Supplementary Material: Discrete Distribution Estimation Under Local Privacy

A. Proof of Theorem 2

As argued in the proof sketch of Theorem 2, it suffices to show that $r_{\ell,\varepsilon,k,n}(\boldsymbol{Q})$ obeys the data processing inequality. Precisely, we need to show that for any row stochastic matrix $\mathbf{W}$, $r_{\ell,\varepsilon,k,n}(\mathbf{W}\boldsymbol{Q}) \geq r_{\ell,\varepsilon,k,n}(\boldsymbol{Q})$. Observe that this is equivalent to showing that $r_{\ell,\varepsilon,k,n}(\boldsymbol{Q}) \geq r_{\ell,k,n}$, where $r_{\ell,k,n}$ is the minimax risk in the non-private setting.

Consider the set of all randomized estimators $\hat{\boldsymbol{p}}$. Under randomized estimators, the minimax risk is given by

$$r_{\ell,k,n} = \inf_{\hat{\boldsymbol{p}}} \sup_{\boldsymbol{p}\in\mathbb{S}^k} \mathop{\mathbb{E}}_{X^n\sim\boldsymbol{p},\hat{\boldsymbol{p}}} \ell(\boldsymbol{p},\hat{\boldsymbol{p}}),$$

where the expectation is taken over the randomness in the observations $X_1,\cdots,X_n$ and the randomness in $\hat{\boldsymbol{p}}$. Under a differentially private mechanism $\boldsymbol{Q}$, the minimax risk is given by

$$r_{\ell,\varepsilon,k,n}(\boldsymbol{Q}) = \inf_{\hat{\boldsymbol{p}}_{\boldsymbol{Q}}} \sup_{\boldsymbol{p}\in\mathbb{S}^k} \mathop{\mathbb{E}}_{Y^n\sim\boldsymbol{p}\boldsymbol{Q},\hat{\boldsymbol{p}}_{\boldsymbol{Q}}} \ell(\boldsymbol{p},\hat{\boldsymbol{p}}_{\boldsymbol{Q}}),$$

where the expectation is taken over the randomness in the private observations $Y_1,\cdots,Y_n$ and the randomness in $\hat{\boldsymbol{p}}_{\boldsymbol{Q}}$.

Assume that there exists a (potentially randomized) estimator $\hat{\boldsymbol{p}}^*_{\boldsymbol{Q}}$ that achieves $r_{\ell,\varepsilon,k,n}(\boldsymbol{Q})$. Consider the following randomized estimator: $\boldsymbol{Q}$ is first applied to $X_1,\cdots,X_n$ individually and $\hat{\boldsymbol{p}}^*_{\boldsymbol{Q}}$ is then jointly applied to the outputs of $\boldsymbol{Q}$. This estimator achieves a risk of $r_{\ell,\varepsilon,k,n}(\boldsymbol{Q})$. Therefore, $r_{\ell,k,n} \leq r_{\ell,\varepsilon,k,n}(\boldsymbol{Q})$.

If there is no estimator that can achieve $r_{\ell,\varepsilon,k,n}(\boldsymbol{Q})$, then there exists a sequence of (potentially randomized) estimators $\{\hat{\boldsymbol{p}}^i_{\boldsymbol{Q}}\}$ such that $\lim_{i\to\infty}\hat{\boldsymbol{p}}^i_{\boldsymbol{Q}}$ achieves the minimax risk. In other words, if $r^i_{\ell,\varepsilon,k,n}(\boldsymbol{Q})$ represents the risk under $\hat{\boldsymbol{p}}^i_{\boldsymbol{Q}}$, then $\lim_{i\to\infty} r^i_{\ell,\varepsilon,k,n}(\boldsymbol{Q}) = r_{\ell,\varepsilon,k,n}(\boldsymbol{Q})$. Using an argument similar to the one presented above, we get that $r_{\ell,k,n} \leq r^i_{\ell,\varepsilon,k,n}(\boldsymbol{Q})$. Taking the limit as i goes to infinity on both sides, we get that $r_{\ell,k,n} \leq r_{\ell,\varepsilon,k,n}(\boldsymbol{Q})$. This finishes the proof.

B. Proof of Proposition 3

Fix $\boldsymbol{Q}$ to $\boldsymbol{Q}_{\text{KRR}}$ and $\hat{\boldsymbol{p}}$ to be the empirical estimator given in (6). In this case, we have that

$$\begin{aligned}
\mathop{\mathbb{E}}_{Y^n\sim\boldsymbol{m}(\boldsymbol{Q}_{\text{KRR}})} ||\hat{\boldsymbol{p}}-\boldsymbol{p}||_2^2 &= \mathop{\mathbb{E}}_{Y^n\sim\boldsymbol{m}(\boldsymbol{Q}_{\text{KRR}})} \left|\left|\frac{e^\varepsilon+k-1}{e^\varepsilon-1}\hat{\boldsymbol{m}}-\frac{1}{e^\varepsilon-1}-\boldsymbol{p}\right|\right|_2^2 \\
&= \mathop{\mathbb{E}}_{Y^n\sim\boldsymbol{m}(\boldsymbol{Q}_{\text{KRR}})} \left|\left|\frac{e^\varepsilon+k-1}{e^\varepsilon-1}(\hat{\boldsymbol{m}}-\boldsymbol{m})\right|\right|_2^2 \\
&= \left(\frac{e^\varepsilon+k-1}{e^\varepsilon-1}\right)^2 \mathop{\mathbb{E}}_{Y^n\sim\boldsymbol{m}(\boldsymbol{Q}_{\text{KRR}})} ||\hat{\boldsymbol{m}}-\boldsymbol{m}||_2^2 \\
&= \left(\frac{e^\varepsilon+k-1}{e^\varepsilon-1}\right)^2 \frac{1-\sum_{i=1}^k m_i^2}{n} \\
&= \frac{1}{n}\left(\frac{e^\varepsilon+k-1}{e^\varepsilon-1}\right)^2 \left(1-\frac{\sum_{i=1}^k\left\{(e^\varepsilon-1)^2p_i^2+2(e^\varepsilon-1)p_i+1\right\}}{(e^\varepsilon+k-1)^2}\right) \\
&= \frac{(e^\varepsilon+k-1)^2-2(e^\varepsilon-1)-k-(e^\varepsilon-1)^2\sum_{i=1}^k p_i^2}{n(e^\varepsilon-1)^2} \\
&= \frac{((e^\varepsilon-1)+k)^2-2(e^\varepsilon-1)-k}{n(e^\varepsilon-1)^2}-\frac{(e^\varepsilon-1)^2}{n\,(e^\varepsilon-1)^2}+\frac{1}{n}-\frac{\sum_{i=1}^k p_i^2}{n} \\
&= \frac{(e^\varepsilon-1)^2+2k\,(e^\varepsilon-1)+k^2-2\,(e^\varepsilon-1)-k-(e^\varepsilon-1)^2}{n(e^\varepsilon-1)^2}+\frac{1-\sum_{i=1}^k p_i^2}{n} \\
&= \frac{2\,(k-1)\,(e^\varepsilon-1)+k\,(k-1)}{n(e^\varepsilon-1)^2}+\frac{1-\sum_{i=1}^k p_i^2}{n} \\
&= \frac{k-1}{n}\left(\frac{2\,(e^\varepsilon-1)+k}{(e^\varepsilon-1)^2}\right)+\frac{1-\sum_{i=1}^k p_i^2}{n},
\end{aligned}$$

and

$$\begin{aligned}
\mathop{\mathbb{E}}_{Y^n\sim\boldsymbol{m}(\boldsymbol{Q}_{\text{KRR}})} ||\hat{\boldsymbol{p}}-\boldsymbol{p}||_1 &= \left(\frac{e^\varepsilon+k-1}{e^\varepsilon-1}\right) \mathop{\mathbb{E}}_{Y^n\sim\boldsymbol{m}(\boldsymbol{Q}_{\text{KRR}})} ||\hat{\boldsymbol{m}}-\boldsymbol{m}||_1 \\
&= \left(\frac{e^\varepsilon+k-1}{e^\varepsilon-1}\right)\sum_{i=1}^k \mathbb{E}\,|m_i-\hat{m}_i| \\
&\approx \left(\frac{e^\varepsilon+k-1}{e^\varepsilon-1}\right)\sum_{i=1}^k \sqrt{\frac{2m_i(1-m_i)}{\pi n}} \\
&= \frac{1}{e^\varepsilon-1}\sum_{i=1}^k \sqrt{\frac{2((e^\varepsilon-1)p_i+1)((e^\varepsilon-1)(1-p_i)+k-1)}{\pi n}}.
\end{aligned}$$

