Optimization for Machine Learning HW 1

Solutions

1. This question provides practice thinking about random variables. Please provide a proof for all answers unless explicitly instructed otherwise.

   (a) Is there a distribution over positive integers that has finite mean but infinite variance? If so, explicitly describe such a distribution and prove that it has the desired properties. If not, prove that no such distribution exists.

   Solution: We will produce such a distribution. First, since $1/n^3$ is a decreasing function of n, we have:

   $$\lim_{N\to\infty} \sum_{n=1}^{N} \frac{1}{n^3} \leq \lim_{N\to\infty} \int_1^{N+1} \frac{dn}{n^3} < \infty$$

   so that $\sum_{n=1}^{\infty} \frac{1}{n^3} = C$ for some constant C (essentially, we use the "integral test" to establish the convergence of the sum). Similarly, we have that $\sum_{n=1}^{\infty} \frac{1}{n^2}$ converges because the integral $\int_1^{\infty} \frac{dn}{n^2}$ converges. Let $\sum_{n=1}^{\infty} \frac{1}{n^2} = K$ (in fact, $K = \frac{\pi^2}{6}$, but its exact value is unimportant in this problem).
   Now, let $p(n) = \frac{1}{Cn^3}$. We have $\sum_{n=1}^{\infty} p(n) = 1$ by definition. Now, the expected value of the distribution specified by probability mass function p is :

   $$\mathbb{E}[n] = \sum_{n=1}^{\infty} \frac{n}{Cn^3} = \sum_{n=1}^{\infty} \frac{1}{Cn^2} = \frac{K}{C} < \infty$$

   However, note that the second moment is:

   $$\mathbb{E}[n^2] = \sum_{n=1}^{\infty} \frac{n^2}{Cn^3} = \sum_{n=1}^{\infty} \frac{1}{Cn} = \infty$$

   since the sum $\sum_{n=1}^{\infty} \frac{1}{n}$ diverges.

   (b) If X is a random variable satisfying $\mathbb{E}[\|X - \mathbb{E}[X]\|^n] \leq \sigma^n$ for some $n > 0$, show that for any $\delta > 0$, with probability at least $1 - \delta$, $\|X - \mathbb{E}[X]\| \leq \frac{\sigma}{\delta^{1/n}}$.

   Solution:
   Notice that $P[\|X - \mathbb{E}[X]\| \geq y] = \mathbb{E}[\|X - \mathbb{E}[X]\|^n \geq y^n]$ for any $y > 0$. Also, $\|X - \mathbb{E}[X]\| \geq 0$ with probability 1. Therefore by Markov inequality, $P[\|X - \mathbb{E}[X]\| \geq y] = \mathbb{E}[\|X - \mathbb{E}[X]\|^n \geq y^n] \leq \frac{\mathbb{E}[\|X - \mathbb{E}[X]\|^n]}{y^n} \leq \frac{\sigma^n}{y^n}$. This implies that for any $y > 0$, with probability at most $1 - \frac{\sigma^n}{y^n}$, we have $\|X - \mathbb{E}[X]\| \leq y$. Set $\frac{\sigma^n}{y^n} = \delta$ and solve for y to obtain $y = \frac{\sigma}{\delta^{1/n}}$.

   (c) Suppose that X is a random variable such that for all real numbers (not just integers!) $n > 0$, $\mathbb{E}[\|X - \mathbb{E}[X]\|^n]^{1/n} \leq \sigma\sqrt{n}$. Show that with probability at least $1 - \delta$, $\|X - \mathbb{E}[X]\| \leq \sigma\sqrt{2\exp(1)\log(1/\delta)}$. Distributions satisfying this property are called *subgaussian*. The Normal distribution is an example of a distribution satisfying this kind of property.

Solution:
By the previous part, we have that for any n, with probability at least $1-\delta$, $\|X - \mathbb{E}[X]\| \leq \frac{\sigma\sqrt{n}}{\delta^{1/n}}$. Since this holds for all n, we are free to pick the "best" n that yields the tightest bound. To find this n, we differentiate $\frac{\sigma\sqrt{n}}{\delta^{1/n}}$:

$$\frac{d}{dn}\frac{\sigma\sqrt{n}}{\delta^{1/n}} = \frac{\sigma\delta^{-1/n}(2\log(\delta)+n)}{n^{3/2}}$$

