proper function and let $x \in \text{dom}(f)$. A subgradient of f at the point x is a vector g such that*

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

Definition 1.9 is essentially saying that for a subgradient g, tangent to the surface of the function at x, there exists a function that underestimates the value of $f(y)$, *i.e.,* the right hand side of the inequality. There may exist an infinite number of subgradients at the point x. As such, we collect them into a set called the subdifferential. A quintessential example of a function that obeys these properties is the absolute value function $f(x) = |x|, \forall x \in \mathbb{R}$. In this case, we can select subgradients $g \in [-1, 1]$ for $x = 0$.

Definition 1.10 (Subdifferential) *The set of all subgradients of f at x is called the subdifferential of f at x and is denoted by $\partial f(x)$:*

$$\partial f(x) = \{g \in \mathbb{R}^n : \forall y \in \mathbb{R}^n, f(y) \geq f(x) + \langle g, y - x \rangle\}$$

Definition 1.11 (Subdifferentiability) *A function $f : \mathbb{R}^n \to (-\infty, +\infty]$ is called subdifferentiable at $x \in \text{dom}(f)$ if $\partial f(x) \neq \emptyset$.*

Turning again to the indicator function, $\delta_C(x)$ for some $x, y \in C$, we may find its subgradients as follows. The final line states the subdifferential of the indicator function. Note that this is also called the normal cone.

$$\begin{aligned} \delta_C(y) &\geq \delta_C(x) + \langle g, y - x \rangle \\ 0 &\geq \langle g, y - x \rangle \\ \partial \delta_C(x) &= \{g \in \mathbb{R}^n : \forall y \in C, 0 \geq \langle g, y - x \rangle\} = N_C(x) \end{aligned}$$

We will now state a series of results concerning the non-emptiness of the subdifferential set. These results essentially state that if a function is convex, it is subdifferentiable over its domain, *i.e.,* it has subgradients. For these results, we have to define the interior of a set and the relative interior of a set.

Definition 1.12 (Interior of a set) *The interior of a subset $S \subseteq X$ of some space X is the union of all subsets of S that are open in X. We define the interior with respect to ϵ-balls $B_\epsilon(x) = \{p \in S : \|x - p\| \leq \epsilon\}$.*

$$\text{int}(S) = \{x \in S : \exists \epsilon > 0, B_\epsilon(x) \subseteq S\}$$

Definition 1.13 (Relative interior of a set) *The relative interior of a subset $S \subseteq X$ of some space X is the interior of the set within the affine hull of the set.*

$$\text{relint}(S) = \{x \in S : \exists \epsilon > 0, B_\epsilon(x) \cap \text{aff}(S) \subseteq S\}$$

Examples of interiors and relative interiors follow. Notice how the relative interior is non-empty when the subset S is a lower-dimensional subset in a higher-dimensional space.

- $S = \{x \in \mathbb{R} : 0 \leq x \leq 1\}$. The interior $\text{int}(S) = (0, 1)$ and the relative interior $\text{relint}(S) = (0, 1)$ are the same because the ϵ-ball in both cases is in a single dimension, overlapping with S.
- $S = \left\{(x, y) \in \mathbb{R}^2 : y = 0, 0 \leq x \leq 1\right\}$. The ϵ-ball is now a set in two dimensions that does not form an entire subset of S. As such, the interior is empty: $\text{int}(S) = \emptyset$. The relative interior, however, is still $\text{relint}(S) = (0, 1)$.
- When S is the probability simplex $\Delta^n = \text{conv}(\{e_1, e_2, \cdots, e_n\})$ in n dimensions (see the right of Figure 1.1, for example), it occupies an $(n-1)$-dimensional subset in n dimensions. Thus, the interior of S will be empty, but the relative interior will still be defined.

We are now ready to state the non-emptiness results.

Theorem 1.14 (Non-emptiness and boundedness at interior points) *Let $f : \mathbb{R}^n \to (-\infty, +\infty]$ be a proper convex function and assume that $x \in \text{int}(\text{dom}(f))$. Then, $\partial f(x)$ is non-empty and bounded. See [Beck, 2017, Theorem 3.14] for the proof.*

$$\text{int}(\text{dom}(f)) \subseteq \text{dom}(\partial f)$$

Theorem 1.15 (Non-emptiness and boundedness in relative interior) *Let $f : \mathbb{R}^n \to (-\infty, +\infty]$ be a proper convex function and assume that $x \in \text{relint}(\text{dom}(f))$. Then, $\partial f(x)$ is non-empty and bounded. See [Beck, 2017, Theorem 3.18].*

$$\text{relint}(\text{dom}(f)) \subseteq \text{dom}(\partial f)$$

Since the relative interior of $\text{dom}(f)$ is always non-empty, we can conclude that there always exists a point in the domain that is subdifferentiable. This result allows us to optimise over lower-dimensional spaces when operating in higher dimensions!