C. Proof of Proposition 4

Fix $\boldsymbol{Q}$ to $\boldsymbol{Q}_{k\text{-RAPPOR}}$ and $\hat{\boldsymbol{p}}$ to be the empirical estimator given in (11), and let $C = \frac{e^{\varepsilon/2}-1}{e^{\varepsilon/2}+1}$, $B = \frac{1}{e^{\varepsilon/2}+1}$, and $A = e^{\varepsilon/2} - 1$. Then $C = BA$, $1 - B = e^{\varepsilon/2}B$, and from Section 4.2 $m_i = p_i C + B$. Using this notation, we have that

$$\begin{aligned}
\mathop{\mathbb{E}}_{Y^n \sim \boldsymbol{m}(\boldsymbol{Q}_{k\text{-RAPPOR}})} ||\hat{\boldsymbol{p}} - \boldsymbol{p}||_2^2 &= \mathop{\mathbb{E}}_{Y^n \sim \boldsymbol{m}(\boldsymbol{Q}_{k\text{-RAPPOR}})} \left|\left| \frac{e^{\varepsilon/2}+1}{e^{\varepsilon/2}-1} \hat{\boldsymbol{m}} - \frac{1}{e^{\varepsilon/2}-1} - \boldsymbol{p} \right|\right|_2^2 \\
&= \mathop{\mathbb{E}}_{Y^n \sim \boldsymbol{m}(\boldsymbol{Q}_{k\text{-RAPPOR}})} \left|\left| \frac{e^{\varepsilon/2}+1}{e^{\varepsilon/2}-1} (\hat{\boldsymbol{m}} - \boldsymbol{m}) \right|\right|_2^2 \\
&= \left(\frac{e^{\varepsilon/2}+1}{e^{\varepsilon/2}-1} \right)^2 \mathop{\mathbb{E}}_{Y^n \sim \boldsymbol{m}(\boldsymbol{Q}_{k\text{-RAPPOR}})} ||\hat{\boldsymbol{m}} - \boldsymbol{m}||_2^2 \\
&= \frac{1}{nC^2} \left(C + kB - \sum_{i=1}^k (p_i C + B)^2 \right) \\
&= \frac{1}{n} \left(1 - \sum_{i=1}^k p_i^2 \right) + \frac{1}{nC^2} \left(C - C^2 + kB - kB^2 - 2CB \right) \\
&= \frac{1}{n} \left(1 - \sum_{i=1}^k p_i^2 \right) + \frac{1}{nBA^2} \left(A - BA^2 + k(1-B) - 2BA \right) \\
&= \frac{1}{n} \left(1 - \sum_{i=1}^k p_i^2 \right) + \frac{1}{n} \frac{k e^{\varepsilon/2}}{(e^{\varepsilon/2}-1)^2},
\end{aligned}$$

and

$$\begin{aligned}
\mathop{\mathbb{E}}_{Y^n \sim \boldsymbol{m}(\boldsymbol{Q}_{\text{KRR}})} ||\hat{\boldsymbol{p}} - \boldsymbol{p}||_1 &= \left(\frac{e^{\varepsilon/2}+1}{e^{\varepsilon/2}-1} \right) \mathop{\mathbb{E}}_{Y^n \sim \boldsymbol{m}(\boldsymbol{Q}_{k\text{-RAPPOR}})} ||\hat{\boldsymbol{m}} - \boldsymbol{m}||_1 \\
&= \left(\frac{e^{\varepsilon/2}+1}{e^{\varepsilon/2}-1} \right) \sum_{i=1}^k \mathbb{E} \left| m_i - \hat{m}_i \right| \\
&\approx \left(\frac{e^{\varepsilon/2}+1}{e^{\varepsilon/2}-1} \right) \sum_{i=1}^k \sqrt{\frac{2m_i(1-m_i)}{\pi n}} \\
&= \sum_{i=1}^k \sqrt{\frac{2((e^{\varepsilon/2}-1)p_i + 1)((e^{\varepsilon/2}-1)(1-p_i)+1)}{\pi n (e^{\varepsilon/2}-1)^2}}.
\end{aligned}$$

D. Proof of Proposition 5

We want to show that for all $\boldsymbol{p} \in \mathbb{S}^k$ and all $\varepsilon \geq \ln k$,

$$\mathbb{E} \, ||\hat{\boldsymbol{p}}_{\text{KRR}} - \boldsymbol{p}||_2^2 \leq \mathbb{E} \, ||\hat{\boldsymbol{p}}_{\text{RAPPOR}} - \boldsymbol{p}||_2^2 , \tag{19}$$

where $\hat{\boldsymbol{p}}_{\text{KRR}}$ is the empirical estimate of $\boldsymbol{p}$ under k-RR, $\hat{\boldsymbol{p}}_{\text{RAPPOR}}$ is the empirical estimate of $\boldsymbol{p}$ under k-RAPPOR, and $\hat{\boldsymbol{p}}$ is the empirical estimator under k-RAPPOR.

From propositions 3 and 4, we have that

$$\mathbb{E} \, ||\hat{\boldsymbol{p}}_{\text{KRR}} - \boldsymbol{p}||_2^2 = \frac{1 - \sum_{i=1}^k p_i^2}{n} + \frac{k-1}{n} \left(\frac{2}{e^\varepsilon - 1} + \frac{k}{(e^\varepsilon - 1)^2} \right),$$

and

$$\mathbb{E} \, ||\hat{\boldsymbol{p}}_{\text{RAPPOR}} - \boldsymbol{p}||_2^2 = \frac{1 - \sum_{i=1}^k p_i^2}{n} + \frac{k e^{\varepsilon/2}}{n(e^{\varepsilon/2}-1)^2}.$$

Therefore, we just have to prove that

$$(k-1)\left(\frac{2}{e^{\varepsilon}-1}+\frac{k}{(e^{\varepsilon}-1)^2}\right)\leq\frac{ke^{\varepsilon/2}}{(e^{\varepsilon/2}-1)^2},$$

for $\varepsilon\geq\ln k$. Alternatively, we can show that

$$f(\varepsilon,k)=\frac{k}{k-1}\left(\frac{e^{\varepsilon}-1}{e^{\varepsilon/2}-1}\right)^2\frac{e^{\varepsilon/2}}{2e^{\varepsilon}+k-2}\geq 1,$$

for $\varepsilon\geq\ln k$. Observe that $f(\varepsilon,k)$ is an increasing function of ε and therefore, it suffices to show that

$$f(\ln k,k)=\frac{k}{k-1}\left(\frac{k-1}{\sqrt{k}-1}\right)^2\frac{\sqrt{k}}{3k-2}=\frac{k}{3k-2}\frac{\sqrt{k}(k-1)}{(\sqrt{k}-1)^2}\geq 1. \tag{20}$$

As a discrete function of $k\in\{2,3,....\}$, $f(\ln k,k)$ admits a unique minimum at $k=7$. Therefore, we just need to verify that $f(\ln 7,7)>1$. Indeed, $f(\ln 7,7)=3.1559>1$.

E. Discrete Distribution Estimation

Consider the $(k-1)$-dimensional probability simplex

$$\mathbb{S}^k=\{\boldsymbol{p}=(p_1,...,p_k)|p_i\geq 0,\sum_{i=1}^k p_i=1\}.$$

The discrete distribution estimation problem is defined as follows. Given a vector $\boldsymbol{p}\in\mathbb{S}^k$, samples $X_1,...,X_n$ are drawn i.i.d according to $\boldsymbol{p}$. Our goal is to estimate the probability vector $\boldsymbol{p}$ from the observation vector $X^n=(X_1,...,X_n)$.

An estimator $\hat{\boldsymbol{p}}$ is a mapping from X^n to a point in $\mathbb{S}^k$. The performance of $\hat{\boldsymbol{p}}$ may be measured via a loss function ℓ that computes a distance-like metric between $\hat{\boldsymbol{p}}$ and $\boldsymbol{p}$. Common loss functions include, among others, the absolute error loss $\ell_1(\boldsymbol{p},\hat{\boldsymbol{p}})=\sum_{i=1}^k|p_i-\hat{p}_i|$ and the quadratic loss $\ell_2^2(\boldsymbol{p},\hat{\boldsymbol{p}})=\sum_{i=1}^k(p_i-\hat{p}_i)^2$. The choice of the loss function depends on the application; for example, ℓ_1 loss is commonly used in classification and other machine learning applications. Given a loss function ℓ, the expected loss under $\hat{\boldsymbol{p}}$ after observing n i.i.d samples is given by

$$r_{\ell,k,n}(\boldsymbol{p},\hat{\boldsymbol{p}})=\mathop{\mathbb{E}}_{X^n\sim\text{Multimial}(n,\boldsymbol{p})}\ell(\boldsymbol{p},\hat{\boldsymbol{p}}). \tag{21}$$

E.1. Maximum likelihood and empirical estimation

In the absence of a prior on $\boldsymbol{p}$, a natural and commonly used estimator of $\boldsymbol{p}$ is the maximum likelihood (ML) estimator. The maximum likelihood estimate $\hat{\boldsymbol{p}}_{\text{ML}}$ of $\boldsymbol{p}$ is defined as

$$\hat{\boldsymbol{p}}_{\text{ML}}=\mathop{\text{argmax}}_{\boldsymbol{p}\in\mathbb{S}^k}\mathbb{P}\left(X_1,...,X_n|\boldsymbol{p}\right)$$

