Math Basics and Properties of Functions

1 Probability Basics

Proposition 1 (Markov Inequality)**.** *Suppose X is a random variable such that $X \geq 0$ with probability 1 and $\mathbb{E}[X] < \infty$. Then for all $y > 0$:*

$$P[X \geq y] \leq \frac{\mathbb{E}[X]}{y}$$

Proof. Notice that $P[X \geq y] = \mathbb{E}[\mathbb{1}[X \geq y]]$, where $\mathbb{1}[X \geq y]$ is 1 when $X \geq y$ and 0 otherwise. Further, since $X \geq 0$, $\mathbb{1}[X \geq y] \leq \frac{X}{y}$. Therefore, $P[X \geq y] = \mathbb{E}[\mathbb{1}[X \geq y]] \leq \mathbb{E}[X/y] = \frac{\mathbb{E}[X]}{y}$.

□

Proposition 2 (Chebychev Inequality)**.** *Suppose X is a random variable with variance $\sigma^2 = \mathbb{E}[(X - \mathbb{E}[X])^2]$. Then for all y:*

$$P[|X - \mathbb{E}[X]| \geq y] \leq \frac{\sigma^2}{y^2}$$

Proof. Define $Z = (X - \mathbb{E}[X])^2$. Observe that $Z \geq 0$ with probability 1, and $\mathbb{E}[Z] = \sigma^2$. Further, $|X - \mathbb{E}[X]| \geq y$ if and only if $Z \geq y^2$. The result now follows from Markov inequality. □

Proposition 3 (Jensen Inequality)**.** *Suppose X is a random variable and f is a convex function. Then*

$$\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$$

Proof. Since f is convex, there exists some g such that for all y,

$$f(y) \geq f(\mathbb{E}[X]) + \langle g, y - \mathbb{E}[X] \rangle$$

(for example, if f is differentiable, $g = \nabla f(\mathbb{E}[X])$). Then

$$\mathbb{E}[f(X)] \geq f(\mathbb{E}[X]) + \mathbb{E}[\langle g, X - \mathbb{E}[X] \rangle] = f(\mathbb{E}[X])$$

□

Jensen inequality is often used in reverse: if f is a *concave* function (that is, $-f$ is convex), then $\mathbb{E}[f(x)] \leq f(\mathbb{E}[X])$. This can be used, for example, to show that for any random variable X, $\mathbb{E}[|X|] \leq \sqrt{\mathbb{E}[X^2]}$.

Also, the following *discrete* statement of Jensen inequality is often useful (note that this is really the same statement, just reformulated a bit):

Corollary 4 (Jensen inequality in discrete case)**.** *Suppose $x_1, \ldots, x_N \in \mathbb{R}^d$ and $a_1, \ldots, a_N$ are non-negative real numbers. Define $\overline{x} = \frac{1}{\sum_{i=1}^N a_i} \sum_{i=1}^N x_i$. Then if f is a convex function:*

$$\frac{1}{\sum_{i=1}^N a_i} \sum_{i=1}^N f(x_i) \geq f(\overline{x})$$

Proof. Consider the random variable X with a discrete distribution that takes on value x_i with probability $\frac{a_i}{\sum_{j=1}^N a_j}$. Notice that $\mathbb{E}[X] = \overline{x}$. The result then follows directly from Jensen inequality. □

Proposition 5 (Union Bound)**.** *Suppose A and B are two events. Then $P[A \text{ or } B] \leq P[A] + P[B]$.*

Proof. Clearly the event represented by A or B can be decomposed into either A or B and not A. The probability of B and not A is at most the probability of B, so the result follows. □

Next, we will often make use of the *bias-variance decomposition*:

Proposition 6 (Bias-variance Decomposition)**.** *Suppose $X \in \mathbb{R}^d$ is some random variable. Then for any deterministic value Y:*

$$\mathbb{E}[\|X - Y\|^2] = \|Y - \mathbb{E}[X]\|^2 + \mathbb{E}[\|X - \mathbb{E}[X]\|^2]$$

Proof. Expanding the square we have:

$$\begin{aligned}
\mathbb{E}[\|X - Y\|^2] &= \mathbb{E}[\|X - \mathbb{E}[X] + \mathbb{E}[X] - Y\|^2] \\
&= \mathbb{E}[\|Y - \mathbb{E}[X]\|^2] + \mathbb{E}[\|X - \mathbb{E}[X]\|^2] + 2\,\mathbb{E}[\langle X - \mathbb{E}[X], \mathbb{E}[X] - Y\rangle] \\
&= \|Y - \mathbb{E}[X]\|^2 + \mathbb{E}[\|X - \mathbb{E}[X]\|^2] + 2\langle \mathbb{E}[X - \mathbb{E}[X]], \mathbb{E}[X] - Y\rangle \\
&= \|Y - \mathbb{E}[X]\|^2 + \mathbb{E}[\|X - \mathbb{E}[X]\|^2]
\end{aligned}$$

where □

Another useful inequality is the Young inequality:

