Note 8: Convexity

Math 198: Math for Machine Learning

1 Optimization Problems

As we alluded in note 7, our primary use of matrix calculus will be for solving optimization problems. These are problems in which we are given a vector-valued function $f : \mathbb{R}^d \to \mathbb{R}$ and attempt to minimize its value. (Note that in practice if we actually want to maximize f, we can instead minimize $-f$.) There may or may not be constraints on the possible inputs to f under consideration; these scenarios are referred to as constrained and unconstrained optimization, respectively. We define the *feasible set* $\mathcal{X} \subseteq \mathbb{R}^d$ to be the set of possible inputs to f subject to the constraints; $\mathcal{X} = \mathbb{R}^d$ if there are no constraints.

2 Convex Sets and Functions

Convexity is an important property of both sets and functions. Informally, a set $\mathcal{X}$ is *convex* if the line segment between any two points is fully contained within the set. We can formalize this for a set $\mathcal{X} \subseteq \mathbb{R}^d$ by stating that it is convex if

$$t\mathbf{x} + (1-t)\mathbf{y} \in \mathcal{X}$$

for any points $\mathbf{x}, \mathbf{y} \in \mathcal{X}$ and all $t \in [0, 1]$. Convex sets are important because optimization problems become much harder to solve if the feasible set is not convex. Of course, $\mathbb{R}^d$ is convex for all d (as it is closed under addition and scalar multiplication), so for unconstrained optimization the feasible set is always complex.

Convexity for functions is defined similarly. A function $f : \mathbb{R}^d \to \mathbb{R}$ is *convex* if

$$f(t\mathbf{x} + (1-t)\mathbf{y}) \leq tf(\mathbf{x}) + (1-t)f(\mathbf{y})$$

for all $\mathbf{x}, \mathbf{y} \in \text{dom} f$ and all $t \in [0, 1]$. Informally, for any points $\mathbf{x}$ and $\mathbf{y}$ in the domain of f, all values of f in between $\mathbf{x}$ and $\mathbf{y}$ will be less than $f(\mathbf{x})$ and $f(\mathbf{y})$. Even more informally, f is bowl-shaped. If the inequality is strict, then f is *strictly convex.* There is an even stronger notion of convexity for functions, appropriately named m-strong convexity. A function f is m-strongly convex if the function $\mathbf{x} \mapsto f(\mathbf{x}) - \frac{m}{2}||\mathbf{x}||_2^2$ is convex.

Convex functions are significantly easier to minimize than non-convex functions, and thus to optimize. For example, for a convex function f and a convex set $\mathcal{X}$, any local minimum of f in $\mathcal{X}$ is also a global minimum. Let $\mathbf{x}^*$ be such a local minimum. Then for some neighborhood $N \subseteq \mathcal{X}$ about $\mathbf{x}^*$, $f(\mathbf{x}) \geq f(\mathbf{x}^*)$ for all $\mathbf{x} \in N$. Suppose there existed another point $\bar{\mathbf{x}} \in \mathcal{X}$ such that $f(\bar{\mathbf{x}}) < f(\mathbf{x}^*)$. Define $x(t) = t\mathbf{x}^* + (1-t)\bar{\mathbf{x}}$. Then for all $t \in (0, 1)$, $x(t) \in \mathcal{X}$ and

$$\begin{aligned} f(x(t)) &= f(t\mathbf{x}^* + (1-t)\bar{\mathbf{x}}) \\ &\leq tf(\mathbf{x}^*) + (1-t)f(\bar{\mathbf{x}}) \\ &< tf(\mathbf{x}^*) + (1-t)f(\mathbf{x}^*) \\ &= f(\mathbf{x}^*) \end{aligned}$$

We can then pick t^* sufficiently close to 1 that $x(t^*) \in N$, and so $f(x(t^*)) \geq f(\mathbf{x}^*)$; but $f(x(t)) < f(\mathbf{x}^*)$ for all $t \in (0, 1)$ by the above. This contradiction implies that no $\bar{\mathbf{x}}$ could have existed, and so $f(\mathbf{x}^*) < f(\mathbf{x})$ for all $x \in \mathcal{X}$.

In the case of strictly convex functions, we can go further and establish that if a strictly convex function f has a local minimum $\mathbf{x}^* \in \mathcal{X}$, it is the only local minimum. Suppose there were another such global minimum $\bar{\mathbf{x}}$. Then both $\mathbf{x}^*$ and $\bar{\mathbf{x}}$ are global minima by the previous result. Therefore $f(\mathbf{x}^*) = f(\bar{\mathbf{x}})$. Consider again $x(t)$, defined as previously. We now have

$$\begin{aligned} f(x(t)) &= f(t\mathbf{x}^* + (1-t)\bar{\mathbf{x}}) \\ &< tf(\mathbf{x}^*) + (1-t)f(\bar{\mathbf{x}}) \\ &= tf(\mathbf{x}^*) + (1-t)f(\mathbf{x}^*) \\ &= f(\mathbf{x}^*) \end{aligned}$$

for all $t \in (0,1)$. Clearly, this contradicts our premise that $f(\mathbf{x}^*)$ is a global minimum, so there must not be any other local minimum $\bar{\mathbf{x}} \in \mathcal{X}$.

