CS 208 - Applied Privacy for Data Science Homework 2

Jason Huang
Spring 2019 - Harvard University

The public Github repo containing all work is at https://github.com/TurboFreeze/cs208hw. All code has also been included in the appendix of this PDF as specified.

Problem 1

(a)

(i) The clamping function is effectively applying a post-processing function to the noisy query result. In other words, Laplace noise is added to the true mean $\bar{x}$, which must be $(\epsilon, 0)$-DP. The following clamping function does not change the privacy characteristics guaranteed by differential privacy, meaning that this mechanism **meets the definition** of $(\epsilon, 0)$-DP (following directly by privacy under post-processing and the proof of Laplace DP).

Note that the scale factor parameter of the Laplace distribution should be set to $s = GS_q/\epsilon$ for differential privacy, meaning that $\epsilon = GS_q/s$. In this case, the global sensitivity GS_q is the maximum change that can be affected to the statistic by a single entry's change, which in this case would be $1/n$ for the mean. Furthermore $s = 2/n$. The $\epsilon = (1/n)/(2/n) \implies \boxed{\epsilon = 0.5}$.

(ii) Constant ratios of Laplace mechanisms

(iii)

$$\frac{P[M(x', q) = r]}{P[M(x, q) = r]} =$$

(iv)

$$\begin{aligned} P[M(x,q) = r] &= P[[\bar{x} + Z]_0^1 = r] \\ &= \\ \frac{P[M(x,q) = r]}{P[M(x', q) = r]} &= \\ P[M(x,q) = r] &= P[\bar{x} + [Z]_{-1}^1 = r] \\ &= \end{aligned}$$

Problem 2

(a) The DGP is the following likelihood of some data vector $k \in \mathbb{N}^n$:

$$P(\mathbf{x} = \mathbf{k}) = \prod_{i=1}^{n} \frac{10^{\mathbf{k}_i} e^{-10}}{\mathbf{k}_i!}$$

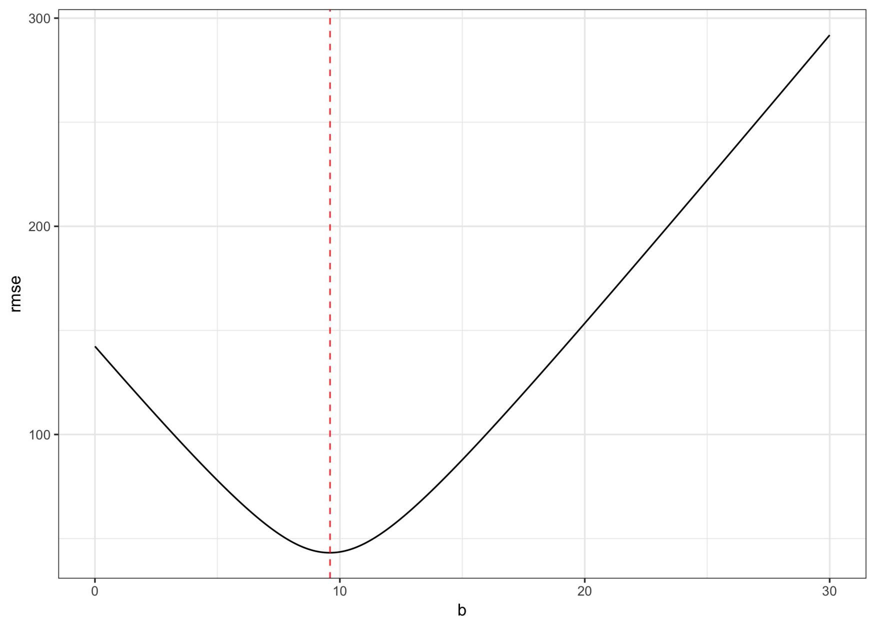
The DGP function was implemented using a Poisson random draw.

See the attached R script `q2.R` for the implementation.

(b) The first mechanism was chosen, involving clamping after Laplace noise has been added.

See the attached R script `q2.R` for the implementation.

(c) The optimal value b^* for b is $\boxed{b^* \approx 10}$. As expected, root mean squared error is indeed high with small clamping regions and decreases as it becomes more appropriate, with large clamping regions yielding high RMSE again.



See the attached R script `q2.R` for the implementation.

(d)

(e)

Problem 3

(a) There are differentially private techniques to release the means $\bar{y}$ and $\bar{x}$ as well as the slope $\hat{\beta}$. However, given the careful considerations needed for the slope $\hat{\beta}$, it may be challenging to come up with a single differentially private mechanism to derive the intercept estimate. However $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$ lends itself nicely to privacy preservation under composition and post-processing. Since there are three differentially private statistics needed here of $\bar{x}, \bar{y}, \hat{\beta}$, for a total epsilon budget of ϵ_t, then calculate differentially private releases of each statistic with $\epsilon = \epsilon_t/3$ to yield $(\epsilon_t/3, 0)$-DP statistics. Since post-processing is allowed without affecting privacy, then this three differentially private statistics will lead to $\epsilon_t/3 + \epsilon_t/3 + \epsilon_t/3 = \epsilon_t$ differential privacy for $\hat{\alpha}$ by composition and $\epsilon_t/3$ differential privacy for $\hat{\beta}$ (which is used in $\hat{\alpha}$ and does not require separate consumption of the budget). Therefore, the overall method for computing both $\hat{\alpha}$ and $\hat{\beta}$ would be ϵ_t-DP, as desired.

Since the data x_i is generated by a Poisson process according to the previous problem, it can be clamped using the optimal value of $b^* \approx 10$ found before. Since there is a linear relationship between x_i and y_i here (and it is known in the following part that the slope is simply 1), then similarly clamp y_i by $b^* \approx 10$.

See the attached R script `q3.R` for the implementation.

(b)

See the attached R script `q3.R` for the implementation.

(c)

See the attached R script `q3.R` for the implementation.

Problem 4

Use linearity of expectations and fundamental bridge to convert between probabilities and expectation of indicators.

$$\begin{aligned}
\mathbb{E}[\#\{i \in [n] : A(M(X))_i = X_i\}/n] &= \mathbb{E}[\mathbb{1}\{i \in [n] : A(M(X))_i = X_i\}/n] \\
&= \mathbb{E}[\sum_{i=1}^{n} \mathbb{1}(A(M(X))_i = X_i)/n] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\mathbb{1}(A(M(X))_i = X_i)] \\
&= \frac{1}{n}\sum_{i=1}^{n} P(A(M(X))_i = X_i)
\end{aligned}$$

Use the definition of (ϵ, δ)-DP

Appendix

Code for Problem 1

Code for Problem 2

Code for Problem 3