In this setting, it is easy to show that the maximum likelihood estimate is equivalent to the empirical estimator of $\boldsymbol{p}$, given by $\hat{p}_i=T_i/n$ where T_i is the frequency of element i. Observe that the empirical estimator is an unbiased estimator for $\boldsymbol{p}$ because $\mathbb{E}[\hat{p}_i]=p_i$ for any k,n, and i. Under maximum likelihood estimation, the ℓ_2^2 loss is the most tractable and simplest to analyze loss function. Because $T_i\sim\text{Binomial}(p_i,n)$, we have $\mathbb{E}[T_i]=np_i$, $\text{Var}(T_i)=np_i(1-p_i)$, and the expected ℓ_2^2 loss of the empirical estimator is given by

$$\begin{aligned}r_{\ell_2^2,k,n}(\boldsymbol{p},\hat{\boldsymbol{p}}_{\text{ML}}) &= \mathbb{E}||\hat{\boldsymbol{p}}_{\text{ML}}-\boldsymbol{p}||_2^2=\sum_{i=1}^k\mathbb{E}\left(\frac{T_i}{n}-p_i\right)^2\\ &= \sum_{i=1}^k\frac{\text{Var}(T_i)}{n^2}=\frac{1-\sum_{i=1}^k p_i^2}{n}.\end{aligned}$$

Let $\boldsymbol{p}_{\mathrm{U}} = \left(\frac{1}{k}, \cdots, \frac{1}{k}\right)$ and observe that

$$r_{\ell_2^2,k,n}(\boldsymbol{p}, \hat{\boldsymbol{p}}_{\mathrm{ML}}) \leq r_{\ell_2^2,k,n}(\boldsymbol{p}_{\mathrm{U}}, \hat{\boldsymbol{p}}_{\mathrm{ML}}) = \frac{1 - \frac{1}{k}}{n}. \tag{22}$$

In other words, the uniform distribution is the worst distribution for the empirical estimator under the ℓ_2^2 loss. From (Kamath et al., 2015), the asymptotic performance of the empirical estimator under the ℓ_1 loss functions is given by

$$r_{\ell_1,k,n}(\boldsymbol{p}, \hat{\boldsymbol{p}}_{\mathrm{ML}}) \approx \sum_{i=1}^{k} \sqrt{\frac{2p_i(1-p_i)}{\pi n}},$$

where $a_n \approx b_n$ means $\lim_{n\to\infty} a_n/b_n = 1$. As in the ℓ_2^2 case, notice that

$$r_{\ell_1,k,n}(\boldsymbol{p}, \hat{\boldsymbol{p}}_{\mathrm{ML}}) \leq r^{\ell_1}_{k,n}(\boldsymbol{p}_{\mathrm{U}}, \hat{\boldsymbol{p}}_{\mathrm{ML}}) \approx \sqrt{\frac{2(k-1)}{\pi n}}, \tag{23}$$

for any $\boldsymbol{p} \in \mathbb{S}^k$. In other words, the uniform distribution is the worst distribution for the empirical estimator under the ℓ_1 loss as well. Observe that the ℓ_1 loss scales as $\sqrt{k/n}$ whereas the ℓ_2^2 loss scales as $1/n$.

E.2. Minimax estimation

Another popular estimator that is widely studied in the absence of a prior is the minimax estimator $\hat{\boldsymbol{p}}_{\mathrm{MM}}$. The minimax estimator minimizes the expected loss under the worst distribution $\boldsymbol{p}$:

$$\hat{\boldsymbol{p}}_{\mathrm{MM}} = \operatorname*{argmin}_{\hat{\boldsymbol{p}}} \max_{\boldsymbol{p}\in\mathbb{S}^k} \mathop{\mathbb{E}}_{X^n\sim\boldsymbol{p}} \ell(\boldsymbol{p}, \hat{\boldsymbol{p}}). \tag{24}$$

The minimax risk is therefore defined as

$$r_{\ell,k,n} = \min_{\hat{\boldsymbol{p}}} \max_{\boldsymbol{p}\in\mathbb{S}^k} \mathop{\mathbb{E}}_{X^n\sim\boldsymbol{p}} \ell(\boldsymbol{p}, \hat{\boldsymbol{p}}).$$

For the ℓ_2^2 loss, it is shown in (Lehmann & Casella, 1998) that

$$\hat{p}_i = \frac{\frac{\sqrt{n}}{k} + \sum_{j=1}^{n} \mathbb{1}_{\{X_j=i\}}}{\sqrt{n}+n} = \frac{\frac{\sqrt{n}}{k} + T_i}{\sqrt{n}+n}, \tag{25}$$

is the minimax estimator, and that the minimax risk is

$$r_{\ell_2^2,k,n} = \frac{1-\frac{1}{k}}{(\sqrt{n}+1)^2}. \tag{26}$$

Observe that unlike the empirical estimator, the minimax estimator is not even asymptotically unbiased. Moreover, it improves on the empirical estimator only slightly (compare Equations (22) to (26)), increasing the the denominator from n to $n + 2\sqrt{n} + 1$ under the worst case distribution (the uniform distribution). This explains why the minimax estimator is almost never used in practice.

The minimax estimator under ℓ_1 loss is not known. However, the minimax risk is known for the case when k is fixed and n is increased. In this case, it is shown in (Kamath et al., 2015) that

$$r_{\ell_1,k,n} = \sqrt{\frac{2(k-1)}{\pi n}} + O\left(\frac{1}{n^{3/4}}\right). \tag{27}$$

Comparing Equations (23) to (27), we see that the worst case loss under the empirical estimator is again roughly as good as the minimax risk.

F. Maximum Likelihood Estimation for k-ary Mechanisms

F.1. k-RR

Proposition 6 *The maximum likelihood estimator of $\boldsymbol{p}$ under k-RR is given by*

$$\hat{p}_i = \left[\frac{T_i}{\lambda} - \frac{1}{e^\varepsilon - 1}\right]^+, \tag{28}$$

where $[x]^+ = \max(0, x)$, T_i is the frequency of element i calculated from Y^n, and λ is chosen so that

$$\sum_{i=1}^{k} \left[\frac{T_i}{\lambda} - \frac{1}{e^\varepsilon - 1}\right]^+ = 1. \tag{29}$$

Moreover, finding λ can be done in $O(k \log k)$ steps.

The proof of the above proposition is provided in Supplementary Section F.2.

F.2. Proof of Proposition 6

The maximum likelihood estimator under k-RR is the solution to

$$\hat{\boldsymbol{p}}_{\text{ML}} = \underset{\boldsymbol{p} \in \mathbb{S}^k}{\text{argmax}}\, \mathbb{P}\left(Y_1, ..., Y_n | \boldsymbol{p}\right),$$

where the Y_i's are the outputs of k-RR. Since the $\log(.)$ function is a monotonic function, the above maximum likelihood estimation problem is equivalent to

$$\hat{\boldsymbol{p}}_{\text{ML}} = \underset{\boldsymbol{p} \in \mathbb{S}^k}{\text{argmax}}\, \log \mathbb{P}\left(Y_1, ..., Y_n | \boldsymbol{p}\right).$$

Given that

$$\begin{aligned}
\mathbb{P}\left(Y_1, ..., Y_n | \boldsymbol{p}\right) &= \prod_{i=1}^{n} \mathbb{P}\left(Y_i | \boldsymbol{p}\right) \\
&= \prod_{i=1}^{n} \left(\sum_{j=1}^{k} \boldsymbol{Q}_{\text{KRR}}(Y_i | X_i = j) p_j \right),
\end{aligned}$$

we have that

$$\log \mathbb{P}\left(Y_1, ..., Y_n | \boldsymbol{p}\right) = \sum_{i=1}^{n} \log \left(\sum_{j=1}^{k} \boldsymbol{Q}_{\text{KRR}}(Y_i | X_i = j) p_j \right).$$

Observe that

$$\begin{aligned}
\sum_{j=1}^{k} \boldsymbol{Q}_{\text{KRR}}(Y_i | X_i = j) p_j &= \boldsymbol{Q}_{\text{KRR}}(Y_i | X_i = Y_i) p_{Y_i} + \sum_{j \neq Y_i} \boldsymbol{Q}_{\text{KRR}}(Y_i | X_i = j) p_j && (30) \\
&= \frac{e^\varepsilon}{e^\varepsilon + k - 1} p_{Y_i} + \frac{1}{e^\varepsilon + k - 1}(1 - p_{Y_i}) && (31) \\
&= \frac{1}{e^\varepsilon + k - 1}\left((e^\varepsilon - 1) p_{Y_i} + 1\right), && (32)
\end{aligned}$$

and therefore,

$$\sum_{i=1}^{n} \log \left(\sum_{j=1}^{k} \boldsymbol{Q}_{\text{KRR}}(Y_i | X_i = j) p_j \right) = \sum_{i=1}^{k} T_i \log \left(\frac{1}{e^\varepsilon + k - 1}\left((e^\varepsilon - 1) p_i + 1\right) \right),$$

where T_i is the number of Y's that are equal to i (i.e., the frequency of element i in the observed sequence Y^n). Thus, the maximum likelihood estimation problem under k-RR is equivalent to

$$\hat{\boldsymbol{p}}_{\text{ML}} = \underset{\boldsymbol{p}\in\mathbb{S}^k}{\text{argmax}} \sum_{i=1}^{k} T_i \log\left((e^{\varepsilon}-1)p_i + 1\right).$$

The above constrained optimization problem is a convex optimization problem that is well studied in the literature under the rubric of water-filling algorithms. From (Boyd & Vandenberghe, 2004), the solution to this problem is given by

$$\hat{p}_i = \left[\frac{T_i}{\lambda} - \frac{1}{e^{\varepsilon}-1}\right]^+,$$

where $[x]^+ = \max(0, x)$ and λ is chosen so that

$$\sum_{i=1}^{k}\left[\frac{T_i}{\lambda} - \frac{1}{e^{\varepsilon}-1}\right]^+ = 1.$$

Given the T_i's, $\boldsymbol{p}$ is computed according to the empirical estimator. If all the $\hat{p}_i$'s are non-negative, then the maximum likelihood estimate is the same as the empirical estimate. If not, $\hat{\boldsymbol{p}}$ is sorted, its negative entries are zeroed out, and lambda is computed according to the above equation. Given lambda, a new $\hat{\boldsymbol{p}}$ can be computed and the above process can be repeated until all the entries of $\hat{\boldsymbol{p}}$ are non-negative. Notice that sorting happens once and the process is repeated at most $k-1$ times. Therefore, the computational complexity of this algorithm is upper bounded by $k\log k + k$ which is $O(k\log k)$.

