Mean Estimation with User-level Privacy under Data Heterogeneity

Rachel Cummings[*]
Columbia University

Vitaly Feldman
Apple

Audra McMillan
Apple

Kunal Talwar
Apple

Abstract

A key challenge for data analysis in the federated setting is that user data is heterogeneous, i.e., it cannot be assumed to be sampled from the same distribution. Further, in practice, different users may possess vastly different number of samples. In this work we propose a simple model of heterogeneous user data that differs in both distribution and quantity of data, and we provide a method for estimating the population-level mean while preserving user-level differential privacy. We demonstrate near asymptotic optimality of our estimator among nearly unbiased estimators. In particular, while the optimal non-private estimator can be shown to be linear, we show that privacy constrains us to use a non-linear estimator.

Large parts of machine learning deal with the problem of building one model that works for all users, based on multiple independent samples from some underlying distribution. In federated learning settings, this i.i.d. assumption is often violated. Users may differ in the distribution from which they sampling, as well as the number of samples they have. This has led to considerable renewed interest in various approaches to personalized models, many of which are compatible with privacy.

Our main contribution is a differentially private algorithm that estimates $p^\star$ and σ_p in this setting. We first study this question in an idealized setting with known σ_p and no privacy constraints. Here the optimal estimator for p_i is simple and linear: it is a weighted linear combination of the individual user means with weights that depend on the k_i's and on σ_p. The variance of this estimate is $\sigma^2_{ideal} \approx (\sum_i \min(k_i, \sigma_p^{-2}))^{-1}$. This expression has a natural interpretation: this is the variance from using $\min(k_i, \sigma_p^2)$ samples from user i and averaging all the Bernoulli samples thus obtained. The restriction on using at most σ_p^{-2} samples from any fixed users ensures that the estimator is not too affected by their personal mean p_i.

We provide a differentially private estimator for $p^\star$ with variance $O(\sigma^2_{ideal})$. Interestingly, the estimator achieving this bound in the private setting is non-linear. Further, we show that σ^2_{ideal} is near-optimal, under some mild technical conditions.

1 Model and Preliminaries

Let $\mathcal{D}$ be a distribution on $[0, 1]$ with (unknown) mean p and variance σ_p^2. We assume a population of $n \in \mathbb{N}$ users, where each user $i \in [n]$ has a hidden variable $p_i \sim \mathcal{D}$ and $k_i \in \mathbb{N}$ samples $x_i^1, \ldots, x_i^{k_i} \sim_{i.i.d.} Ber(p_i)$. That is, the samples of user i are i.i.d. from a Bernoulli distribution with parameter p_i, which we will denote $\mathcal{D}_i = \text{Ber}(p_i)$. Assume without loss and for ease of notation that individuals are sorted by their k_i, so that $k_1 \geq \cdots \geq k_n$. The samples x_i^j and hidden variables p_i of

[*]Part of this work was completed while the author was at Apple.

Preprint. Under review.

each user are unknown to the analyst. For simplicity, we will concentrate on the case where the k_i's are public. We defer to the full version the general case where the k_i's are also private.

The analyst's goal is to estimate the population mean p with an estimator of minimum variance in a manner that is differentially private with respect to user data (p_i and $\{x_i^j\}$). Each user provides their own estimate of their p_i to the analyst based on their data x_i: $\hat{p}_i = \frac{1}{k_i}\sum_{j=1}^{k_i} x_i^j$. The analyst can then aggregate these (possibly along with other information) into her estimate of p. Let $\mathcal{D}_{\text{Ber},k}$ be the distribution that first samples $p_i \sim \mathcal{D}$, then samples $x_1, \cdots, x_k \sim Ber(p_i)$ and finally outputs $\hat{p}_i = \frac{1}{k}\sum_{i=1}^{k} x_i$.

Differential privacy (DP) [1] informally limits the inferences that can be made about an individual as a result of computations on a large dataset containing their data. In our setting where users have heterogeneous quantities of data, we distinguish between *user-level* and *example-level* DP. The former considers changing all data points associated with a single user, whereas the latter considers changing only a single data point, regardless of the number of data points contributed by that user. In this work, we provide user-level DP guarantees.

