Supplementary Material for Priv'IT: *Priv*ate and Sample Efficient *I*dentity *T*esting

Bryan Cai
EECS, MIT
bcai@mit.edu

Constantinos Daskalakis
EECS and CSAIL, MIT
costis@csail.mit.edu

Gautam Kamath
EECS and CSAIL, MIT
g@csail.mit.edu

1 Proof of Theorem 3

We will prove the theorem for the case where $\beta = 1/3$, the general case follows at the cost of a multiplicative $\log(1/\beta)$ in the sample complexity from a standard amplification argument. To be more precise, we can consider splitting our dataset into $O(\log(1/\beta))$ sub-datasets and run the $\beta = 1/3$ test on each one independently. We return the majority result – since each test is correct with probability $\geq 2/3$, correctness of the overall test follows by Chernoff bound. It remains to argue privacy – note that a neighboring dataset will only result in a single sub-dataset being changed. Since we take the majority result, conditioning on the result of the other sub-tests, the result on this sub-dataset will either be irrelvant to or equal to the overall output. In the former case, any test is private, and in the latter case, we know that the individual test is ε-differentially private. Overall privacy follows by applying the law of total probability.

We require the following two claims, which give bounds on the random variables N_i and Y_i. Note that, due to the fact that we draw $Poisson(m)$ samples, each $N_i \sim Poisson(mp_i)$ independently.

Claim 1. $|Y_i| \leq \frac{2}{c_2\varepsilon} \log\left(\frac{1}{1-(1-c_2)^{1/|\mathcal{A}|}}\right)$ *simultaneously for all* $i \in \mathcal{A}$ *with probability exactly* $1 - c_2$.

Proof. The survival function of the folded Laplace distribution with parameter $2/c_2\varepsilon$ is $\exp(-c_2\varepsilon x/2)$, and the probability that a sample from it exceeding the value $\frac{2}{c_2\varepsilon} \log\left(\frac{1}{1-(1-c_2)^{1/|\mathcal{A}|}}\right)$ is equal to $1-(1-c_2)^{1/|\mathcal{A}|}$. The probability that probability that it does not exceed this value is $(1-c_2)^{1/|\mathcal{A}|}$, and since the Y_i's are independent, the probability that none exceeds this value is $1 - c_2$, as desired. □

Claim 2. $|N_i - mp_i| \leq \max\left\{4\sqrt{mp_i \log n}, \log n\right\}$ *simultaneously for all* $i \in \mathcal{A}$ *with probability at least* $1 - \frac{2}{n^{0.84}} - \frac{1.1}{n}$.

Proof. We consider this in two cases. Let X be a $Poisson(\lambda)$ random variable. First, assume that $\lambda \geq e^{-3} \log n$. By Bennett's inequality, we have the following tail bound [Pol15, Can17]:

$$\Pr\left[|X - \lambda| \geq x\right] \leq 2\exp\left(-\frac{x^2}{2\lambda}\psi\left(\frac{x}{\lambda}\right)\right),$$

where

$$\psi(t) = \frac{(1+t)\log(1+t) - t}{t^2/2}.$$

Consider $x = 4\sqrt{\lambda \log n}$. At this point, we have

$$\psi(x/\lambda) = \psi(4\sqrt{\log n/\lambda}) \geq \psi(4e^{3/2}) \geq 0.23.$$

Thus,

$$\begin{aligned}\Pr\left[|X-\lambda| \geq 4\sqrt{\lambda \log n}\right] &\leq 2\exp\left(-0.23 \cdot 8 \log n\right) \\ &\leq 2n^{-1.84}.\end{aligned}$$

Now, we focus on the other case, where $\lambda \leq e^{-3}\log n$. Here, we appeal to Proposition 1 of [Kla00], which implies the following via Stirling's approximation:

$$\Pr\left[|X-\lambda| \geq k\lambda\right] \leq \frac{k}{k-1}\exp(-\lambda + k\lambda - k\lambda \log k).$$

We set $k\lambda = \log n$, giving the upper bound

$$\frac{k}{k-1}n^{1-\log k} \leq 1.1 \cdot n^{-2}.$$

We conclude by taking a union bound over $[n]$, with the argument for each $i \in [n]$ depending on whether $\lambda = mp_i$ is large or small. □

We proceed with proving the two desiderata of this algorithm, correctness and privacy.

Correctness. We use the following two properties of the statistic $Z(D)$, which rely on the condition that $m = \Omega(\sqrt{n}/\alpha^2)$. The proofs of these properties are identical to the proofs of Lemma 2 and 3 in [ADK15], and are omitted.

Claim 3. *If $p = q$, then $\mathbf{E}[Z] = 0$. If $d_{\mathrm{TV}}(p,q) \geq \alpha$, then $\mathbf{E}[Z] \geq 1$.*

Claim 4. *If $p = q$, then $\mathbf{Var}[Z] \leq 1/1000$. If $d_{\mathrm{TV}}(p,q) \geq \alpha$, then $\mathbf{Var}[Z] \leq 1/1000 \cdot \mathbf{E}[Z]^2$.*

First, we note that, by Claim 1, the probability that we return in line 5 is exactly c_2. We now consider the case where $p = q$. We note that by Claim 2, the probability that we output "$p \neq q$" in line 10 is $o(1)$, and thus negligible. By Chebyshev's inequality, we get that $Z \leq 1/10$ with probability at least $9/10$, and we output "$p = q$" with probability at least $c_2/2 + (1-c_2)\cdot(9/10 - c_2)^2 \geq 2/3$ (note that we subtract c_2 from $9/10$ since we are conditioning on an event with probability $1 - c_2$, and by union bound). Similarly, when $d_{\mathrm{TV}}(p,q) \geq \alpha$, Chebyshev's inequality gives that $Z \geq 9/10$ with probability at least $9/10$, and therefore we output "$p \neq q$" with probability at least $2/3$.