Setting this equal to zero, and using the fact that $\frac{\sigma\delta^{-1/n}}{n^{3/2}} \neq 0$, we obtain that the optimal n satisfies $0 = 2\log(\delta) + n$, so $n = -2\log(\delta)$. Now, plug this back into $\frac{\sigma\sqrt{n}}{\delta^{1/n}}$ to obtain:

$$\begin{aligned}\frac{\sigma\sqrt{n}}{\delta^{1/n}} &= \frac{\sigma\sqrt{2\log(1/\delta)}}{\delta^{-1/(2\log(\delta))}} \\ &= \frac{\sigma\sqrt{2\log(1/\delta)}}{\exp\left(-\frac{\log(\delta)}{2\log(\delta)}\right)} \\ &= \sigma\sqrt{2\exp(1)\log(1/\delta)}\end{aligned}$$

2. This question provides practice in some linear algebra ideas.

   (a) For any matrix M, the *operator norm* of M is $\|M\|_{\text{op}} = \sup_{\|v\|=1}\|Mv\|$. Prove that for all matrices A and B of the same dimensions, $\|A+B\|_{\text{op}} \leq \|A\|_{\text{op}} + \|B\|_{\text{op}}$.

   Solution: Let x be any unit vector. Then

   $$\|(A+B)x\| = \|Ax+Bx\| \leq \|Ax\| + \|Bx\| \leq \|A\|_{\text{op}} + \|B\|_{\text{op}}$$

   The first inequality is triangle inequality, and the second is by definition of operator norm. Since this holds for all unit vectors x, it must hold for the supremum over x as well and so we are done.

   (b) Most of the matrices we will discuss in this class are *symmetric* matrices. The *real spectral theorem* states that any symmetric matrix $M \in \mathbb{R}^{d\times d}$ has an orthonormal basis of eigenvectors. That is, there exists $v_1, \ldots, v_d$ such that each v_i has norm 1, $\langle v_i, v_j\rangle = 0$ for all $i \neq j$, and $Mv_i = \lambda_i v_i$ for some real numbers $\lambda_1, \ldots, \lambda_d$. Prove that $\|M\|_{\text{op}} = \max_i |\lambda_i|$.

   Solution:
   Without loss of generality, assume $|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_d|$, so that $\max_i |\lambda_i| = |\lambda_1|$. Now, observe that $\|Mv_1\| = \|\lambda_1 v_1\| = |\lambda_1|$, so that $\|M\|_{\text{op}} \geq |\lambda_1|$.
   Next, let x be an arbitrary vector with $\|x\| = 1$. Since $v_1, \ldots, v_d$ forms an orthonormal basis, we can write $x = \sum_{i=1}^d c_i v_i$ for some real numbers c_i. Further, $\|x\|^2 = \sum_{i=1}^d c_i^2$. Also

   $$\|Mx\|^2 = \|\sum_{i=1}^d \lambda_i c_i v_i\|^2 = \sum_{i=1}^d \lambda_i^2 c_i$$

   Define $z_i = c_i^2$. Notice that $\sum_{i=1}^d z_i = \|x\|^2 = 1$. Thus:

   $$\|Mx\|^2 = \sum_{i=1}^d \lambda_i^2 c_i^2$$

   Since $z_i \geq 0$ and $\lambda_i^2 \leq \lambda_1^2$:

   $$\begin{aligned}&\leq \lambda_1^2 \sum_{i=1}^d z_i \\ &= \lambda_1^2\end{aligned}$$

Since this holds for arbitrary x, we have Thus $\|M\|_{\text{op}} = \sqrt{\sup_{\|x\|=1} \|Mx\|^2} \leq \sqrt{\lambda_1^2} = |\lambda_1|$ and so we are done.

(c) Prove a generalized Young's inequality: for any vectors a and b and any positive scalar λ, $\langle a, b\rangle \leq \frac{\|a\|^2}{2\lambda} + \frac{\lambda\|b\|^2}{2}$ (hint: take a derivative).