Formally, let $D_i = \{x_i^1, \cdots, x_i^{k_i}\}$ be the data of user i for each $i \in [n]$. We say that two datasets $D = \{D_i\}_{i\in[n]}$ and $D' = \{D_i'\}_{i\in[n]}$ are *neighboring* if $|D_i| = |D_i'|$ for all $i \in [n]$, and there exists an index i such that for all $j \in [n]\backslash\{i\}$, $D_j = D_j'$. That is, the entire local dataset of a single user is changed.

Definition 1.1 (User-level (ϵ, δ)-Differential Privacy [1])**.** *Given $\epsilon \geq 0$, $\delta \in [0, 1]$ a randomized mechanism $\mathcal{M} : \mathcal{X}^{k_1} \times \cdots \times \mathcal{X}^{k_n} \to \mathcal{Y}$ is* (ϵ, δ)-differentially private *if for all neighboring datasets $D \sim D' \in \mathcal{X}^{k_1} \times \cdots \times \mathcal{X}^{k_n}$, and all events $E \subseteq \mathcal{Y}$, $\Pr[\mathcal{M}(D) \in E] \leq e^\epsilon \cdot \Pr[\mathcal{M}(D') \in E] + \delta$, where the probabilities are taken over the random coins of $\mathcal{M}$.*

2 A Non-Private Estimator

We begin by illustrating the procedure for computing an optimal estimator $\hat{p}$ in the non-private setting. The analyst will compute the population-level mean estimate $\hat{p}$ as a weighted linear combination of the user-level estimates $\hat{p}_i$.[2] Let σ_i^2 be the variance of $\hat{p}_i$. In an idealized setting where the σ_i^2 are all known, the following is an optimal and unbiased estimator [2]:

$$\hat{p}^{\text{ideal}} = \sum_{i=1}^{n} w_i^* \hat{p}_i \text{ where } w_i^* = \frac{1/\sigma_i^2}{\sum_{j=1}^{n} 1/\sigma_j^2}. \tag{1}$$

In practice, the σ_is are unknown, so the analyst must rely on estimates to assign weights. Fortunately, the user-level variance σ_i^2 can be expressed as a function of k_i and the population statistics p and σ_p^2: $\sigma_i^2 = \frac{1}{k_i}(p - p^2) + (1 - \frac{1}{k_i})\sigma_p^2$. Now, p and σ_p^2 are also unknown but since they are population statistics, we can use simple estimators to obtain initial estimates. These initial statistics can then be used to define the weights, resulting in a refined estimate of the mean p. Under some mild conditions on $\mathcal{D}$, and provided that n is large enough, the error incurred by $\text{var}(\hat{p}^{\text{realistic}}) \leq C \cdot \text{var}(\hat{p}^{\text{ideal}})$ for some constant C.[3]

3 A Framework for Private Estimators

We now turn to our main result, which is a framework for designing differentially private estimators of the mean p. As in Section 2 we will need an initial estimate of p and σ_p^2 in order to decide how to weight the contributions of the users. In the non-private setting, there are canonical, and optimal, choices of these estimators, the empirical mean and empirical variance. In the private setting, these choices are not canonical and the optimal estimators are setting dependent. There is a considerable literature exploring the performance of various mean and variance estimators for the homogeneous, single data point per user setting. As such, we leave the choice of the specific initial mean and

[2] In the non-private setting, this restriction is without loss since the optimal estimator takes this form. In the private setting this is still near-optimal; see Section 4 for more details.

[3] This can be observed by viewing the non-private setting as a simplified version of the setting studied in Section 4, which proves near-optimality of (truncated) linear estimators for this problem.

variance estimators as parameters of the framework. This allows us to focus on the nuances of the heterogeneous setting, not addressed in prior work.

We begin with a discussion of the ideal estimator $\hat{p}_\epsilon^{\text{ideal}}$ if the σ_i were known, which has two key differences to $\hat{p}^{\text{ideal}}$. The first main distinction is that Laplacian noise is added to achieve differential privacy. A natural solution would be to add noise directly to the non-private estimator $\hat{p}^{\text{ideal}}$, but the sensitivity of this statistic is too high. Thus, the first change we make is to limit the weight on any individual's contribution. We do this with simple truncation parameter T that trades-off the variance of the weighted sum of individual estimates (which is minimized by assigning high weight to low variance estimators) and variance of the noise added for privacy (which is minimized by assigning roughly equal weight to all users).