Proposition 7 (Young inequality)**.** *Suppose X and Y are arbitrary vectors. Then for any $\lambda > 0$ and any non-negative real numbers p and q satisfying $\frac{1}{p} + \frac{1}{q} = 1$,*

$$\langle X, Y\rangle \leq \frac{\|X\|^p}{p\lambda^p} + \frac{\lambda^q \|Y\|^q}{q}$$

Proof. Notice that we must have both p and q at least 1 (otherwise $1/p + 1/q > 1$, which is not allowed). Now, differentiating the RHS with respect to λ yields:

$$-\frac{\|X\|^p}{\lambda^{p+1}} + \lambda^{q-1}\|Y\|^q$$

Differentiating again yields:

$$(p+1)\frac{\|X\|^p}{\lambda^{p+2}} + (q-1)\lambda^{q-2}\|Y\|^q \geq 0$$

where we have used $q \geq 1$. Thus, we can find a minimum value for the RHS by setting the first derivative to 0. Solving $-\frac{\|X\|^p}{\lambda^{p+1}} + \lambda^{q-1}\|Y\|^q = 0$ for λ then gives $\lambda = \frac{\|X\|^{p/(p+q)}}{\|Y\|^{q/(p+q)}}$, which yields a minimum value of

$$\frac{\|X\|^{p-\frac{p^2}{p+q}}\|Y\|^{\frac{pq}{p+q}}}{p} + \frac{\|Y\|^{q-\frac{q^2}{p+q}}\|X\|^{\frac{pq}{p+q}}}{q} = \frac{\|X\|^{\frac{pq}{p+q}}\|Y\|^{\frac{pq}{p+q}}}{p} + \frac{\|Y\|^{\frac{pq}{p+q}}\|X\|^{\frac{pq}{p+q}}}{q}$$

Now, notice that

$$1 = \frac{1}{p} + \frac{1}{q} = \frac{p+q}{pq}$$

So that the minimum of the RHS is in fact:

$$\frac{\|X\|\|Y\|}{p} + \frac{\|Y\|\|X\|}{q} = \|X\|\|Y\|$$

Now to conclude observe that $\langle X, Y\rangle \leq \|X\|\|Y\|$ by Cauchy-Schwarz. □

Finally, we will also often need the following generalization of Cauchy-Schwarz:

Proposition 8 (Cauchy-Schwarz for random variables)**.** *Suppose $A \in \mathbb{R}$ and $B \in \mathbb{R}$ are arbitrary random variables. Then:*

$$\mathbb{E}[AB] \leq \sqrt{\mathbb{E}[A^2]\,\mathbb{E}[B^2]}$$

Proof. By Young inequality (Proposition 7), we have

$$\begin{aligned}\mathbb{E}[AB] &\leq \mathbb{E}\left[\frac{A^2}{2\lambda^2} + \frac{\lambda^2 B^2}{2}\right] \\ &= \frac{\mathbb{E}[A^2]}{2\lambda^2} + \frac{\lambda^2\,\mathbb{E}[B^2]}{2}\end{aligned}$$

for any constant $\lambda > 0$. Now, set $\lambda = \frac{\mathbb{E}[A^2]}{\mathbb{E}[B^2]}$ to conclude the desired result. □

2 Convexity, Smoothness and Lipschitzness

Many classic algorithms work for *convex* functions:

Definition 9. *A function $\mathcal{L}$ is* convex *if for all x, there exists some g_x such that for all y:*

$$\mathcal{L}(y) \geq \mathcal{L}(x) + \langle g_x, y - x\rangle$$

When $\mathcal{L}$ is differentiable, $g_x = \nabla\mathcal{L}(x)$.

When $\mathcal{L}$ is twice-differentiable, the following is an equivalent condition for convexity:

Proposition 10. *Suppose $\mathcal{L}$ is twice-differentiable. Then $\mathcal{L}$ is convex if and only if $\nabla^2\mathcal{L}(\mathbf{w}) \succeq 0$ for all $\mathbf{w}$.*

Next, we will usually assume that loss functions $\mathcal{L}$ iare *Lipschitz*:

