Homework 4

Due: 1 November, 2024

Each problem is graded on the coarse scale of $\checkmark^+$, $\checkmark$, $\checkmark^-$ and no $\checkmark$. It is also assigned a multiplier, denoting the relative importance of the problem. Both correctness and presentation are grading criteria.

Please read and make sure you understand the collaboration policy on the course missive. Extra credit problems are clearly marked below (see course missive for the details of grade calculations).

Remember to prove all your (non-elementary and not shown in class) mathematical claims, unless stated otherwise.

Each group of students should submit only 1 pdf to the corresponding Canvas assignment.

Problem 1

(2 $\checkmark$s)

Here is an argument involving the Johnson-Lindenstrauss lemma, but part of it is wrong. Identify the part and explain why it is wrong, as motivation for why we had to use Kirszbraun's extension theorem.

> Consider a set of n points $\{q_1, \ldots, q_n\}$, and we wish to estimate the quality of the best 1-median solution: find a point p^* that minimises the cost, that is the sum of distances between p and each of the n points. We project the points down to $O(\frac{\log n}{\epsilon^2})$ dimensions using a random project Π, with the big-O chosen to preserve norms in $O(n^2)$ directions, and find the optimal solution $\hat{p}$ in the low dimensionality. The goal, as we saw in class, is to show (incorrectly in this example) that the cost of p^* and $\hat{p}$ are within a $1 \pm O(\epsilon)$ factor of each other.
>
> First, we claim that the cost of $\hat{p}$ in the low dimensionality at most $1+O(\epsilon)$ times the cost of p^* (in high dimensionality). The observation is that, by the Johnson-Lindenstrauss lemma, p^* will have its distances with q_i approximately preserved (with high probability) over the projection Π. Thus Πp^* is a good quality solution in the low dimensionality, having cost that is no more than $1 + O(\epsilon)$ times the cost of p^* in the original dimensionality. As the cost of $\hat{p}$ is no more than that of Πp^*, it is also upper bounded by the cost of p^* up to the same $1 + O(\epsilon)$ factor.
>
> Second, we claim that the cost of $\hat{p}$ (in low dimensionality) cannot be more than a $1 - O(\epsilon)$ factor smaller than the cost of p^*. Observe that $\hat{p}$ is within the convex hull of $\{\Pi q_1, \ldots, \Pi q_n\}$, and so there must be some point $\tilde{p}$ in the convex hull of $\{q_1, \ldots, q_n\}$ such that $\Pi\tilde{p} = \hat{p}$. By the Johnson-Lindenstrauss lemma, the distances between $\tilde{p}$ and the points $q_1, \ldots, q_n$ is preserved (since the set $\{q_1, \ldots, q_n\} \cup \{p^*, \tilde{p}\}$ has size $O(n)$), and so the cost of $\hat{p}$ is within a factor of $1 \pm O(\epsilon)$ of the cost of $\tilde{p}$. This means that the cost of $\hat{p}$ is at least a $1 - O(\epsilon)$ factor of the cost of $\tilde{p}$, which in turn is at least the cost of p^* by the optimality of p^*.

Problem 2

(3 ✓s total)

The Johnson-Lindenstrauss lemma states that random projections preserve norms and distances. How about other basic geometric quantities? Does a projection preserving norms and distances necessarily preserve the following as well?

1. Angles: consider three points defining one angle by two of the line segments. If the projection preserves the length of the segments by a $(1+\epsilon)$ factor, does it also preserve the angle to a $(1+f(\epsilon))$ factor for some function f that is increasing in ϵ and tends to 0 as $\epsilon \to 0$? If so, give a proof. If not, give a counterexample.
2. Triangle area: setting as above.

Problem 3

(1 ✓)

We define the KL-divergence between distributions $\mathbf{p}$ and $\mathbf{q}$ as follows:

$$D_{\mathrm{KL}}(\mathbf{p}||\mathbf{q}) = \sum_i p_i \log \frac{p_i}{q_i}$$

Be careful that the KL-divergence is *not* symmetric, and furthermore is not a metric—it does not satisfy the triangle inequality over the set of distributions.

