Online Appendix for "Information Design for Differential Privacy"

Ian M. Schmutte and Nathan Yoder*

December 6, 2022

S.1 Differential Privacy and Learning

Here, we offer an interpretation of differential privacy as a bound on Bayesian updating. Proposition S.1 shows that differential privacy is equivalent to a bound on the amount that learning the mechanism's output can cause an observer who believes respondents' types are independent to update their beliefs about a specific respondent's type θ_n. In particular, differential privacy limits the proportional change in the odds that the respondent is type 1. The argument mirrors that of Theorem 6.1 in Kifer and Machanavajjhala (2014). However, Proposition S.1 differs in that it bounds the change in the odds that the respondent has one type instead of another, rather than the change in the odds that it has a certain type instead of being absent from the data altogether.

Proposition S.1 (Interpretations of Differential Privacy)**.** *The following are equivalent:*

i. (S, m) *is* ϵ*-differentially private.*

ii. If an agent's prior $\hat{\pi}_0 \in \Delta(\Theta)$ *is a product distribution which places positive probability on both* $\theta_n = 1$ *and* $\theta_n = 0$*, then after observing a realization* s *from* (S, m)*, the log odds of the event* $\{\theta : \theta_n = 1\}$ *under the agent's posterior* $\hat{\pi}$ *can differ by no more than* ϵ *from its log odds under* $\hat{\pi}_0$*:*

$$\left| \log\left(\frac{\hat{\pi}(\{\theta : \theta_n = 1\})}{\hat{\pi}(\{\theta : \theta_n = 0\})}\right) - \log\left(\frac{\hat{\pi}_0(\{\theta : \theta_n = 1\})}{\hat{\pi}_0(\{\theta : \theta_n = 0\})}\right) \right| \leq \epsilon.$$

*Schmutte: University of Georgia, Terry College of Business, Department of Economics; E-mail: schmutte@uga.edu. Yoder: University of Georgia, Terry College of Business, Department of Economics; E-mail: nathan.yoder@uga.edu.

Proof. ((i)⇒(ii)): Suppose that (S, m) is differentially private and that an agent's prior $\hat{\pi}_0$ is a product distribution. Then for $t \in \{0,1\}$ we can write

$$\hat{\pi}(\{\theta : \theta_n = t\}) = \frac{\sum_{\theta:\theta_n=t} m(s|\theta)\hat{\pi}_0(\theta)}{\sum_{\theta\in\{0,1\}^N} m(s|\theta)\hat{\pi}_0(\theta)} = \frac{\sum_{\theta:\theta_n=t} m(s|\theta)\hat{\pi}_0(\{\hat{\theta} : \hat{\theta}_n = t\})\hat{\pi}_0(\{\hat{\theta} : \hat{\theta}_{-n} = \theta_{-n}\})}{\sum_{\theta\in\{0,1\}^N} m(s|\theta)\hat{\pi}_0(\theta)}$$

Hence

$$\begin{aligned}\left|\log\left(\frac{\hat{\pi}(\{\theta : \theta_n = 1\})}{\hat{\pi}(\{\theta : \theta_n = 0\})}\right) - \log\left(\frac{\hat{\pi}_0(\{\theta : \theta_n = 1\})}{\hat{\pi}_0(\{\theta : \theta_n = 0\})}\right)\right| &= \left|\log\left(\frac{\sum_{\theta:\theta_n=1} m(s|\theta)\hat{\pi}_0(\{\hat{\theta} : \hat{\theta}_{-n} = \theta_{-n}\})}{\sum_{\theta:\theta_n=0} m(s|\theta)\hat{\pi}_0(\{\hat{\theta} : \hat{\theta}_{-n} = \theta_{-n}\})}\right)\right| \\ &= \left|\log\left(\frac{\sum_{\theta_{-n}\in\{0,1\}^{N-1}} m(s|(1,\theta_{-n}))\hat{\pi}_0(\{\hat{\theta} : \hat{\theta}_{-n} = \theta_{-n}\})}{\sum_{\theta_{-n}\in\{0,1\}^{N-1}} m(s|(0,\theta_{-n}))\hat{\pi}_0(\{\hat{\theta} : \hat{\theta}_{-n} = \theta_{-n}\})}\right)\right|.\end{aligned}$$

