CSCI 4961/6961: Homework 1

Assigned Thursday September 10 2020. Due by 11:59pm Thursday September 17 2020.

Your answers must be LEGIBLE, clearly labeled, and provide well-written, clear, and convincing arguments. Please start early so you have ample time to see me during office hours.

Motivation: we will use sampling a lot in this course, to speed up computations. This HW looks at the types of sampling we might choose to use.

1. Often we want to approximate a large sum of numbers, $S = \sum_{i=1}^n x_i$ (imagine that $n = O(10^6)$). Here are two ways to approximate this sum, given an integer $k \ll n$:

   - **(Set sampling)** Sample uniformly from the set of all k-element subsets of $[n] := \{1, \ldots, n\}$ to select a set I of indices, then form the approximation
   $$\hat{S}_{\text{set}} = c_{\text{set}} \sum_{i \in I} x_i.$$
   The constant c_{set} is chosen so that $\hat{S}_{\text{set}}$ is an unbiased estimate of S, that is, $\mathbb{E}\hat{S}_{\text{set}} = S$.
   - **(Bernoulli sampling)** Sample n i.i.d. Bernoulli random variables $\alpha_i \sim \text{Bern}(\frac{k}{n})$ and form the approximation
   $$\hat{S}_{\text{bern}} = c_{\text{bern}} \sum_{i=1}^n \alpha_i x_i,$$
   where similarly to before, the constant c_{bern} is chosen so that $\hat{S}_{\text{bern}}$ is an unbiased estimate of S, that is, $\mathbb{E}\hat{S}_{\text{bern}} = S$.

   Compute the constants c_{set} and c_{bern} and compute the variances of $\hat{S}_{\text{set}}$ and $\hat{S}_{\text{bern}}$. Which method do you expect to give more accurate estimates of S, and why?

2. Let ν be the percentage of non-zero coefficients α_i in the Bernoulli sampling approach. What are the expectation and variance of ν?

3. Suppose instead $\mathbf{S} = \sum_{i=1}^n \mathbf{v}_i$ is the sum of vectors in $\mathbb{R}^d$. We can approximate this sum using the set sampling and Bernoulli sampling approaches. What are the constants c_{set} and c_{bern} in this case, and what are the covariance matrices of $\hat{\mathbf{S}}_{\text{set}}$ and $\hat{\mathbf{S}}_{\text{bern}}$?

4. To implement the set sampling method, we need to be able to sample uniformly from the set of k-element subsets of $[n]$.

   Here is an algorithm that attempts to do so: sample a number ℓ_1 uniformly at random from $[n]$, then sample a number ℓ_2 uniformly at random from $[n] \setminus \{\ell_1\}$, then sample a number ℓ_3 uniformly at random from $[n] \setminus \{\ell_1, \ell_2\}$; continue in this manner until you have finally sampled a number ℓ_k uniformly at random from $[n] \setminus \{\ell_1, \ldots, \ell_{k-1}\}$.

   Show that the set $\{\ell_1, \ldots, \ell_k\}$ returned by this procedure is indeed uniformly sampled from the set of k-element subsets of $[n]$:
   $$\mathbb{P}(\{\ell_1, \ldots, \ell_k\} = \{i_1, \ldots, i_k\}) = \frac{1}{\binom{n}{k}} \quad \text{for any } k\text{-element subset } \{i_1, \ldots, i_k\} \text{ of } [n].$$

5. (CSCI6961 students) If our goal is to minimize the variance of a sampling-based estimate of S while selecting on average only k terms in the sum, we can use a non-uniform Bernoulli sampling scheme
   $$\hat{S}_{\text{imp}} = \sum_{i=1}^n \frac{\alpha_i}{p_i} x_i \quad \text{where} \quad \alpha_i \sim \text{Bern}(p_i)$$
   and use *importance sampling probabilities* p_i that are chosen to minimize the variance while selecting only k non-zero summands on average. Let us determine these importance sampling probabilities.

(a) First, explain why the importance sampling probabilities can be determined by finding the p_i that minimize $\sum_{i=1}^{n} \frac{x_i^2}{p_i}$ subject to the constraint that $\sum_{i=1}^{n} p_i = k$, assuming that the solution of this optimization problem yields valid probabilities.

(b) Now, solve this optimization problem and state the importance sampling probabilities. To do so, it may help to recall the Lagrange multiplier method for minimizing $f(\mathbf{p})$ subject to the constraint that $g(\mathbf{p}) = 0$. It states that the minimizer satisfies the condition $\nabla f(\mathbf{p}) = \lambda \nabla g(\mathbf{p})$ for some $\lambda \in \mathbb{R}$. This optimality condition, plus the constraint $\sum_{i=1}^{n} p_i = k$, are sufficient to determine the importance sampling probabilities.

(c) Look at the form of these importance sampling probabilities and give an intuitive explanation for why they take that form.