Definition 11. *A function $\mathcal{L} : \mathbb{R}^d \to \mathbb{R}$ is G-Lipschitz if for all $x, y \in \mathbb{R}^d$,*

$$|\mathcal{L}(x) - \mathcal{L}(y)| \leq G\|x - y\|$$

This definition may be a little strange looking, but it's actually fairly likely to hold. The intuition here is that the G in G-Lipschitzness is actually measuring the degree to which $\mathcal{L}$ is continuous. In fact, if W is a compact set, then any continuous $\mathcal{L}$ must be G-Lipschitz for some G. When $\mathcal{L}$ is differentiable, we can phrase G-Lipschitzness in the following way:

Proposition 12. *If $\mathcal{L}$ is differentiable, then $\mathcal{L}$ is G-Lipschitz if and only if $\|\nabla\mathcal{L}(x)\| \leq G$ for all x.*

Proof. First, suppose $\|\nabla\mathcal{L}(x)\| \geq G$ for some $x \in \mathbb{R}^d$. By definition of gradient for any vector v, we have:

$$\lim_{\delta\to 0} \frac{\mathcal{L}(x + \delta v) - \mathcal{L}(x) - \delta\langle v, \nabla\mathcal{L}(x)\rangle}{\delta} = 0$$

Let $v = \frac{\nabla\mathcal{L}(x)}{\|\nabla\mathcal{L}(x)\|}$. Then this implies:

$$\lim_{\delta\to 0} \frac{\mathcal{L}(x + \delta v) - \mathcal{L}(x)}{\delta} = \|\nabla\mathcal{L}(x)\|$$

Thus, if $\|\nabla\mathcal{L}(x)\| \geq G$, there must be some δ such that

$$|\mathcal{L}(x + \delta v) - \mathcal{L}(x)| \geq G\delta$$

and so $\mathcal{L}$ is no G-Lipschitz. Therefore G-Lipschitzness implies $\|\nabla\mathcal{L}(x)\| \leq G$. Now suppose $\|\nabla\mathcal{L}(x)\| \leq G$ for all x. Then by the fundamental theorem of calculus,

$$\begin{aligned}\mathcal{L}(x) - \mathcal{L}(y) &= \int_0^1 \frac{d}{dt}\mathcal{L}(y + t(x-y))\, dt \\ &= \int_0^1 \langle \nabla\mathcal{L}(y + t(x-y)), x - y\rangle\, dt \\ &\leq \int_0^1 \|\nabla\mathcal{L}(y + t(x-y))\|\|x-y\|\, dt \\ &\leq G\|x-y\|\end{aligned}$$

so that $\mathcal{L}$ is G-Lipschitz. □

For most of this course, we will consider exclusively losses $\mathcal{L}$ that are *smooth*:

Definition 13. *A function $\mathcal{L} : \mathbb{R}^d \to \mathbb{R}$ is H-smooth if $\mathcal{L}$ is differentiable at all $x \in \mathbb{R}^d$, and for all $x, y \in W$,*

$$\|\nabla\mathcal{L}(x) - \nabla\mathcal{L}(y)\| \leq H\|x - y\|$$

Let's be clear: there are plenty of losses out there that are not smooth, or are only smooth with a very large value of H. However, without *some* kind of assumption about the loss it is very difficult to prove anything about how an algorithm will operate. Similarly to the way that G-Lipschitzness is the same as the gradient being bounded by G when $\mathcal{L}$ is differentiable, H-smoothness is the same as the Hessian being bounded by H when $\mathcal{L}$ is twice-differentiable:

Proposition 14. *Suppose $\mathcal{L}$ is twice differentiable. Then $\mathcal{L}$ is H-smooth if and only if $\|\nabla^2\mathcal{L}(\mathbf{w})\|_{op} \leq H$ for all $\mathbf{w}$.*

The proof is essentially the same as the proof of Proposition 12, although with two integrals instead of one. We will leave it as an exercise.

Smoothness is at least approximately satisfied by essentially any continuous function, in the sense that any contuous function can be replaced with a approximation that is smooth. The quality of the approximation can be traded off for the amount of smoothness. Formally, the following theorem holds:

Theorem 15. *Let $\mathcal{L} : \mathbb{R}^d \to \mathbb{R}$ be a G-Lipschitz function. Then for any H, consider*

$$\hat{\mathcal{L}}(x) = \inf_y \mathcal{L}(y) + \frac{H}{2}\|x - y\|^2$$

Then $\hat{\mathcal{L}}$ is H-smooth and G-Lipschitz, and for all $x \in W$,

$$|\hat{\mathcal{L}}(x) - \mathcal{L}(x)| \leq \frac{G^2}{H}$$

To extend the idea of smoothness, we have the notion of *second-order smoothness*:

Definition 16. *A function $\mathcal{L} : \mathbb{R}^d \to \mathbb{R}$ is ρ-second-order smooth if $\mathcal{L}$ is twice-differentiable at all $x \in \mathbb{R}^d$, and for all x, y, v:*

$$\|(\nabla^2\mathcal{L}(x) - \nabla^2\mathcal{L}(y))v\| \leq \rho\|x - y\|\|v\|$$