F.3. k-RAPPOR

Proposition 7 *The maximum likelihood estimator of $\boldsymbol{p}$ under k-RAPPOR is*

$$\underset{\boldsymbol{p}\in\mathbb{S}^k}{\textit{argmax}} \sum_{j=1}^{k}(n-T_j)\log\left((1-\delta)-(1-2\delta)p_j\right) + T_j\log\left((1-2\delta)p_j+\delta\right)$$

where $T_j = \sum_{i=1}^{n} Y_i^{(j)}$ and $\delta = 1/(e^{\varepsilon/2}+1)$.

The proof of the above proposition is provided in Supplementary Section F.4. Observe that unlike k-RR, a k-dimensional convex program has to be solved in this case to determine the maximum likelihood estimate of $\boldsymbol{p}$.

F.4. Proof of Proposition 7

The maximum likelihood estimator under k-RAPPOR is the solution to

$$\hat{\boldsymbol{p}}_{\text{ML}} = \underset{\boldsymbol{p}\in\mathbb{S}^k}{\text{argmax}}\ \mathbb{P}\left(Y_1, ..., Y_n|\boldsymbol{p}\right),$$

where the Y_i's are the outputs of k-RAPPOR. Since the $\log(.)$ function is a monotonic function, the above maximum likelihood estimation problem is equivalent to

$$\hat{\boldsymbol{p}}_{\text{ML}} = \underset{\boldsymbol{p}\in\mathbb{S}^k}{\text{argmax}}\ \log\mathbb{P}\left(Y_1, ..., Y_n|\boldsymbol{p}\right).$$

Recall that under k-RAPPOR, $Y_i = [Y_i^{(1)}, \cdots, Y_i^{(k)}]$ is a k-dimensional binary vector, which implies that

$$\mathbb{P}(Y_i^{(j)} = 1) = \left(\frac{e^{\varepsilon/2}-1}{e^{\varepsilon/2}+1}\right)p_j + \frac{1}{e^{\varepsilon/2}+1}, \tag{33}$$

for all $i \in \{1, \cdots, n\}$ and $j \in \{1, \cdots, k\}$. Therefore,

$$\begin{aligned}\log \mathbb{P}\left(Y_1, ..., Y_n | \boldsymbol{p}\right) &= \log \prod_{i=1}^{n} \prod_{j=1}^{k} \left(Y_i^{(j)}(p_j(1-\delta) + (1-p_j)\delta) + (1 - Y_i^{(j)})(p_j\delta + (1-p_j)(1-\delta)) \right) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{k} \log \left(Y_i^{(j)}(p_j(1-\delta) + (1-p_j)\delta) + (1 - Y_i^{(j)})(p_j\delta + (1-p_j)(1-\delta)) \right) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{k} \log \left((1-2\delta)(2Y_i^{(j)} - 1)p_j - Y_i^{(j)}(1-2\delta) + (1-\delta) \right),\end{aligned}$$

where $\delta = 1/(1 + e^{\varepsilon/2})$. Therefore, under k-RAPPOR, the maximum likelihood estimation problem is given by

$$\underset{\boldsymbol{p} \in \mathbb{S}^k}{\operatorname{argmax}} \sum_{j=1}^{k} (n - T_j) \log \left((1-\delta) - (1-2\delta)p_j \right) + T_j \log \left((1-2\delta)p_j + \delta \right)$$

where $T_j = \sum_{i=1}^{n} Y_i^{(j)}$.

G. Conditions for Accurate Decoding under k-RR

For accurate decoding, we must satisfy three criteria: (i) k and C must be large enough that the input strings to be distinguishable, (ii) k and C must be large enough that the linear system in (18) is not underconstrained, and (iii) n must be large enough that the variance on estimated probability vector $\hat{\boldsymbol{p}}$ is small.

Let us first consider string distinguishability. Each string $s \in \mathcal{S}$ is associated with a C-tuple of hashes it can produces in the various cohorts: $\text{HASH}_{\mathcal{C}}^{(k)}(s) = \langle \text{HASH}_1^{(k)}(s), \text{HASH}_2^{(k)}(s), \cdots, \text{HASH}_C^{(k)}(s) \rangle \in \mathcal{X}^C$. Two strings $s_i \in \mathcal{S}$ and $s_j \in \mathcal{S}$ are distinguishable from one another under the encoding scheme if $\text{HASH}_{\mathcal{C}}^{(k)}(s_i) \neq \text{HASH}_{\mathcal{C}}^{(k)}(s_j)$, and a string s is distinguishable within the set $\mathcal{S}$ if $\text{HASH}_{\mathcal{C}}^{(k)}(s) \neq \text{HASH}_{\mathcal{C}}^{(k)}(s_j) \forall s_j \in \mathcal{S} \backslash s$.

Because $\text{HASH}_{\mathcal{C}}^{(k)}(s)$ is distributed uniformly over $\mathcal{X}^C$, $\mathbb{P}(\text{HASH}_{\mathcal{C}}^{(k)}(s) = \boldsymbol{x}_{\mathcal{C}}) \approx \frac{1}{k^C}$ for all $\boldsymbol{x}_{\mathcal{C}} \in \mathcal{X}^C$. It follows that the probability of two strings being distinguishable is also $\frac{1}{k^C}$. Furthermore, the probability that exactly one string from $\mathcal{S}$ produces the hash tuple $\boldsymbol{x}_{\mathcal{C}}$ is:

$$\text{Binomial}(1; \frac{1}{k^C}, S) = \frac{S(k^C - 1)^{S-1}}{(k^C)^S}$$

Thus, the expected number of $\boldsymbol{x}_{\mathcal{C}} \in \mathcal{X}^C$ associated with exactly one string in $\mathcal{S}$, which is also the expected number of distinguishable strings in a set $\mathcal{S}$ is:

$$\sum_{\boldsymbol{x}_{\mathcal{C}} \in \mathcal{Y}^C} \left(\frac{S(k^C - 1)^{S-1}}{(k^C)^S} \right) = S \left(\frac{k^C - 1}{k^C} \right)^{S-1} \tag{34}$$

and the probability that a string s is distinguishable within the set $\mathcal{S}$ is $\left(\frac{k^C - 1}{k^C} \right)^{S-1}$.

Consider a probability distribution $\boldsymbol{p} \in \mathbb{S}^S$. The expected recoverable probability mass is the the mass associated with the distinguishable strings within the set $\mathcal{S}$ is $\sum_{s \in \mathcal{S}} p_s \left(\frac{k^C - 1}{k^C} \right)^{S-1} = \left(\frac{k^C - 1}{k^C} \right)^{S-1}$ Therefore, if we hope to recover at least P_t of the probability mass, we require $\left(\frac{k^C - 1}{k^C} \right)^{S-1} \geq P_t$, or equivalently, $k^C \geq \frac{1}{1 - P_t^{\frac{1}{S-1}}}$.

Now consider ensuring that the linear system in (18) is not underconstrained. The system has S variables and kC independent equations. Thus, the system is not underconstrained so long as $kC \geq S$.

H. Supplementary Figures

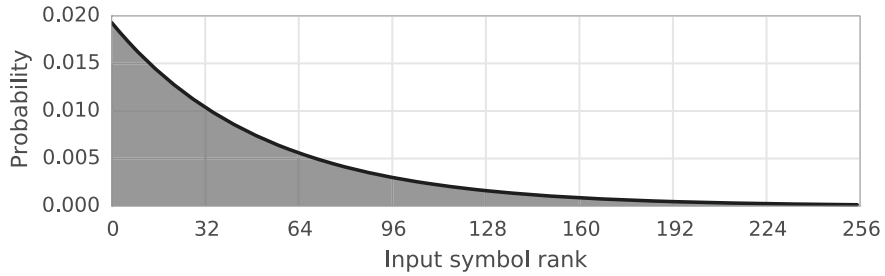


Figure 4: The true input distribution $\boldsymbol{p}$ for open-set and closed-set experiments in sections 4.4 and 5 is the geometric distribution with mean at $|\text{input alphabet}|/5$, truncated and renormalized. In the k-ary experiments of Section 4.4, the input alphabet is size k; in the open alphabet experiments of Section 5, the input alphabet is size $S = 256$.