Privacy. We will prove $(0, c_2\varepsilon/2)$-differential privacy. By Claim 1, the probability that we return in line 5 is exactly c_2. Thus the minimum probability of any output of the algorithm is at least $c_2/2$, and therefore $(0, c_2\varepsilon/2)$-differential privacy implies $(\varepsilon, 0)$-differential privacy.

We first consider the possibility of rejecting in line 11. Consider two neighboring datasets D and D', which differ by 1 in the frequency of symbol i. Coupling the randomness of the Y_j's on these two datasets, the only case in which the output differs is when Y_i is such that the value of $|N_i + Y_i - mq_i|$ lies on opposite sides of the threshold for the two datasets. Since N_i differs by 1 in the two datasets, and the probability mass assigned by the PDF of Y_i to any interval of length 1 is at most $c_2\varepsilon/4$, the probability that the outputs differ is at most $c_2\varepsilon/4$. Therefore, this step is $(0, c_2\varepsilon/4)$-differentially private.

We next consider the value of Z for two neighboring datasets D and D', where D' has one fewer occurrence of symbol i. We only consider the case where we have not already returned in line 11,

as otherwise the value of Z is irrelevant for determining the output of the algorithm.

$$\begin{aligned}
&Z(D) - Z(D') \\
&= \frac{1}{m\alpha^2}\left[\frac{(N_i - mq_i)^2 - N_i}{mq_i} - \frac{(N_i - 1 - mq_i)^2 - (N_i - 1)}{mq_i}\right] \\
&= \frac{1}{m\alpha^2}\left[\frac{(N_i - mq_i)^2 - N_i}{mq_i} - \frac{(N_i - mq_i)^2 - 2(N_i - mq_i) + 1 - N_i + 1}{mq_i}\right] \\
&= \frac{2(N_i - mq_i - 1)}{m^2\alpha^2 q_i}.
\end{aligned}$$

Since we did not return in line 11,

$$\begin{aligned}
|N_i - mq_i| &\le \frac{4}{c_2\varepsilon}\log\left(\frac{1}{1-(1-c_2)^{1/n}}\right) + \max\left\{4\sqrt{mq_i\log n}, \log n\right\} \\
&\le \frac{4\log(n/c_2)}{c_2\varepsilon} + \max\left\{4\sqrt{mq_i\log n}, \log n\right\}
\end{aligned}$$

This implies that

$$\begin{aligned}
|Z(D) - Z(D')| &= \frac{2|N_i - mq_i - 1|}{m^2\alpha^2 q_i} \\
&\le \frac{2}{m^2\alpha^2 q_i}\left(\frac{6\log(n/c_2)}{c_2\varepsilon} + 4\sqrt{mq_i\log n}\right).
\end{aligned}$$

We will enforce that each of these terms are at most $c_2\varepsilon/8$.

$$\begin{aligned}
\frac{12\log(n/c_2)}{m^2\alpha^2 q_i c_2\varepsilon} \le \frac{c_2\varepsilon}{8} &\Rightarrow m \ge \sqrt{\frac{96}{c_2^2 c_1}}\frac{\sqrt{n\log(n/c_2)}}{\alpha^{1.5}\varepsilon} \\
\frac{8\sqrt{\log n}}{m^{1.5}\alpha^2\sqrt{q_i}} \le \frac{c_2\varepsilon}{8} &\Rightarrow m \ge \left(\frac{64}{c_2\sqrt{c_1}}\right)^{2/3}\frac{(n\log n)^{1/3}}{\alpha^{5/3}\varepsilon^{2/3}}
\end{aligned}$$

Since both terms are at most $c_2\varepsilon/8$, this step is $(0, c_2\varepsilon/4)$-differentially private. Combining with the previous step gives the desired $(0, c_2\varepsilon/2)$-differential privacy, and thus (as argued at the beginning of the privacy section of this proof) ε-pure differential privacy.

References

[ADK15] Jayadev Acharya, Constantinos Daskalakis, and Gautam Kamath. Optimal testing for properties of distributions. In *Advances in Neural Information Processing Systems 28*, NIPS '15, pages 3577–3598. Curran Associates, Inc., 2015.

[Can17] Clément L. Canonne. A short note on Poisson tail bounds. `http://www.cs.columbia.edu/~ccanonne/files/misc/2017-poissonconcentration.pdf`, 2017.

[Kla00] Bernhard Klar. Bounds on tail probabilities of discrete distributions. *Probability in the Engineering and Informational Sciences*, 14(02):161–171, 2000.

[Pol15] David Pollard. A few good inequalities. `http://www.stat.yale.edu/~pollard/Books/Mini/Basic.pdf`, 2015.