The second main modification is to lower the sensitivity of the weighted statistic. Inspired by the Gaussian mean estimator of [3], we truncate the individual contributions $\hat{p}_i$. The truncation intervals $[a_i, b_i]$ are chosen to be as small as possible (to reduce noise added for privacy), while simultaneously ensuring that $\hat{p}_i \in [a_i, b_i]$ with high probability (to avoid truncating relevant information for the estimation). In order to achieve this, we need a tail bound on the distribution $\mathcal{D}$. To maintain generality, we assume there exists a known function $f_D^k(n, \cdot, \beta)$ that describes high-probability concentration guarantees of $\hat{p}_i$ around p, and is defined in the following way: $\Pr\left(\forall i, |\hat{p}_i - p| \leq f_{\mathcal{D}}^{k_i}(n, \sigma_p^2, \beta)\right) \geq 1 - \beta$. We can now describe the ideal estimator $\hat{p}_\epsilon^{\text{ideal}}$:

$$\hat{p}_\epsilon^{\text{ideal}} = \textstyle\sum_{i=1}^n w_i [\hat{p}_i]_{a_i}^{b_i} + \text{Lap}(\frac{\max_i w_i |b_i - a_i|}{\epsilon}), \tag{2}$$

where $[\hat{p}_i]_{a_i}^{b_i}$ denotes the projection of $\hat{p}_i$ onto the interval $[a_i, b_i]$ and

$$a_i = p - f_{\mathcal{D}}^{k_i}(n, \sigma_p^2, \beta), \quad b_i = p + f_{\mathcal{D}}^{k_i}(n, \sigma_p^2, \beta), \quad \text{and} \quad \tilde{w_i}^* = \frac{\min\{1/\sigma_i^2, T^*/\sigma_i\}}{\sum_{j=1}^n \min\{1/\sigma_j^2, T^*/\sigma_j\}}. \tag{3}$$

The weight truncation parameter T is then chosen to minimise the variance. Since all the parameters are known, this is a simple optimisation problem.

Algorithm 1 Private Heterogeneous Mean Estimation $\hat{p}_\epsilon^{\text{realistic}}$

Input: (ϵ, δ)-DP mean estimator $\texttt{mean}_{\epsilon,\delta}$, (ϵ, δ)-DP variance estimator $\texttt{variance}_{\epsilon,\delta}$, num. of users n, num. of samples from each user $(k_1, \ldots, k_n)$ $(k_i \geq k_{i+1})$, user-level ests $(\hat{p}_1, \cdots, \hat{p}_n)$, $\alpha > 0$.

Initial Estimates

$\hat{p}^{\text{initial}} = \texttt{mean}_{\epsilon,\delta}(\mathrm{x}^1_{9\mathrm{n}/10}, \cdots, \mathrm{x}^1_{\mathrm{n}})$ ▷ Initial mean estimate

$\hat{\sigma}_p^2 = \texttt{variance}_{\epsilon,\delta}(\hat{p}_1, \cdots, \hat{p}_{\log n})$ ▷ Initial variance estimate

Defining weights and truncation

for $i = \log n$ to $9n/10$ **do**

  Compute $\hat{\sigma}_i^2 = \frac{1}{k_i}(\hat{p}^{\text{initial}} - (\hat{p}^{\text{initial}})^2) + (1 - \frac{1}{k_i})\hat{\sigma}_p^2$. ▷ Estimate individual variances

  $\hat{a_i} = \hat{p}_\epsilon^{\text{initial}} - \alpha - f_{\mathcal{D}}(n, \hat{\sigma_p^2}, \beta/2) - f_{\text{Bin}}(k_i, \hat{p} + \alpha + f_{\mathcal{D}}(n, \hat{\sigma_p}, \beta/2), \beta/n)$

  $\hat{b_i} = \hat{p}_\epsilon^{\text{initial}} + \alpha + f_{\mathcal{D}}(n, \hat{\sigma_p^2}, \beta/2) + f_{\text{Bin}}(k_i, \hat{p} + \alpha + f_{\mathcal{D}}(n, \hat{\sigma_p^2}, \beta/2), \beta/n)$.