Figure 5: The improvement in ℓ_2 decoding of the projected k-RR decoder (left) and projected k-RAPPOR decoder (right). This figure demonstrates that the same patterns hold in ℓ_2 as in ℓ_1 for the conditions shown in Figure 1.

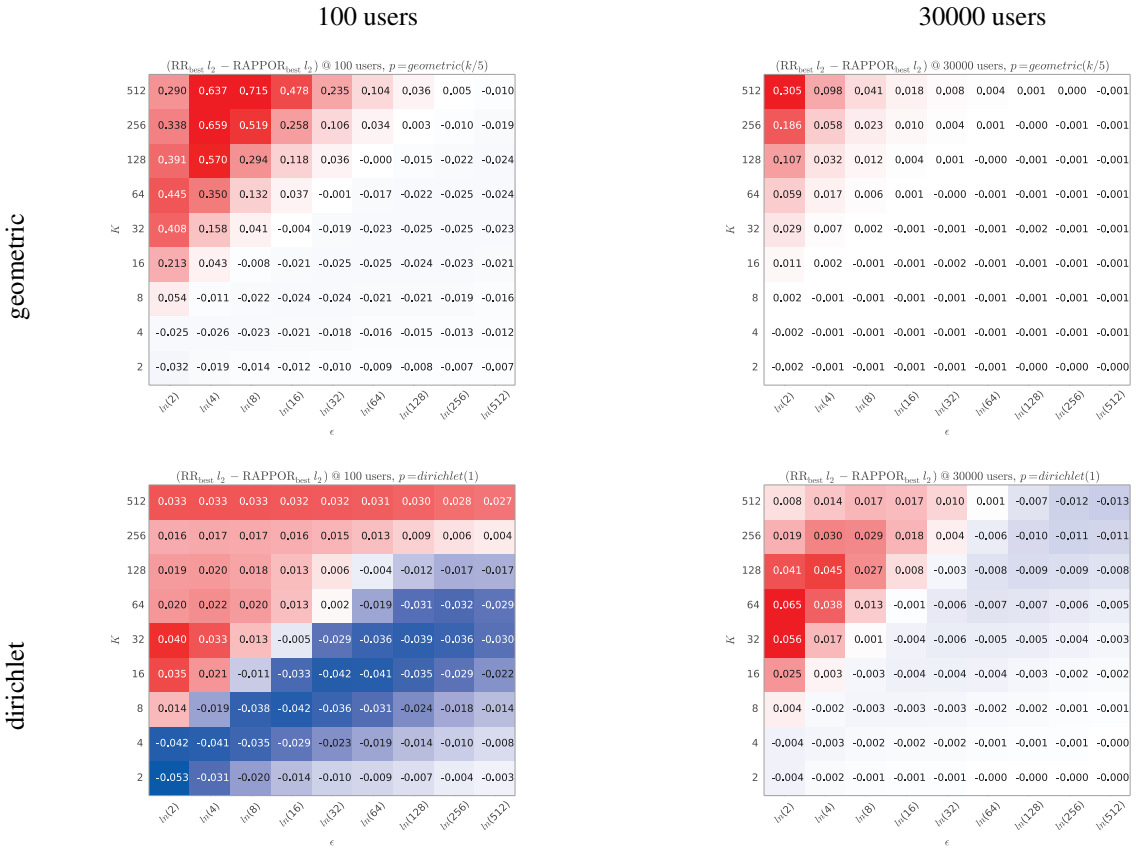
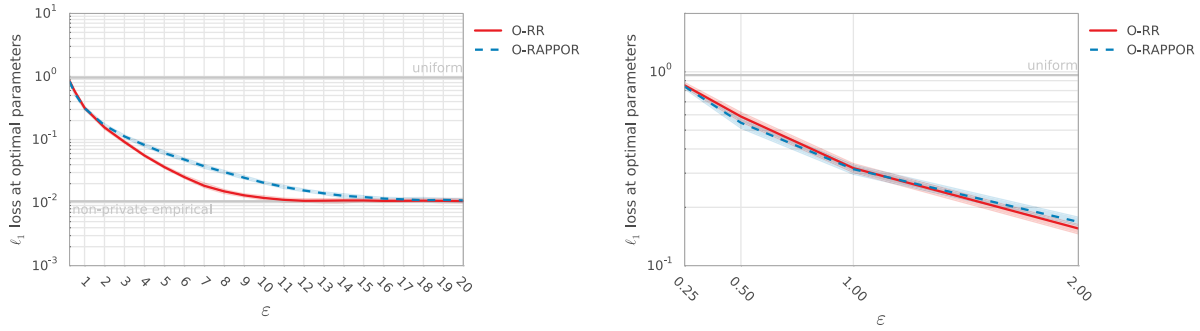


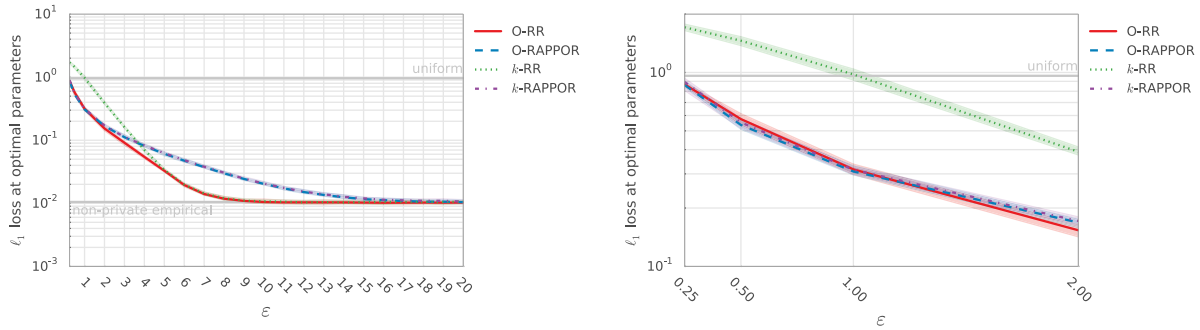
Figure 6: The improvement (negative values, blue) of the best k-RR decoder over the best k-RAPPOR decoder varying the size of the alphabet k (rows) and privacy parameter ε (columns). This figure demonstrates that the same patterns hold in ℓ_2 as in ℓ_1 for the conditions shown in Figure 2.



(a) Full ε range.

(b) Low ε range.

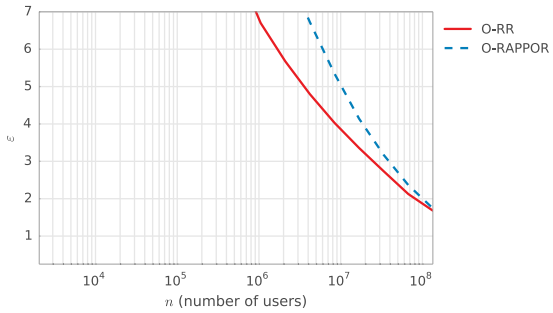
Figure 7: ℓ_1 loss when decoding open alphabets using the O-RR and O-RAPPOR for $n = 10^6$ users with input drawn from an alphabet of $S = 256$ symbols under a geometric distribution with mean=$S/5$, as depicted in Figure 4. Free parameters are set via grid search over $k \in [2, 4, 8, \ldots, 2048, 4096]$, $c \in [1, 2, 4, \ldots, 512, 1024]$, $h \in [1, 2, 4, 8, 16]$ to minimize the median loss over 50 samples at the given ε value. Lines show median ℓ_1 loss while the (narrow) shaded regions indicate 90% confidence intervals (over 50 samples). Baselines indicate expected loss from (1) using an empirical estimator directly on the input $\boldsymbol{s}$ and (2) using the uniform distribution as the $\hat{\boldsymbol{p}}$ estimate.



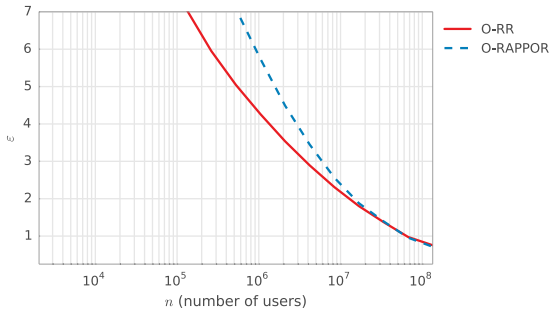
(a) Full ε range.

(b) Low ε range.

