Another Math Primer: Convexity (and Concavity)

1 Continuous Optimization

Recall the form of a(n unconstrained) continuous optimization problem: find $\boldsymbol{x}^* \in \mathbb{R}^d$ such that

$$f(\boldsymbol{x}^*) = \min_{\boldsymbol{x} \in \mathbb{R}^d} f(\boldsymbol{x})$$

Once again, $\mathbb{R}^d$ is a finite-dimensional vector space ($d \in \mathbb{N}$) with elements $\boldsymbol{x} = (x_1, \ldots, x_d)$ and $f : \mathbb{R}^d \to \mathbb{R}$.

Also recall the definitions of global and local minima. A **global minimum** $\boldsymbol{x}^* \in \mathbb{R}^d$ is a point such that $f(\boldsymbol{x}^*) \leq f(\boldsymbol{x})$, for all $\boldsymbol{x} \in \mathbb{R}^d$. More generally, a **local minimum** $\tilde{x} \in \mathbb{R}^d$ is a point such that $f(\tilde{x}) \leq f(\boldsymbol{x})$, for all $\boldsymbol{x} \in \mathbb{R}^d$ within some small distance (i.e., neighborhood) of $\tilde{x}$.

This lecture begins our discussion on convexity. In it, we define **convex functions**. In the next lecture, we will cover **convex sets**. Convex functions have the special property that all local minima are in fact global, which, as we shall see, comes in very handy when solving unconstrained continuous optimization problems via gradient descent. Convex sets are important in constrained optimization problems, the subject of the next lecture.

2 Convex Functions

Formally, for all $\lambda \in [0, 1]$ and all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, a convex function is one that satisfies the following:

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

This definition ensures that the value the function takes on between any two points $\boldsymbol{x}$ and $\boldsymbol{y}$ (the LHS) is no greater than the value of the line segment connecting $\boldsymbol{x}$ and $\boldsymbol{y}$ (the RHS). Intuitively, a function is convex if the line segment connecting any two points on the surface of the function lies above (or on) the function (see Figure 1).

[1] This notes were compiled in conjunction with Professor Amy Greenwald.

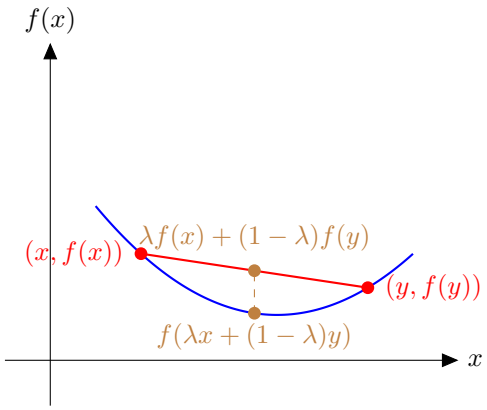

Figure 1: A convex function: the line segment connecting any two points x and y lies above the function.

The function depicted in Figure 2 is not a convex function, because there exists two points (e.g., the local and global minima) such that the line segment that connects them falls below the function. There are many such pairs of points; Figure 2 depicts but one example.

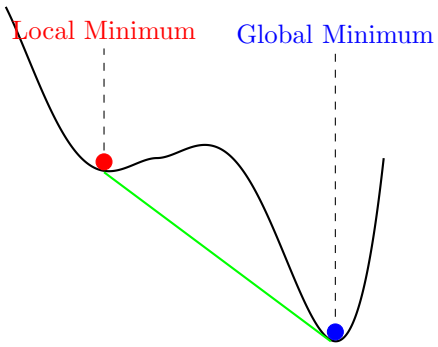

Figure 2: An example of a function with distinct local and global minima. This function is non-convex, because there exist two points (e.g., the local and global minima) such that the line segment connecting them (the green line) falls below the function (the black curve).

So why do we care about convexity? If we can roll a ball down a function starting from anywhere, and it always funnels to the same trough, then there is hope that gradient descent can likewise find its way to that sole trough. Intuitively, we are seeking a property that precludes the possibility of the ball—or gradient descent—getting stuck in a local trough (i.e., at a local minimum) that is not also a global one. (Visualize a ball rolling down a positive quadratic function, or any other function shaped like a bowl e.g., x^4.)

Theorem All local minima of a convex function f are global minima.

Proof Sketch Assume we are given two points, x_L, which is a local minimum, and x_G, which is a global minimum, of some convex function f, and that $f(x_G) < f(x_L)$, i.e., x_L is not a global minimizer of f. Consider the line segment between x_L and x_G given by $\lambda x_L + (1-\lambda)x_G$, for some $\lambda \in [0, 1]$. This line segment must slope downwards from x_L to x_G, since $f(x_G) < f(x_L)$, which implies that the value of the function at all points on that line segment is strictly less than $f(x_L)$. But this is a contradiction, as x_L cannot be a local minimum if there are points near x_L at which the function value is lower than it is at x_L. Therefore, a convex function cannot have any local minima that are not also global minima.

N.B. *The converse of this theorem is not necessarily true.* Even if all local minima are global minima, a

function need not be convex. This is in fact good news, because it means that there exist non-convex functions for which gradient descent (with proper hyperparameter tuning) can find a globally optimal solution.

This theorem explains why convexity is such an important property of functions in continuous optimization. When optimizing a convex function via gradient descent, so long as an appropriate learning rate can be chosen,[2] gradient descent can be guaranteed to converge to a global minimum in polynomial time.

Non-convex optimization problems are very prevalent in practice (e.g., training deep neural networks). While finding a global optimum of a non-convex function via gradient descent is computationally intractable in general, we will see that gradient descent can still be effective, as long as we enhance it with some of the same tools we developed for solving discrete optimization problems via local search (e.g., randomized restarts).

Examples of Convex Functions:

- Affine (sometimes called linear) functions: $f(x) = ax + b$
- Quadratic functions: $f(x) = ax^2 + bx + c$, if $a \geq 0$
- Exponential function: $f(x) = e^{ax}$
- Negative Logarithm: $f(x) = -\log g(x)$, for $x > 0$

Some Operations that Preserve Convexity:

- Non-negative weighted sums: If f_1, f_2 are convex and $w_1, w_2 \geq 0$, then $w_1 f_1 + w_2 f_2$ is convex
- Composition with affine functions: If f is convex, then $f(Ax + b)$ is convex
- Pointwise maximum: If f_1, f_2 are convex, then $f(x) = \max\{f_1(x), f_2(x)\}$ is convex

Convex functions are useful for minimization. For maximization problems, we have a similar definition for *concave* functions. A function $f : \mathbb{R}^n \to \mathbb{R}$ is concave if and only if $-f$ is convex.

[2]Choosing such a rate scientifically depends on another property called Lipschitz continuity that is outside the scope of this lecture. But in practice, choosing these rates (i.e., optimizing hyperparameters) is much more of an art than a science.