Application: Proofs With Convexity

In this section we will present some useful proofs involving convexity; a few more will be left as homework problems.

We briefly introduced norms in note 2, and since then have primarily considered the p-norms. We now give the definition formally. A norm on a real vector space V is a function $||\cdot|| : V \to \mathbb{R}$ which satisfies the following properties:

(a) $||\mathbf{x}|| \geq 0$, where $||\mathbf{x}|| = 0$ iff $\mathbf{x} = \mathbf{0}$

(b) $||a\mathbf{x}|| = |a|||\mathbf{x}||$

(c) $||\mathbf{x} + \mathbf{y}|| \leq ||\mathbf{x}|| + ||\mathbf{y}||$

Using these properties, we can prove that all norms are convex. Let $||\cdot||$ be a norm on V, $\mathbf{x}, \mathbf{y} \in V$, and $t \in [0, 1]$. Then

$$\begin{aligned} ||t\mathbf{x} + (1-t)\mathbf{y}|| &\leq ||t\mathbf{x}|| + ||(1-t)\mathbf{y}|| \\ &= |t|||\mathbf{x}|| + |(1-t)|||\mathbf{y}|| \\ &= t||\mathbf{x}|| + (1-t)||\mathbf{y}|| \end{aligned}$$

This is a useful result, as we often seek to optimize a norm, as in gradient descent where we optimize the 2-norm.

We next consider convexity proofs for functions f which we already know to be convex. Firstly, if we have some $\alpha \geq 0$, then αf is convex:

$$\begin{aligned} (\alpha f)(t\mathbf{x} + (1-t)\mathbf{y}) &= \alpha f(t\mathbf{x} + (1-t)\mathbf{y}) \\ &\leq \alpha(tf(\mathbf{x}) + (1-t)f(\mathbf{y})) \\ &= t(\alpha f(\mathbf{x})) + (1-t)(\alpha f(\mathbf{y})) \\ &= t(\alpha f)(\mathbf{x}) + (1-t)(\alpha f)(\mathbf{y}) \end{aligned}$$

Additionally, if we have another convex function g, then $f + g$ is convex:

$$\begin{aligned} (f+g)(t\mathbf{x} + (1-t)\mathbf{y}) &= f(t\mathbf{x} + (1-t)\mathbf{y}) + g(t\mathbf{x} + (1-t)\mathbf{y}) \\ &\leq tf(\mathbf{x}) + (1-t)f(\mathbf{y}) + tg(\mathbf{x}) + (1-t)g(\mathbf{y}) \\ &= t(f(\mathbf{x}) + g(\mathbf{x})) + (1-t)(f(\mathbf{y}) + g(\mathbf{y})) \\ &= t(f+g)(\mathbf{x}) + (1-t)(f+g)(\mathbf{y}) \end{aligned}$$

Combining these two proofs, we can show that for n convex functions $f_1, \ldots, f_n$ and constants $\alpha_1, \ldots, \alpha_n \geq 0$, then

$$\sum_{i=1}^{n} \alpha_i f_i$$

is convex as well.

Finally, we show that for convex f and a matrix $\mathbf{A}$ and vector $\mathbf{b}$ of appropriate dimension, $g(\mathbf{x}) = f(\mathbf{A}\mathbf{x} + \mathbf{b})$ is convex:

$$\begin{aligned} g(t\mathbf{x} + (1-t)\mathbf{y}) &= f(\mathbf{A}(t\mathbf{x} + (1-t)\mathbf{y}) + \mathbf{b}) \\ &= f(t\mathbf{A}\mathbf{x} + (1-t)\mathbf{A}\mathbf{y} + \mathbf{b}) \\ &= f(t(\mathbf{A}\mathbf{x} + (1-t)\mathbf{A}\mathbf{y} + t\mathbf{b} + (1-t)\mathbf{b}) \\ &= f(t(\mathbf{A}\mathbf{x} + \mathbf{b}) + (1-t)(\mathbf{A}\mathbf{y} + \mathbf{b})) \\ &\leq tf(\mathbf{A}\mathbf{x} + \mathbf{b}) + (1-t)f(\mathbf{A}\mathbf{y} + \mathbf{b}) \\ &= tg(\mathbf{x}) + (1-t)g(\mathbf{y}) \end{aligned}$$