  ▷ Estimate truncation parameters

Compute $\hat{T}^* = \arg\min_T \frac{\sum_{i=\log n+1}^{9n/10} \min\{\frac{1}{\hat{\sigma_i}^2}, T^2\} + \max_{\log n+1 \leq i \leq 9n/10} \frac{\min\{1/\hat{\sigma_i}^4, T^2/\hat{\sigma_i}^2\}|\hat{b_i} - \hat{a_i}|^2}{\epsilon^2}}{(\sum_{i=\log n+1}^{9n/10} \min\{1/\hat{\sigma_j^2}, T/\hat{\sigma_i}\})^2}$

  ▷ Compute weight truncation

for $i = \log n$ to $9n/10$ **do**

  $\hat{w_i}^* = \frac{\min\{1/\hat{\sigma_i^2}, \hat{T}^*/\hat{\sigma_i}\}}{\sum_{j=\log n+1}^{9n/10} \min\left\{1/\hat{\sigma_j^2}, \hat{T}^*/\hat{\sigma_i}\right\}}$ ▷ Compute weights

Final Estimate

$\Lambda = \max_{i \in [\log n, 9n/10]} \frac{\min\{1/\hat{\sigma_i^2}, \hat{T}^*/\hat{\sigma_i}\}|\hat{b_i} - \hat{a_i}|}{\sum_{j=\log n+1}^{9n/10} \min\left\{1/\hat{\sigma_j^2}, \hat{T}^*/\hat{\sigma_i}\right\}}$ ▷ Compute sensitivity

Sample $Y \sim \text{Lap}\left(\frac{\Lambda}{\epsilon}\right)$ ▷ Sample noise added for privacy

return $\hat{p}_\epsilon^{\text{realistic}} = \sum_{i=\log n+1}^{9n/10} \hat{w_i}^* [\hat{p}_i]_{\hat{a_i}}^{\hat{b_i}} + Y$ ▷ Final estimate

As in the non-private setting, in order to translate $\hat{p}_\epsilon^{\text{ideal}}$ into a realisable estimator we need to obtain estimates of p and σ_p^2. We will divide the individuals into three groups. The first group, consisting of the $n/10$ individuals with the lowest k_i will be used to compute the initial mean estimate $\hat{p}_\epsilon^{\text{initial}}$. The $\log n$ individuals with the largest k_i will be used to compute an the initial variance estimate $\hat{\sigma}_p^2$. These initial estimates will be plugged into expressions to compute $\hat{\sigma}_i^2$, $\hat{a}_i$, and $\hat{b}_i$ for the remaining individuals $\log n+1 \le i \le 9n/10$. Let $\texttt{mean}_{\epsilon,\delta}$ and $\texttt{variance}_{\epsilon,\delta}$ be the (ϵ,δ)-DP mean and variance estimators used to give the initial estimates.

The following theorem gives the properties of the estimators $\texttt{mean}_{\epsilon,\delta}$ and $\texttt{variance}_{\epsilon,\delta}$ necessary to ensure that the error of $\hat{p}_\epsilon^{\text{realistic}}$ is within a constant factor of the error of $\hat{p}_\epsilon^{\text{ideal}}$.

Theorem 3.1. *For any $\epsilon > 0$, $\delta \in [0,1]$, Algorithm 1 is (ϵ,δ)-DP. If with probability $1-\beta$,*

- $\texttt{mean}_{\epsilon,\delta}$ *is such that given $n/10$ samples from $\mathcal{D}$, $|p-\hat{p}| \le \alpha \le f_{\mathcal{D}}^{k_i}(n,\sigma_p^2,\beta)$ and $\hat{p}_\epsilon^{\text{initial}}(1-\hat{p}_\epsilon^{\text{initial}}) \in \left[\frac{1}{2}p(1-p), \frac{3}{2}p(1-p)\right]$,*
- $\texttt{variance}_{\epsilon,\delta}$ *is such that given $\log n$ samples from $\mathcal{D}_{Ber,k}$, $\hat{\sigma}_p^2 \in [\text{var}(\mathcal{D}_{Ber,k}),\ 8\cdot\text{var}(\mathcal{D}_{Ber,k})]$,*
- *$\mathcal{D}$ is s.t. $f_{\mathcal{D}}^{k_i}(n,\sigma_p^2,\beta) \ge \Omega(f_{\mathcal{D}}(n,\sigma_p^2,\beta) + f_{\text{Bin}}(k_i, \min\{1/2, p+f_{\mathcal{D}}(n,\sigma_p,\beta/2)\},\beta/n))$*
- *the distribution of the k_i is such that $\frac{k_1}{k_{n/2}} \le \frac{n/2-\log n}{\log n}$*

then with probability $1-2\beta$, $\text{var}(\hat{p}_\epsilon^{\text{realistic}}) \le C\cdot\text{var}(\hat{p}_\epsilon^{\text{ideal}})$ for some absolute constant C.