Solution:
First, by Cauchy-Schwarz, $\langle a, b\rangle \leq \|a\|\|b\|$, so it suffices to prove $\|a\|\|b\| \leq \frac{\|a\|^2}{2\lambda} + \frac{\lambda\|b\|^2}{2}$ for all λ. This again is equivalent to $\|a\|\|b\| \leq \inf_\lambda \frac{\|a\|^2}{2\lambda} + \frac{\lambda\|b\|^2}{2}$. To compute the infimum, we differentiate with respect to λ, yielding $-\frac{\|a\|^2}{2\lambda^2} + \frac{\|b\|^2}{2}$, which suggests that the optimal λ is $\frac{\|a\|}{\|b\|}$, so that $\inf_\lambda \frac{\|a\|^2}{2\lambda} + \frac{\lambda\|b\|^2}{2} = \|a\|\|b\|$ as desired.
There are a couple minor caveats: first, to check that this is indeed the minimum, we can compute the second derivative, which is $\frac{\|a\|^2}{\lambda^3} \geq 0$, so that it is indeed a minimum rather than a maximum in λ. Second, if $\|b\| = 0$, we can see that the result reduces to showing that $0 \leq \frac{\|a\|^2}{2\lambda}$, which holds since $\lambda > 0$ and $\|a\| \geq 0$.

Alternate Solution
Several students submitted a clever solution along the following lines:
Consider the quantity $\left\| \frac{a}{\sqrt{2\lambda}} - \frac{b\sqrt{\lambda}}{\sqrt{2}} \right\|^2$. Since the euclidean norm is non-negative, we must have:

$$0 \leq \left\| \frac{a}{\sqrt{\lambda 2}} - \frac{b\sqrt{\lambda}}{\sqrt{2}} \right\|^2$$

now expand the RHS:

$$= \frac{\|a\|^2}{2\lambda} - \langle a, b\rangle + \frac{\lambda\|b\|^2}{2}$$

rearrange terms:

$$\langle a, b\rangle \leq \frac{\lambda\|a\|^2}{2} + \frac{\|b\|^2}{2\lambda}$$

3. Suppose you are working for a online store and you need to predict the probability that a person who visits your homepage will buy something. You have a dataset $z_1, \ldots, z_N$ where z_i is 1 if the ith visitor to the homepage bought something, and 0 otherwise. You assume that each z_i is an independent and indentically distributred random variable, so your task is just to learn $p = P[z = 1] = \mathbb{E}[z]$. You decide to use the simple estimate $\hat{p} = \frac{1}{N}\sum_{t=1}^N z_i$. Show that for any N and any p, $\mathbb{E}[|\hat{p} - p|] \leq \frac{\sqrt{p(1-p)}}{\sqrt{N}}$.

Solution:
Since $x \mapsto -\sqrt{x}$ is convex, Jensen inequality tells us that for any random variable, $\mathbb{E}[\sqrt{X}] \leq \sqrt{\mathbb{E}[X]}$. Therefore if we set $X = (\hat{p} - p)^2$, we have $\mathbb{E}[|\hat{p} - p|] \leq \sqrt{\mathbb{E}[(\hat{p} - p)^2]}$. Let's compute $\mathbb{E}[(\hat{p} - p)^2]$:

$$\begin{aligned}
\mathbb{E}[(\hat{p} - p)^2] &= \frac{1}{N^2}\,\mathbb{E}\left[\left(\sum_{i=1}^N z_i - p\right)^2\right] \\
&= \frac{1}{N^2}\,\mathbb{E}\left[\sum_{i=1}^N (z_i^2 - p)^2 + \sum_{i\neq j}(z_i - p)(z_j - p)\right]
\end{aligned}$$

Now, notice that since $\mathbb{E}[z_i] = p$ and z_i and z_j are independent, $\mathbb{E}[(z_i - p)(z_j - p)] = 0$, and $\mathbb{E}[(z_i^2 - p)] = p(p-1)$, so that $\mathbb{E}[(\hat{p} - p)^2] = \frac{p(1-p)}{N}$. Taking the square root now completes the proof.

BONUS This question is an exercise in expectations and sums and logs. Many of these ideas and techniques are also useful in analysis of algorithms. We will prove upper and lower bounds on the expected displacement of a random walk. Let $X_1, \ldots, X_N$ be independent random variables with $X_i = \pm 1$ with equal probability. Let $S_n = \sum_{i=1}^{n} X_n$. Show that there exists a constant C such that for all N, $C\sqrt{N} \leq \mathbb{E}[|S_N|] \leq \sqrt{N}$.

