Privacy in Statistics and Machine Learning Spring 2021
In-class Exercises for Lecture 16 (Synthetic data generation and MW-EM)
March 25/26, 2021

Adam Smith and Jonathan Ullman

Problems with marked with an asterisk **(*)** *are more challenging or open-ended.*

1. Consider the class of *3-way marginal queries*. Suppose data records are d-bit vectors (so $\mathcal{U} = \{0,1\}^d$). For any three distinct features $a, b, c \in [d]$, the marginal table $t_{a,b,c}(\mathbf{x})$ is an 8-dimensional vector with the frequencies in $\mathbf{x}$ of all $8 = 2^3$ possible combinations of values for features a, b, c. We can think ofthe table as 8 linear queries. The set of 3-way marginal queries contains all possible 3-way marginal tables.
   (a) What are m and K for this class of queries?
   (b) For error α, what is the sample size required by the accuracy bound we proved for simple Gaussian noise, the projection mechanism, and MW-EM?
   (c) If α is constant, which of these methods provides hte best ℓ_∞ guarantee? Does the answer change if we are just interested in an ℓ_2 guarantee?

2. What is the running time of MW-EM? Assume T is given, and that it takes $\Theta(1)$ time to evaluate $\varphi_i(x)$ for each $i \in [k]$ and x in $\mathcal{U}$. Assume that real arithmetic operations (exponentiation, summation, etc) take constant time. [If it is easier, assume the exponential mechanism is replaced with report-noisy-max.]

   Compare the running times of the Guassian and MW-EM mechanisms on three-way marginal queries.

3. What kinds of synthetic data distributions can MW-EM generate?
   (a) Show that the distributions $\mathbf{p}^t$ generated by the MW updates have the following feature: for each $x \in \mathcal{U}$, the probability $p^t(x)$ depends only on the values $(\varphi_i(x))_{i\in[k]}$.
   (b) Fix the data universe to $\mathcal{U} = \{0,1\}^d$ and consider the class of *one-way marginals*: these simply ask for the frequency of 1's in each of the d binary attributes.
   Show that running MW-EM for this class of queries can only produce distributions $\mathbf{p}^t$ under which the d attributes are independent.

4. What loss of generality is there in producing synthetic data as output for query release? Not much, it turns out. Suppose we have a differentially private algorithm that, for every $\mathbf{x} \in \mathcal{U}^n$, produces as output a ist of k values $\mathbf{a} = (a_1, \ldots, a_k)$ such that $\|\mathbf{a} - \mathbf{F}\mathbf{h}_\mathbf{x}\|_\infty \le \alpha$.
   (a) Show that we can post-process the algorithm's outputs to produce a synthetic data distribution $\hat{\mathbf{p}}$ such that $\|\mathbf{F}\hat{\mathbf{p}} - \mathbf{F}\mathbf{h}_\mathbf{x}\|_\infty \le 2\alpha$. [*Hint:* Search over $\Delta([m])$ to find a $\hat{\mathbf{p}}$ for which $\|\mathbf{F}\hat{\mathbf{p}} - \mathbf{a}\|_\infty$ is as small as possible.]
   (b) **(*)** Give a post-processing algorithm that runs in time $poly(m, k, 1/\alpha)$ and produces a $\hat{\mathbf{p}}$ such that $\|\mathbf{F}\hat{\mathbf{p}} - \mathbf{F}\mathbf{h}_\mathbf{x}\|_\infty \le 3\alpha$.