The final three results state some important facts surrounding the relationship between subgradients and gradients, the definition of a descent direction, and an important optimality condition.

Theorem 1.16 (Subdifferential at points of differentiability) *Let $f : \mathbb{R}^n \to (-\infty, +\infty]$ be a proper convex function and let $x \in \text{int}(\text{dom}(f))$. If f is differentiable at x, then $\partial f(x) = \{\nabla f(x)\}$. Conversely, if f has a unique subgradient at x, then f is differentiable at x and $\partial f(x) = \{\nabla f(x)\}$. See [Beck, 2017, Theorem 3.33].*

Definition 1.17 (Descent property of descent directions) *Let $f : \mathbb{R}^n \to (-\infty, +\infty]$ be a proper convex function and let $x \in \text{int}(\text{dom}(f))$ and assume that $d \neq 0 \in \mathbb{R}^n$ is a descent direction of f at x. Then there exists some $\epsilon > 0$ such that $x + \eta d \in \text{dom}(f)$ $\eta \in (0, \epsilon]$ and*

$$f(x + \eta d) \leq f(x)$$

See [Beck, 2017, Lemma 8.2] for more.

Theorem 1.18 (Fermat's optimality condition) *Let $f : \mathbb{R}^n \to (-\infty, +\infty]$ be a proper convex function. Then*

$$x^\star \in \arg\min \{f(x) : x \in \mathbb{R}^n\}$$

if and only if $0 \in \partial f(x^\star)$.

Proof: From the definition of a subgradient, we have that

$$f(x) \geq f(x^\star) + \langle g, x - x^\star \rangle$$

But the only way for this to be true is if $g = 0$. Thus, $0 \in \partial f(x^\star)$. ■

1.4 Closing example

Say that we are given a constrained optimisation problem with a constrained convex set $C \subseteq \mathbb{R}^n$ and a proper convex function $f : \mathbb{R}^n \to (-\infty, +\infty]$:

$$\min_{x \in C} f(x) \text{ s.t. } f(x) = \sum_{i=1}^{N} f_i(x) \tag{1.1}$$

It is quite ordinary in machine learning to construct losses as sums over data points, i, in a data set of size N. It is also common to constrain the solution set, for instance, we often apply "soft constraints" (*i.e.,* regularisation) to prevent an explosion in gradients and to make our functions "nicer".

We may restate Equation 1.1 using the indicator function to enforce the constraints.

$$\min f(x) + \delta_C(x) \tag{1.2}$$

This form allows for the classical approach to finding a solution: take a (sub)gradient and solve for zero! According to Fermat's optimality condition, Theorem 1.18, global minima must have $0 \in \partial \left(f(x^\star) + \delta_C(x^\star) \right)$. However, we require one more result to split subdifferential between the sum.

Theorem 1.19 (Sum rule for subdifferential calculus) *Let $f_1, f_2, \cdots, f_m : \mathbb{R}^n \to (-\infty, +\infty]$ be proper convex functions and assume that $\bigcap_{i=1}^m \text{relint}(\text{dom}(f_i)) \neq \emptyset$ (i.e., the functions have overlapping domains and subgradients by Theorem 1.15). Then for any $x \in \mathbb{R}^n$*

$$\partial \left(\sum_{i=1}^{m} f_i \right)(x) = \sum_{i=1}^{m} \partial f_i(x)$$

See [Beck, 2017, Theorem 3.40] for more.

Using this theorem, we achieve the following.

$$0 \in \partial \left(f(x^\star) + \delta_C(x^\star) \right)$$
$$0 \in \partial f(x^\star) + \partial \delta_C(x^\star)$$

Clearly, some vector $g \in \mathbb{R}^n$ must exist such that $g \in \partial f(x^\star)$ and $-g \in \partial \delta_C(x^\star)$. However, we note that $\partial \delta_C(x) = \{g \in \mathbb{R}^n : \forall y \in C, 0 \geq \langle g, y - x \rangle\} = N_C(x)$ and since we assume that $x, y \in C, x \neq y$, then the only setting of g is zero! Hence, if a global minimum of f is within the constrained set C, the minimum will remain unchanged by the application of constraints.

References

Amir Beck. *First-Order Methods in Optimization*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2017. doi: 10.1137/1.9781611974997. URL https://epubs.siam.org/doi/abs/10.1137/1.9781611974997.