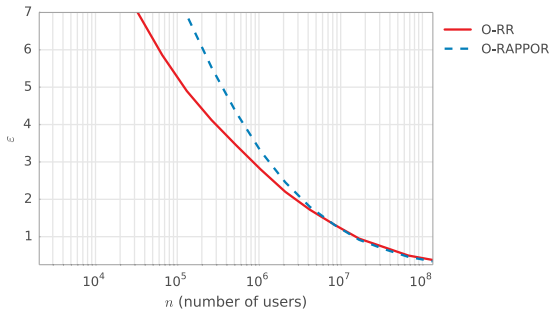
Figure 8: ℓ_1 loss when decoding decoding a known alphabet using the O-RR and O-RAPPOR (via permutative perfect hash functions) for $n = 10^6$ users with input drawn from an alphabet of $S = 256$ symbols under a geometric distribution with mean=$S/5$, as depicted in Figure 4. Free parameters are set via grid search over $k \in [2, 4, 8, \ldots, 2048, 4096]$, $c \in [1, 2, 4, \ldots, 512, 1024]$, $h \in [1, 2, 4, 8, 16]$ to minimize the median loss over 50 samples at the given ε value. Lines show median ℓ_1 loss while the (narrow) shaded regions indicate 90% confidence intervals (over 50 samples). Note that the k-RAPPOR and O-RAPPOR lines in (b) are nearly indistinguishable. Baselines indicate expected loss from (1) using an empirical estimator directly on the input $\boldsymbol{s}$ and (2) using the uniform distribution as the $\hat{\boldsymbol{p}}$ estimate.



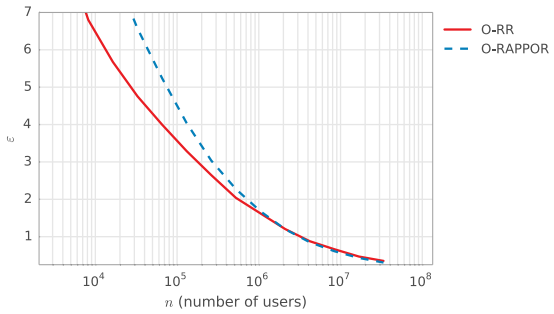
(a) $\ell_1 = 0.02$



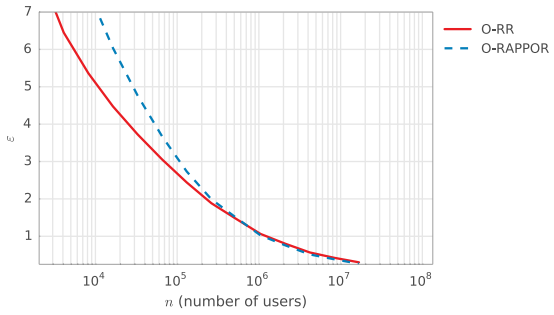
(b) $\ell_1 = 0.05$



(c) $\ell_1 = 0.10$

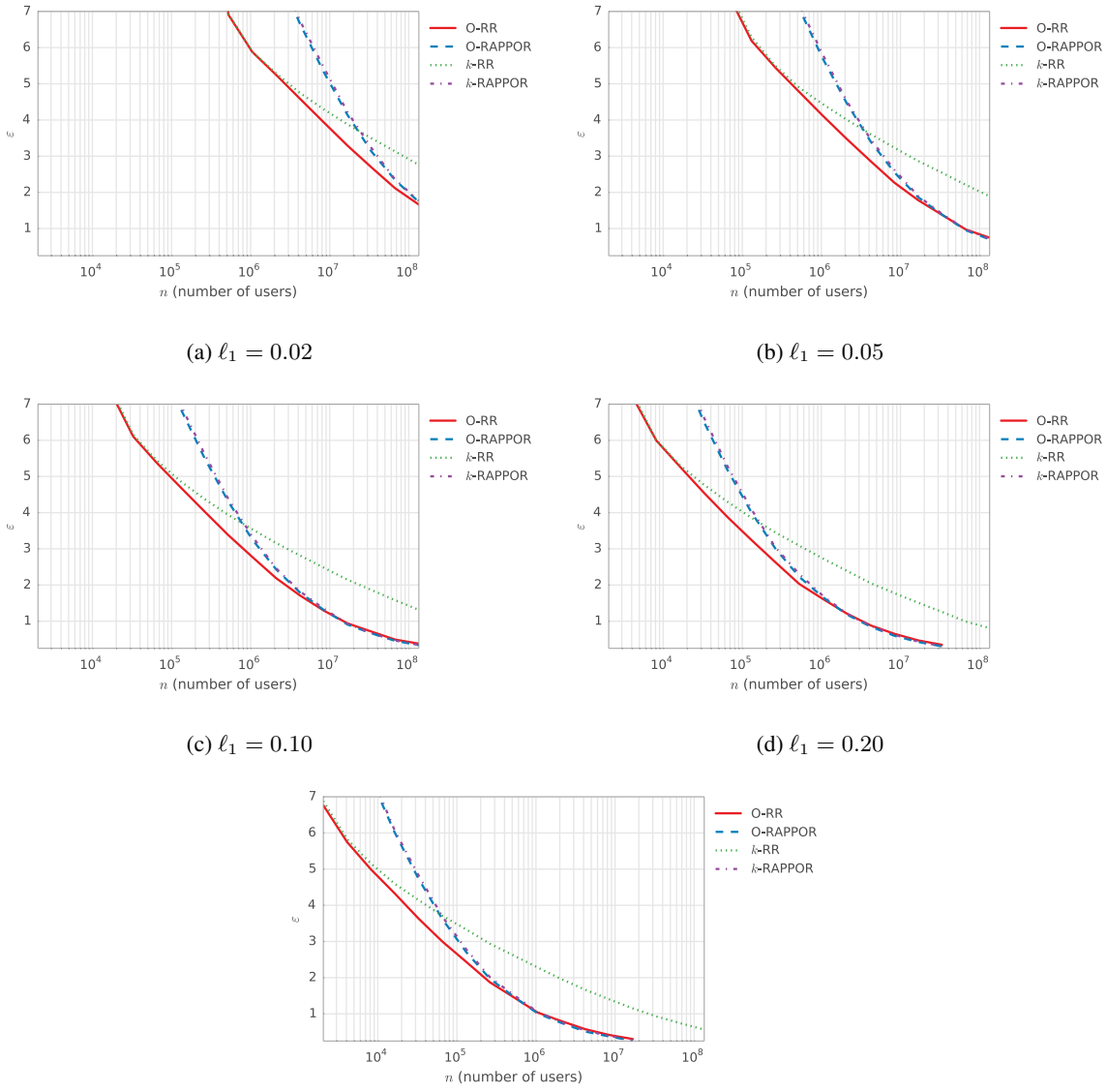


(d) $\ell_1 = 0.20$



(e) $\ell_1 = 0.30$

Figure 9: Taking ℓ_1 loss (the utility) and n (the number of users) as fixed requirements (as is the case in many practical scenarios), we approximate the degree of privacy ε that can be obtained under O-RR and O-RAPPOR for open alphabets (lower ε is better). Input is generated from an alphabet of $S = 256$ symbols under a geometric distribution with mean=$S/5$, as depicted in Figure 4. Free parameters are set via grid search to minimize the median loss over 50 samples at the given ε and fixed parameter values.



(a) $\ell_1 = 0.02$

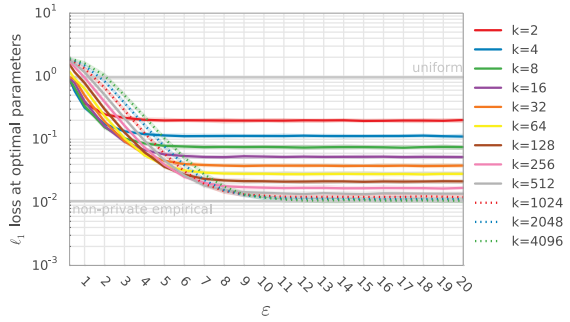
(b) $\ell_1 = 0.05$

(c) $\ell_1 = 0.10$

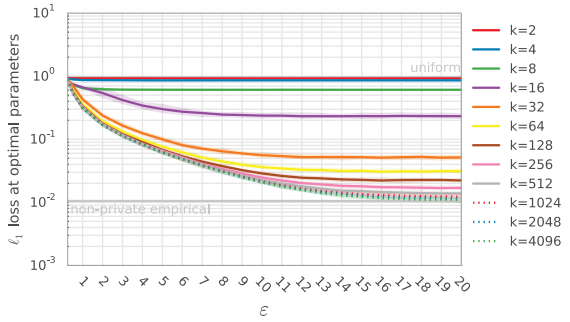
(d) $\ell_1 = 0.20$

(e) $\ell_1 = 0.30$

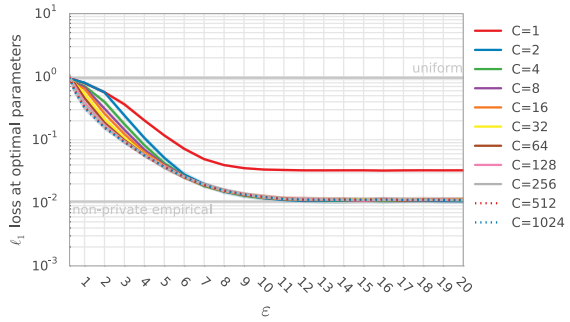
Figure 10: Taking ℓ_1 loss (the utility) and n (the number of users) as fixed requirements (as is the case in many practical scenarios), we approximate the degree of privacy ε that can be obtained under O-RR and O-RAPPOR for closed alphabets (lower ε is better). Input is generated from an alphabet of $S = 256$ symbols under a geometric distribution with mean=$S/5$, as depicted in Figure 4. Free parameters are set via grid search to minimize the median loss over 50 samples at the given ε and fixed parameter values.