Just as Lipschitzness relates to the first derivative and smoothness to the second derivative, second-order smoothness means that the *third* derivative is bounded. The third derivative of a function $f : \mathbb{R}^d \to \mathbb{R}$ is a bit of a complicated object. Formally, it is a tensor in $\mathbb{R}^{d\times d\times d}$. Concretely, it can be represented by a three-dimensional matrix whose ijk entry is $\frac{\partial f}{\partial x_i \partial x_j \partial x_k}$. This 3-dimensional matrix turns vectors into regular 2-dimensional matrices: given a 3-dimensional matrix T and a vector v, we write the 2-dimensional matrix Tv as:

$$Tv[i, j] = \sum_k T[i, j, k]v[k]$$

We say $\|T\|_{\text{op}} = \sup_{\|v\|\leq 1} \|Tv\|_{\text{op}}$. With this notation, a thrice-differentiable function is ρ-second-order smooth if and only if its third derivative has operator norm at most ρ.

2.1 Critical points and Local minima

Because we will not be assuming that $\mathcal{L}$ is convex, in general we will not be able to actually show that our algorithms in fact minimizes $\mathcal{L}$. Instead, we will often simply try to approach a *critical point*:

Definition 17. *A* critical point*, or* first-order stationary point *of $\mathcal{L}$ is a point x such that $\nabla\mathcal{L}(x) = 0$.*

The search for critical points is motivated by the observation that any minimizer of $\mathcal{L}$ must also be a critical point, so that finding a critical point is a necessary but not sufficient condition for actually minimizing $\mathcal{L}$. There is some active research investigating the degree to which this condition may actually be sufficient. In fact, one of the key desirable properties of convex functions is that any critical point must also minimize $\mathcal{L}$. Sometimes we will be more ambitious and instead try to find a *local minimum*:

Definition 18. *A* local minimum*, or a* second-order stationary point *of $\mathcal{L}$ is a point x such that for some ϵ and all y with $\|y - x\| \leq \epsilon$, $\mathcal{L}(y) \geq \mathcal{L}(x)$.*

As is typical in optimization, instead of actually identifying critical points or local minima, we will identify approximate critical points or local minima. The notion of an approximate critical point is relatively straightforward: we simply try to make $\|\nabla\mathcal{L}(x)\|$ as small as possible:

Definition 19. *A ϵ-approximate critical point of $\mathcal{L}$ is a point x such that $\|\nabla\mathcal{L}(x)\| \leq \epsilon$*

The appropriate way to define an approximate local minimum is less clear however. To think about this, we will need to use some information about the second derivative. One particular idea that is important is the notion of being *positive semidefinite*:

Definition 20. *A square matrix $M \in \mathbb{R}^{d\times d}$ is positive semidefinite if for all $v \in \mathbb{R}^d$, $v^\top Mv \geq 0$. Further, for any matrices A and B, we use the notation $A \succeq B$ to indicate that $v^\top Av \geq v^\top Bv$ for all v. Then, M is positive semidefinite if and only if $M \succeq 0$.*

We will adopt the following standard definition:

Definition 21. *Suppose $\mathcal{L}$ is twice-differentiable. Then x is an (ϵ, δ)-approximate local minimum if $\|\nabla\mathcal{L}(x)\| \leq \epsilon$ and $\nabla^2\mathcal{L}(x) \succeq -\delta I$. Frequently, we will consider the case $\delta = \sqrt{\epsilon}$.*

Let's get some intuition behind why we impose the condition on $\nabla^2\mathcal{L}(x)$. First, observe that if $\mathcal{L}$ is twice differentiable, then at any local minimum we must have $\nabla^2\mathcal{L}(x) \succeq 0$. To see this intuitively, observe that by a second-order Taylor expansion, for all v very close to 0 we have:

$$\mathcal{L}(x + v) \approx \mathcal{L}(x) + \langle\nabla\mathcal{L}(x), v\rangle + \frac{v^\top\nabla^2\mathcal{L}(x)v}{2}$$

We have that $\nabla\mathcal{L}(x) = 0$ at a local minimum, so:

$$= \mathcal{L}(x) + \frac{v^\top\nabla^2\mathcal{L}(x)v}{2}$$

Now, since we are at a local minimum, $\mathcal{L}(y) \geq \mathcal{L}(x)$ so that we must have $v^\top\nabla^2\mathcal{L}(x)v \geq 0$ for all v in some neighborhood of 0, which implies that $\nabla^2\mathcal{L}(x) \succeq 0$. Therefore, the δ part of Definition 21 is measuring the degree to which this inequality is violated.