Privacy in Statistics and Machine Learning — Spring 2023
In-class Exercises for Lecture 7 (Recap)
February 9, 2023

Adam Smith (based on materials developed with Jonathan Ullman)

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

1. **Medians.** Suppose we want to find the median of a list of real numbers $\mathbf{x} = (x_1, \ldots, x_n)$ that lie in the set $\{1, \ldots, R\}$.

   Consider an instantiation of the exponential mechanism based on the following score function: For every $y \in \{1, \ldots, R\}$, let

   $$q(y; \mathbf{x}) = -\left|\sum_{i=1}^{n} sign(y - x_i)\right|$$

   where

   $$sign(z) = \begin{cases} 1 & \text{if } z > 0\,, \\ 0 & \text{if } z = 0\,, \\ -1 & \text{if } z < 0\,. \end{cases}$$

   If all the input values are distinct, this score is 0 exactly when y is a valid median for $\mathbf{x}$. In general, the score will be minimized at the true median.

   (a) Show that q has sensitivity at most 1 when neighboring data sets are allowed to differ by the insertion or deletion of one entry.

   (b) Let A_ε be the algorithm one gets by instantiating the exponential mechanism with score q, parameter ε and output set $\mathcal{Y} = \{1, \ldots, R\}$. Show that there is a constant $c > 0$ such that: for every data set $\mathbf{x}$, for every R and $\varepsilon < 1$, and for every $\beta \in (0, 1)$, the probability that $A_\varepsilon(\mathbf{x})$ samples a value y with $|rank_{\mathbf{x}}(y) - n/2| > c \cdot \frac{\ln(R)+\ln(1/\beta)}{\varepsilon}$ is at most β. Here $rank_{\mathbf{x}}(y) \in \{0, 1, \ldots, n\}$ is the position y would have in the sorted order of $\mathbf{x}$.

   For this part, it is ok to assume distinct data values, so that the rank of a value is uniquely defined.

   [*Hint:* How does $rank_{\mathbf{x}}(\cdot)$ relate to $q(\cdot; \mathbf{x})$? Look at the ratio between the probability mass of a true median and the probability mass of an element with very low or high rank.]