Now we have

$$\begin{aligned}&\left(\min_{\theta_{-n}\in\{0,1\}^{N-1}}\left\{\frac{m(s|(1,\theta_{-n}))}{m(s|(0,\theta_{-n}))}\right\}\right)\left(\sum_{\theta_{-n}\in\{0,1\}^{N-1}} m(s|(0,\theta_{-n}))\hat{\pi}_0(\{\hat{\theta} : \hat{\theta}_{-n} = \theta_{-n}\})\right) \\ &\quad\le \sum_{\theta_{-n}\in\{0,1\}^{N-1}} m(s|(1,\theta_{-n}))\hat{\pi}_0(\{\hat{\theta} : \hat{\theta}_{-n} = \theta_{-n}\}) \\ &\quad\le \left(\max_{\theta_{-n}\in\{0,1\}^{N-1}}\left\{\frac{m(s|(1,\theta_{-n}))}{m(s|(0,\theta_{-n}))}\right\}\right)\left(\sum_{\theta_{-n}\in\{0,1\}^{N-1}} m(s|(0,\theta_{-n}))\hat{\pi}_0(\{\hat{\theta} : \hat{\theta}_{-n} = \theta_{-n}\})\right),\end{aligned}$$

and so

$$\begin{aligned}&\min_{\theta_{-n}\in\{0,1\}^{N-1}}\left\{\frac{m(s|(1,\theta_{-n}))}{m(s|(0,\theta_{-n}))}\right\} \le \frac{\sum_{\theta_{-n}\in\{0,1\}^{N-1} m(s|(1,\theta_{-n}))\hat{\pi}_0(\{\hat{\theta}:\hat{\theta}_{-n}=\theta_{-n}\})}}{\sum_{\theta_{-n}\in\{0,1\}^{N-1} m(s|(0,\theta_{-n}))\hat{\pi}_0(\{\hat{\theta}:\hat{\theta}_{-n}=\theta_{-n}\})}} \le \max_{\theta_{-n}\in\{0,1\}^{N-1}}\left\{\frac{m(s|(1,\theta_{-n}))}{m(s|(0,\theta_{-n}))}\right\}; \\ &\Rightarrow \log\left(\frac{\sum_{\theta_{-n}\in\{0,1\}^{N-1} m(s|(1,\theta_{-n}))\hat{\pi}_0(\{\hat{\theta}:\hat{\theta}_{-n}=\theta_{-n}\})}}{\sum_{\theta_{-n}\in\{0,1\}^{N-1} m(s|(0,\theta_{-n}))\hat{\pi}_0(\{\hat{\theta}:\hat{\theta}_{-n}=\theta_{-n}\})}}\right) \\ &\quad\le \max\left\{\left|\log\left(\min_{\theta_{-n}\in\{0,1\}^{N-1}}\left\{\frac{m(s|(1,\theta_{-n}))}{m(s|(0,\theta_{-n}))}\right\}\right)\right|, \left|\log\left(\max_{\theta_{-n}\in\{0,1\}^{N-1}}\left\{\frac{m(s|(1,\theta_{-n}))}{m(s|(0,\theta_{-n}))}\right\}\right)\right|\right\} \le \epsilon,\end{aligned}$$

as desired.

((ii)⇒(i)): Let $n \in \{1,\dots,N\}$ and let $\theta, \theta' \in \Theta$ be such that $\theta_{-n} = \theta'_{-n}$. Let $\hat{\pi}_0$ be such that $\hat{\pi}_0(\{\hat{\theta} : \hat{\theta}_{-n} = \theta_{-n}\}) = 1$ and $\hat{\pi}_0(\{\hat{\theta} : \hat{\theta}_n = 1\}) \in (0,1)$. Then for $t \in \{0,1\}$, $\hat{\pi}_0(\{\theta : \theta_n = t\}) = \hat{\pi}_0(\theta_{-n}, t)$ and $\hat{\pi}(\{\theta : \theta_n = t\}) = \frac{m(s|(t,\theta_{-n}))\hat{\pi}_0(\theta_{-n},t)}{m(s|(t,\theta_{-n}))\hat{\pi}_0((t,\theta_{-n}))+m(s|(1-t,\theta_{-n}))\hat{\pi}_0((1-t,\theta_{-n}))}$. Hence

$$\epsilon \ge \left|\log\left(\frac{\hat{\pi}(\{\theta : \theta_n = 1\})}{\hat{\pi}(\{\theta : \theta_n = 0\})}\right) - \log\left(\frac{\hat{\pi}_0(\{\theta : \theta_n = 1\})}{\hat{\pi}_0(\{\theta : \theta_n = 0\})}\right)\right| = \left|\log\left(\frac{m(s|(1,\theta_{-n}))}{m(s|(0,\theta_{-n}))}\right)\right| = \left|\log\left(\frac{m(s|\theta')}{m(s|\theta)}\right)\right|.$$

Since this holds for any $s \in S$, (i) follows. □

References

KIFER, D. AND A. MACHANAVAJJHALA (2014): "Pufferfish: A Framework for Mathematical Privacy Definitions," *ACM Transactions on Database Systems (TODS)*, 39, 1–36.