One important property of the KL-divergence is that it can be used to upper bound the ℓ_1 distance between two distributions. The following inequality is known as Pinsker's inequality:

Fact 1. *Given any two distributions* $\mathbf{p}, \mathbf{q}$ *over the same domain, we have*

$$\frac{1}{2}||\mathbf{p}-\mathbf{q}||_1 = d_{\mathrm{TV}}(\mathbf{p},\mathbf{q}) \leq \sqrt{\frac{1}{2} D_{\mathrm{KL}}(\mathbf{p}||\mathbf{q})}$$

The question is, can there be a converse to Pinsker's inequality in a qualitative sense, that is, some monotonic function f of the total variation distance that upper bounds the KL-divergence, which satisfies $\lim_{x\to 0} f(x) = 0$? If so, state an inequality and prove it (it is ok if it is an "obvious" or uninteresting inequality). If not, give a proof why not.

(Note: As we saw in class, the Hellinger distance does enjoy having both inequalities upper and lower bounding the total variation distance.)

Problem 4

In addition to Pinsker's inequality, another very important property of the KL divergence is that it is *additive* under taking products of distributions. Explicitly, for distributions $\mathbf{p}_1, \mathbf{p}_2, \mathbf{q}_1, \mathbf{q}_2$:

$$D_{\mathrm{KL}}(\mathbf{p}_1 \otimes \mathbf{p}_2, \mathbf{q}_1 \otimes \mathbf{q}_2) = D_{\mathrm{KL}}(\mathbf{p}_1, \mathbf{q}_1) + D_{\mathrm{KL}}(\mathbf{p}_2, \mathbf{q}_2)$$

This fact, combined with Pinsker's inequality, allows us to show lower bounds on the number of samples needed to distinguish between two distributions, much like how we use the Hellinger distance to do that as shown in class. Furthermore, the KL-divergence is sometimes a lot easier to calculate than the Hellinger distance, because of the fraction (inside the logarithm) in its definition.

Consider the geometric distributions $\text{Geom}(p)$ and $\text{Geom}(p+\epsilon)$ where $p < \frac{1}{3}$ and $\epsilon \ll p$.

1. (2 ✓, **Extra credit**) Show that $D_{\text{KL}}(\text{Geom}(p)||\text{Geom}(p+\epsilon)) = O(\epsilon^2/p^2)$.

   Warning: The calculations are somewhat annoying, and be careful to use *consistent* definitions of the Geometric distribution (there are 2 slightly different definitions).

   Hint: You will need to take the second order Taylor expansion of logarithms at some point. You may assume without proof that because $\epsilon \ll p$, the second order expansion is sufficient for this analysis. Specifically, you can replace $\log(1+x)$ with $x - x^2/2$ for the purposes of this analysis, even if this is technically incorrect but fixable.

2. (1 ✓) Hence show that it takes at least $\Omega(p^2/\epsilon^2)$ samples to distinguish between the two geometric distributions with probability 2/3.

3. (2 ✓) It is also possible to use the KL-divergence to show high probability lower bounds. The high probability Pinsker inequality states that, for any pair of distributions $\mathbf{p}, \mathbf{q}$ and for any event A (and its complement $\bar{A}$),

$$\mathbf{p}(A) + \mathbf{q}(\bar{A}) \geq \frac{1}{2} e^{-D_{\text{KL}}(\mathbf{p}||\mathbf{q})}$$

   Use this inequality to show that it takes $\Omega(\frac{p^2}{\epsilon^2} \log \frac{1}{\delta})$ samples to distinguish the two geometric distributions with probability at least $1 - \delta$. Depending on how you prove this, you may use the *data processing inequality* for the KL-divergence without proof.

(Note: The high probability Pinsker inequality is a more versatile tool for lower bounds than the lower bound discussed in class based on the squared Hellinger distance, particularly in situations where the sampling is more complicated than a fixed number of i.i.d. samples. As an example, a KL-divergence+(high probability) Pinsker's inequality argument can recover the same Bernoulli mean estimation lower bound we showed in class, up to constants.)