In the full version of this paper we give an example instantiation of Algorithm 1 with particular $\texttt{mean}_{\epsilon,\delta}$ and $\texttt{variance}_{\epsilon,\delta}$, and we show that provided $\mathcal{D}$ and $\mathcal{D}_{\text{Ber},k}$ are sufficiently well-behaved, this instantiation meets the requirements of Theorem 3.1.

4 Optimality of $\hat{p}_\epsilon^{\text{realistic}}$

In this section we show that the estimator $\hat{p}_\epsilon^{\text{realistic}}$ discussed in Section 3 is minimax optimal up to logarithmic factors. Under the assumptions of Theorem 3.1, it is sufficient to show that the estimator $\hat{p}_\epsilon^{\text{ideal}}$, defined by Eqns (2) and (3) is minimax optimal up to logarithmic factors. The following class will act as an intermediary in our proof that the estimator $\hat{p}_\epsilon^{\text{ideal}}$ is optimal up to logarithmic factors:

$$\texttt{NLE} = \{M_{\texttt{NL}}(x_1,\cdots,x_n;\mathbf{w}) = \textstyle\sum_{i=1}^n w_i x_i + \text{Lap}\left(\frac{\max_i w_i\sigma_i}{\epsilon}\right) \mid w_1,\cdots,w_n \in [0,1], \sum_{i=1}^n w_i = 1\}.$$

Similar to $\hat{p}_\epsilon^{\text{ideal}}$, this class of estimators is not realizable since we only have access to an estimate of $\sigma_i = \text{var}(\mathcal{D}_p(k_i))$. The estimators in NLE are also not ϵ-DP (unless $\sigma_i = 1$ for all $i \in [n]$). Firstly, we will argue that the variance of $\hat{p}_\epsilon^{\text{ideal}}$ is only a polylog factor larger than the variance of the optimal estimator in NLE. Finally, we'll show that under some mild conditions, the optimal unbiased estimator lies in NLE.

Lemma 4.1. $\text{var}(\hat{p}_\epsilon^{\text{ideal}}) = \tilde{\Theta}(\inf_{M\in\texttt{NLE}}\text{var}(M))$.

The main component of this proof is showing that the weights given in Eqn 3 are optimal for NLE. Let us turn to final component of our optimality proof.

Lemma 4.2. *Let $\mathcal{P}$ be a parameterized family of distributions $p \mapsto \mathcal{D}_p$ and suppose that $M : [0,1]^n \to [0,1]$ is an estimator such that for all $p \in [0,1]$, (1) M is almost unbiased, $\mathbb{E}_{\forall i\in[n],x_i\sim\mathcal{D}_p(k_i),M}[M(x_1,\cdots,x_n)] \in [p-\alpha, p+\alpha]$, and (2) the Fisher information of ϕ_{p,k_i} is inversely proportional to the variance $\text{var}(\mathcal{D}_p(k_i))$, $\int\left(\frac{\partial}{\partial p}\log\phi_{p,k_i}(x_i)\right)^2\phi_{p,k_i}(x_i)dx_i = O\left(\frac{1}{\text{var}(\mathcal{D}_p(k_i))}\right)$, then there exists an estimator in the class NLE such that $\max_{p\in[1/3,2/3]}[\text{var}_{\mathcal{D}_p}(M_{\texttt{NL}})] \le O\left(\max_{p\in[1/3,2/3]}[\text{var}_{\mathcal{D}_p}(M)]\right)$ where the constant depends on α.*

Finally, combining Theorem 3.1, Lemma 4.1, and Lemma 4.2 we can conclude that under some mild conditions, and assuming the existence of sufficiently nice estimators $\texttt{mean}_{\epsilon,\delta}$ and $\texttt{variance}_{\epsilon,\delta}$, $\hat{p}_\epsilon^{\text{realistic}}$ is an optimal nearly unbiased estimator up to logarithmic factors.

References

[1] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. volume Vol. 3876, pages 265–284, 01 2006.

[2] J. Hartung, G. Knapp, and B. Sinha. Statistical meta-analysis with applications. 08 2008.

[3] V. Karwa and S. P. Vadhan. Finite sample differentially private confidence intervals. volume abs/1711.03908 of *Innovations in Theoretical Computer Science '18*, 2018.