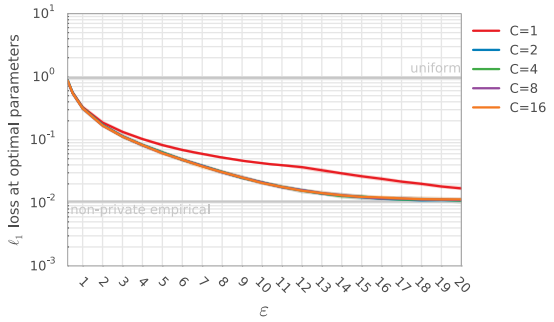
(a) O-RR varying k



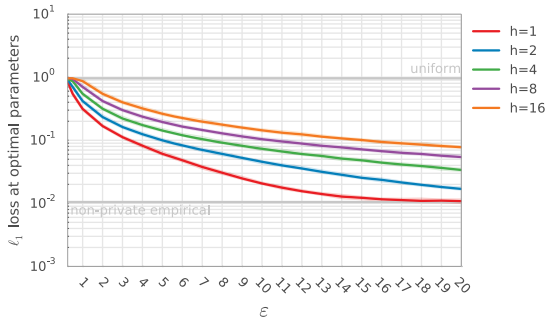
(b) O-RAPPOR varying k



(c) O-RR varying C

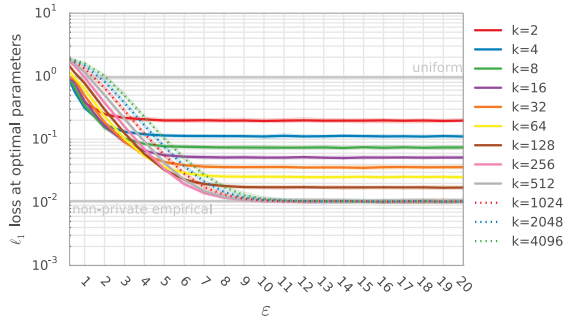


(d) O-RAPPOR varying C

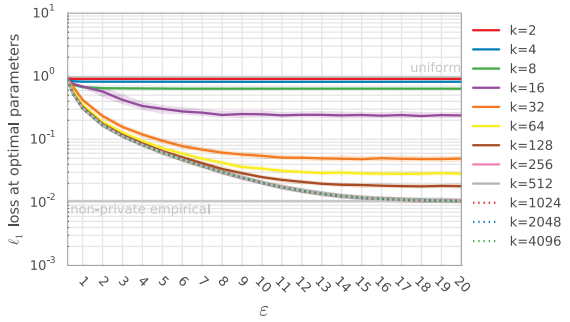


(e) O-RAPPOR varying h

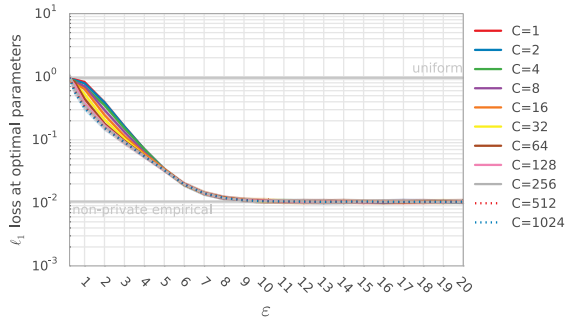
Figure 11: ℓ_1 loss when decoding open alphabets using O-RR and O-RAPPOR under various parameter settings, for $n = 10^6$ users with input drawn from an alphabet of $S = 4096$ symbols under a geometric distribution with mean=$S/5$. Remaining free parameters are set via grid search to minimize the median loss over 50 samples at the given ε and fixed parameter values. Lines show median ℓ_1 loss while the (narrow) shaded regions indicate 90% confidence intervals (over 50 samples for the optimal parameter settings.)



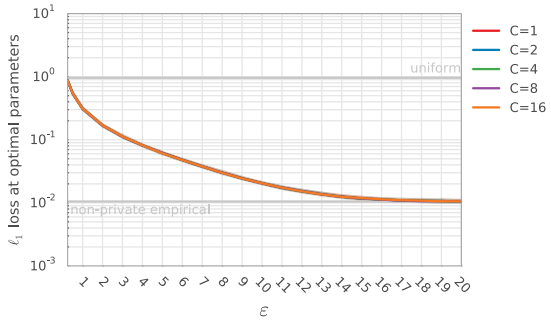
(a) O-RR varying k



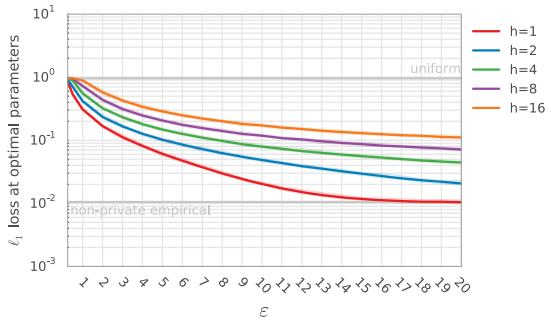
(b) O-RAPPOR varying k



(c) O-RR varying C

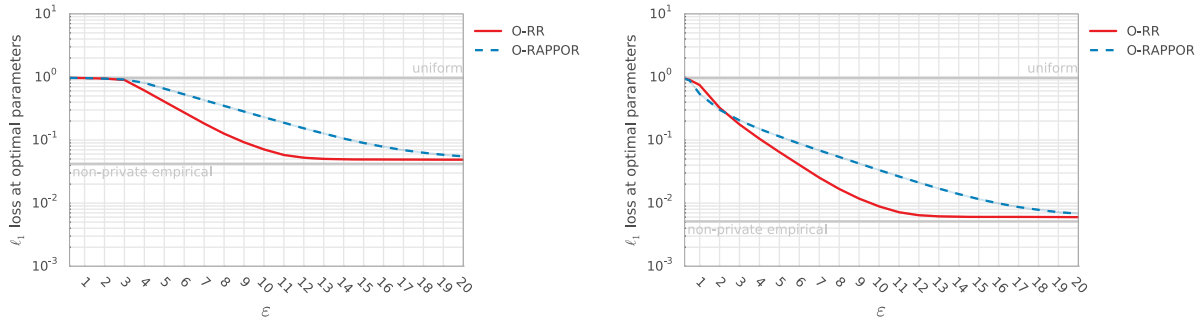


(d) O-RAPPOR varying C



(e) O-RAPPOR varying h

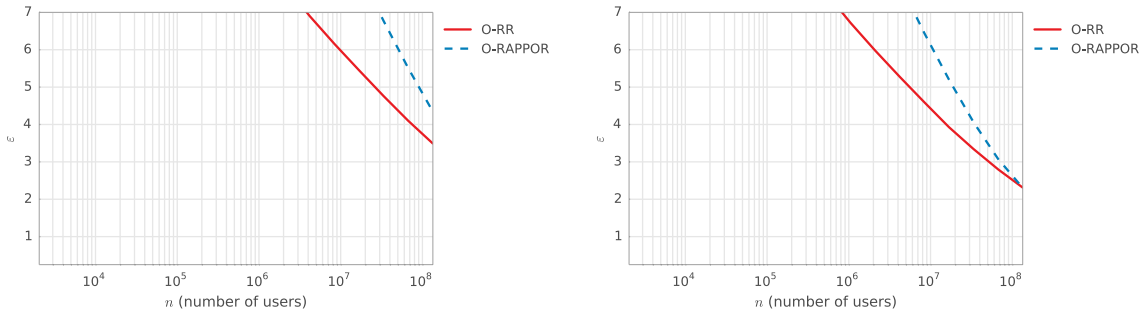
Figure 12: ℓ_1 loss when decoding closed alphabets using the O-RR and O-RAPPOR under various parameter settings, for $n = 10^6$ users with input drawn from an alphabet of $S = 4096$ symbols under a geometric distribution with mean=$S/5$. Remaining free parameters are set via grid search to minimize the median loss over 50 samples at the given ε and fixed parameter values. Lines show median ℓ_1 loss while the (narrow) shaded regions indicate 90% confidence intervals (over 50 samples for the optimal parameter settings.)



(a) $n = 10^6$ users

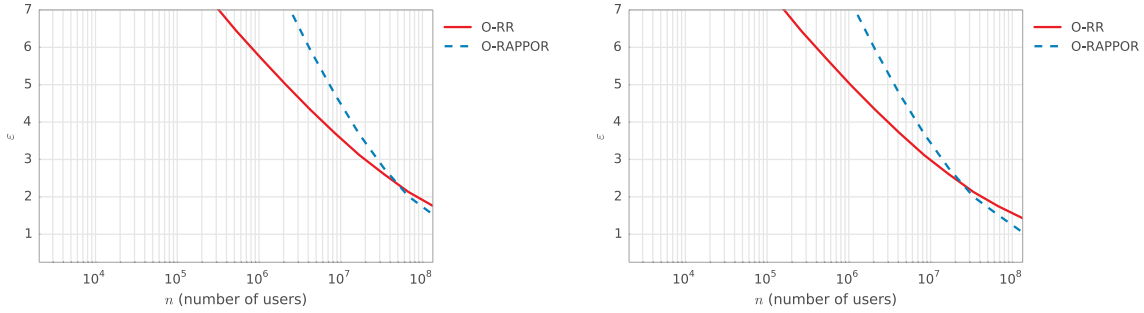
(b) $n = 10^8$ users

Figure 13: ℓ_1 loss when decoding open alphabets using the O-RR and O-RAPPOR, with input drawn from an alphabet of $S = 4096$ symbols under a geometric distribution with mean=$S/5$. Free parameters are set via grid search over $k \in [2, 4, 8, \ldots, 8192, 16384]$, $c \in [1, 2, 4, \ldots, 512, 1024]$, $h \in [1, 2]$ to minimize the median loss over 50 samples at the given ε value. Lines show median ℓ_1 loss while the (narrow) shaded regions indicate 90% confidence intervals (over 50 samples). Baselines indicate expected loss from (1) using an empirical estimator directly on the input $\boldsymbol{s}$ and (2) using the uniform distribution as the $\hat{\boldsymbol{p}}$ estimate.