(a) First, we will show that $\mathbb{E}[S_N^2] = N$, and that $\mathbb{E}[|S_N|] \leq \sqrt{N}$.
We have $S_N^2 = \sum_{i=1}^{N} X_i^2 + \sum_{i \neq j} X_i X_j = N + + \sum_{i \neq j} X_i X_j$. Since $\mathbb{E}[X_i] = 0$ and X_i is independent of X_j, $\mathbb{E}[X_i X_j] = 0$ for $i \neq j$, so $\mathbb{E}[S_N^2] = N$. Now by Jensen inequality, $\mathbb{E}[|S_N|] \leq \sqrt{\mathbb{E}[S_N^2]} = \sqrt{N}$.
The rest of this question is designed to guide you through showing $\mathbb{E}[|S_N|] \geq C\sqrt{N}$ for some constant C.

(b) Next, we let $Z_n = P[S_n = 0]$, and define $Z_0 = 1$. We will show that for $N \geq 2$, $\mathbb{E}[|S_N|] = \sum_{i=0}^{N-1} Z_i$
We proceed by induction on n. The statement is clearly true for $n = 1$, so suppose it is true for some n. By definition of expectation, we have:

$$\begin{aligned}
\mathbb{E}[|S_n|] &= \sum_{x=-\infty}^{\infty} |x+1| P[S_{n-1} = x, X_n = 1] + |x-1| P[S_{n-1} = x, X_n = -1] \\
&= \sum_{x=-\infty}^{\infty} \frac{|x+1| + |x-1|}{2} P[S_{n-1} = x] \\
&= P[S_{n-1} = 0] + \sum_{x \neq 0} |x| P[S_{n-1} = x] \\
&= \mathbb{E}[S_{n-1}] + Z_{n-1} \\
&= \sum_{i=0}^{n-1} Z_i
\end{aligned}$$

(c) Now, we show for any odd $n \geq 1$, $Z_n = 0$, and for any even n, $Z_n = \frac{n!}{(n/2)!(n/2)!2^n}$.
In order for S_n to be zero, the number of X_i for $i \leq n$ with $X_i = 1$ must be the same as the number with $X_i = 0$, so the total n must be twice the number of X_i with $X_i = 1$, which is impossible if n is odd, so $Z_n = 0$ for odd n.
For even n, the total number of possible ways to assign values to the individual X_i is 2^n. S_n will be zero if exactly $n/2$ of these are equal to 1, so there are $\binom{n}{n/2} = \frac{n!}{(n/2)!(n/2)!}$ ways to assign a positive value to $n/2$ of the indices.

(d) This expression will allow us to show that for any m, $\log(Z_{2m}) = \sum_{i=1}^{m} \log\left(1 - \frac{1}{2i}\right)$.
Using the expression for Z_n, we have

$$\begin{aligned}
Z_{2i} &= \frac{(2i)(2i-1)\ldots(i+1)}{i!2^{2i}} \\
Z_{2i-1} &= \frac{(2i-2)(2i-3)\ldots i}{(i-1)!2^{2i-2}} \\
\frac{Z_{2i-1}}{Z_{2i}} &= \frac{2i(2i-1)}{i^2 2^2} = 1 - \frac{1}{2i}
\end{aligned}$$

Thus, by unrolling this recursion for m rounds we have $Z_{2m} = \prod_{i=1}^{m} \log\left(1 - \frac{1}{2i}\right)$. Take the logarithm of both sides to conclude:

$$\log(Z_{2m}) = \sum_{i=1}^{m} \log\left(1 - \frac{1}{2i}\right)$$

(e) Next we need a few technical identities. First, for all $x \le 1/2$, $\log(1-x) \ge -x - x^2$.
Let $f(x) = \log(1-x)$. First, if $x < 0$, then $f(x) \ge 0 > -x - x^2$. So let us consider only $x \in [0, 1/2]$. By the Taylor expansion for $f(x)$ around 0 (which is valid in $|x| < 1$):

$$\begin{aligned} f(x) &= -x - \frac{x^2}{2} - \frac{x^3}{3} - \frac{x^4}{4} - \dots \\ &\ge -x - \frac{x^2}{2} - \frac{x^2}{3}\sum_{i=1}^{\infty} x^i \\ &\ge -x - \frac{x^2}{2} - \frac{x^2}{3}\sum_{i=1}^{\infty} \frac{1}{2^i} \\ &\ge -x - x^2 \end{aligned}$$

(f) As a second technical identity, we show that $\sum_{i=2}^{m} \frac{1}{i} \le \int_1^m \frac{dx}{x}$.
Since $\frac{1}{x}$ is a decreasing function, we have that $\int_{a-1}^{a} \frac{dx}{x} \ge \int_{a-1}^{a} \frac{dx}{a} = \frac{1}{a}$. Therefore $\sum_{i=2}^{m} \frac{1}{i} \le \sum_{i=2}^{m} \int_{i-1}^{i} \frac{dx}{x} = \int_1^m \frac{dx}{x} = \log(m)$. Then we have $\sum_{i=1}^{m} \frac{1}{i} = 1 + \sum_{i=2}^{m} \frac{1}{i} \le 1 + \log(m)$.

(g) Show that $\sum_{n=1}^{N-1} \frac{1}{\sqrt{n}} \le \int_0^{N-1} \frac{dx}{\sqrt{x}} = 2\sqrt{N-1}$.

Solution:
There is a typo in this question, it should say $\sum_{n=1}^{N-1} \frac{1}{\sqrt{n}} \le 1 + \int_1^{N-1} \frac{dx}{\sqrt{x}} = 2\sqrt{N-1} - 1$.
Now, by the same argument as in the previous part, since $\frac{1}{\sqrt{x}}$ is decreasing, the sum is upper-bounded by the integral and so $\sum_{n=2}^{N-1} \frac{1}{\sqrt{n}} \le \int_1^{N-1} \frac{dx}{\sqrt{x}} = 2\sqrt{N-1} - 2$. Then observe that $\sum_{n=1}^{N-1} \frac{1}{\sqrt{n}} = 1 + \sum_{n=2}^{N-1} \frac{1}{\sqrt{n}}$ to conclude the result.

(h) Show that for all even $n \ge 1$, $Z_n \ge \frac{\exp(-1/2 - \pi^2/24)\sqrt{2}}{\sqrt{n}}$. You may use without proof the identity $\sum_{i=1}^{\infty} \frac{1}{i^2} = \frac{\pi^2}{6}$.

Solution:
Let $n = 2m$. By part (d), we have $\log(Z_{2m}) = \sum_{i=1}^{m} \log\left(1 - \frac{1}{2i}\right)$:

$$\log(Z_n) = \log(Z_{2m}) = \sum_{i=1}^{m} \log\left(1 - \frac{1}{2i}\right)$$

By part (e)

$$\begin{aligned} &\ge \sum_{i=1}^{m} \left[-\frac{1}{2i} - \frac{1}{4i^2}\right] \\ &\ge -\sum_{i=1}^{m} \frac{1}{2i} - \sum_{i=1}^{\infty} \frac{1}{4i^2} \\ &\ge -\frac{1}{2}\sum_{i=1}^{m} \frac{1}{i} - \frac{\pi^2}{24} \end{aligned}$$

By part (f):

$$\ge -\frac{1}{2} - \frac{\log(m)}{2} - \frac{\pi^2}{24}$$

exponentiate both sides:

$$\begin{aligned} Z_n &\geq \exp\left[-\frac{1}{2} - \frac{\pi^2}{24}\right] / \exp\left[\frac{\log(m)}{2}\right] \\ &= \frac{\exp\left[-\frac{1}{2} - \frac{\pi^2}{24}\right]}{\sqrt{m}} \end{aligned}$$

plug in the definition $m = n/2$:

$$= \frac{\exp\left[-\frac{1}{2} - \frac{\pi^2}{24}\right]\sqrt{2}}{\sqrt{n}}$$

(i) Conclude that there exists some absolute constant C such that $\mathbb{E}[|S_N|] \geq C\sqrt{N}$.

Solution:
By part (b), we have

$$\mathbb{E}[|S_N|] = \sum_{i=0}^{N-1} Z_i$$

Apply part (h):

$$\begin{aligned} &\geq \sum_{i=0}^{N-1} \frac{\exp\left[-\frac{1}{2} - \frac{\pi^2}{24}\right]\sqrt{2}}{\sqrt{n}} \\ &\geq \exp\left[-\frac{1}{2} - \frac{\pi^2}{24}\right]\sqrt{2} \sum_{i=0}^{N-1} \frac{1}{\sqrt{n}} \end{aligned}$$

Apply part (g):

$$\begin{aligned} &\geq \exp\left[-\frac{1}{2} - \frac{\pi^2}{24}\right]\sqrt{2}(2\sqrt{N-1} - 1) \\ &\geq 2\exp\left[-\frac{1}{2} - \frac{\pi^2}{24}\right]\sqrt{2N} \end{aligned}$$

So the result holds with $C = 2\exp\left[-\frac{1}{2} - \frac{\pi^2}{24}\right]\sqrt{2}$.