(a) $\ell_1 = 0.10$

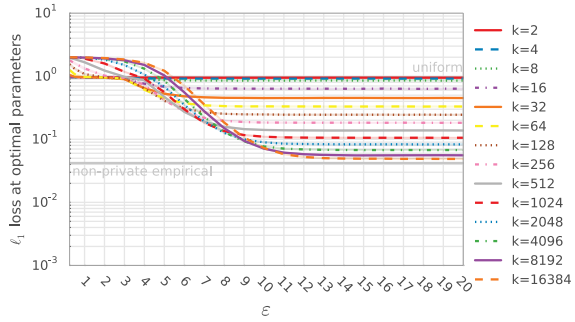
(b) $\ell_1 = 0.20$



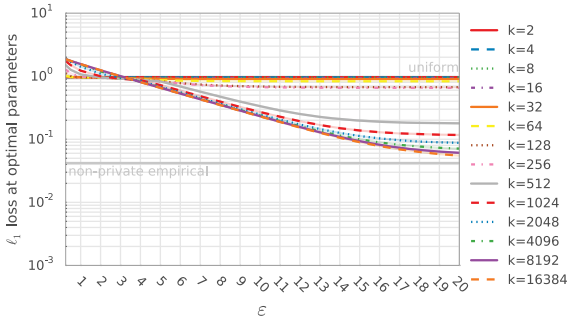
(c) $\ell_1 = 0.30$

(d) $\ell_1 = 0.40$

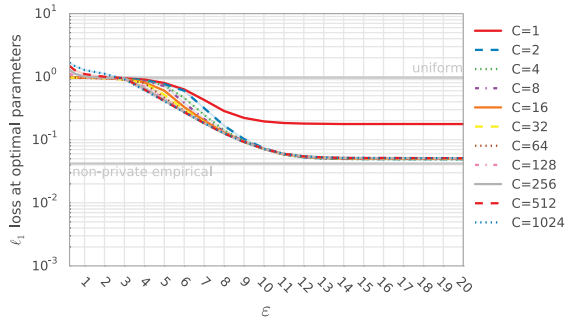
Figure 14: Taking ℓ_1 loss (the utility) and n (the number of users) as fixed requirements (as is the case in many practical scenarios), we approximate the degree of privacy ε that can be obtained under O-RR and O-RAPPOR for open alphabets (lower ε is better). Input is generated from an alphabet of $S = 4096$ symbols under a geometric distribution with mean=$S/5$, as depicted in Figure 4. Free parameters are set via grid search to minimize the median loss over 50 samples at the given ε and fixed parameter values.



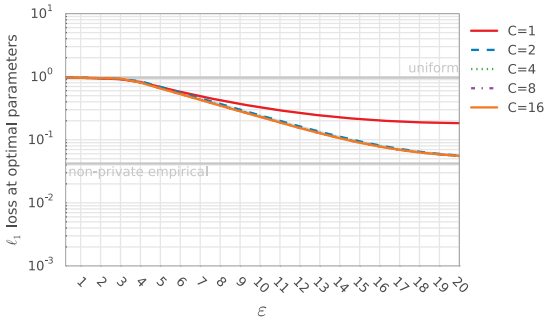
(a) O-RR varying k



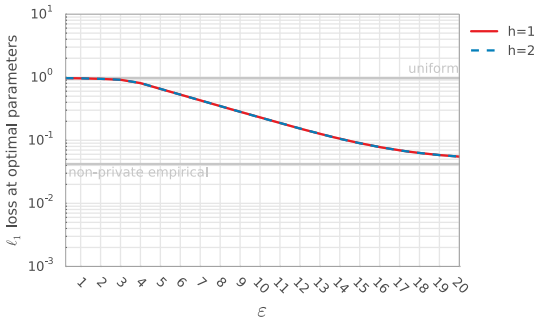
(b) O-RAPPOR varying k



(c) O-RR varying C

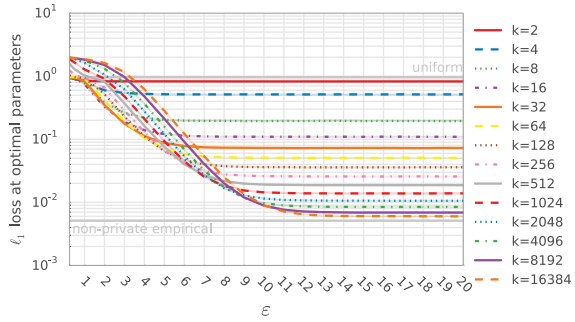


(d) O-RAPPOR varying C

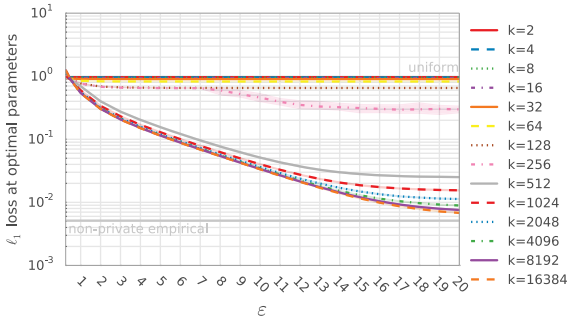


(e) O-RAPPOR varying h

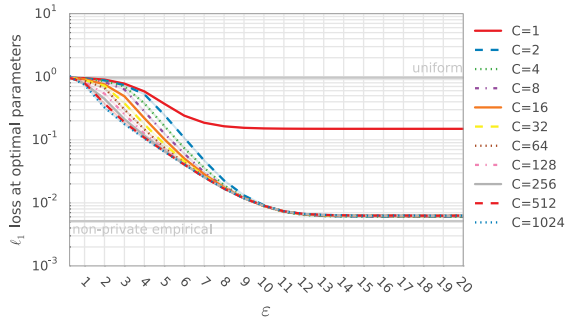
Figure 15: ℓ_1 loss when decoding open alphabets using O-RR and O-RAPPOR under various parameter settings, for $n = 10^6$ users with input drawn from an alphabet of $S = 4096$ symbols under a geometric distribution with mean=$S/5$. Remaining free parameters are set via grid search to minimize the median loss over 50 samples at the given ε and fixed parameter values. Lines show median ℓ_1 loss while the (narrow) shaded regions indicate 90% confidence intervals (over 50 samples for the optimal parameter settings.)



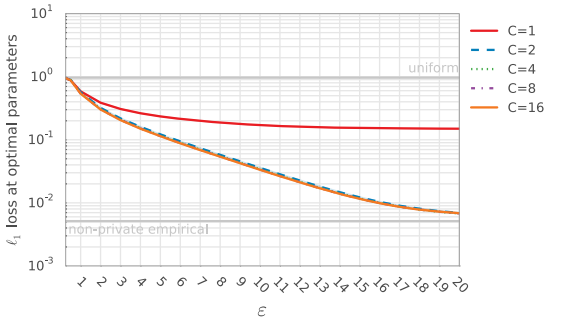
(a) O-RR varying k



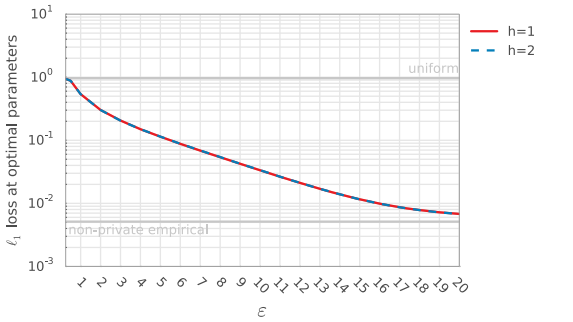
(b) O-RAPPOR varying k



(c) O-RR varying C



(d) O-RAPPOR varying C



(e) O-RAPPOR varying h

Figure 16: ℓ_1 loss when decoding open alphabets using O-RR and O-RAPPOR under various parameter settings, for $n = 10^8$ users with input drawn from an alphabet of $S = 4096$ symbols under a geometric distribution with mean=$S/5$. Remaining free parameters are set via grid search to minimize the median loss over 50 samples at the given ε and fixed parameter values. Lines show median ℓ_1 loss while the (narrow) shaded regions indicate 90% confidence intervals (over 50 samples for the optimal parameter settings.)