A New Analysis of Differential Privacy's Generalization Guarantees

Christopher Jung
University of Pennsylvania, Philadelphia, PA, USA
chrjung@seas.upenn.edu

Katrina Ligett
The Hebrew University, Jerusalem, Israel
katrina@cs.huji.ac.il

Seth Neel
University of Pennsylvania, Philadelphia, PA, USA
sethneel@wharton.upenn.edu

Aaron Roth
University of Pennsylvania, Philadelphia, PA, USA
aaroth@cis.upenn.edu

Saeed Sharifi-Malvajerdi
University of Pennsylvania, Philadelphia, PA, USA
saeedsh@wharton.upenn.edu

Moshe Shenfeld
The Hebrew University, Jerusalem, Israel
moshe.shenfeld@mail.huji.ac.il

Abstract

We give a new proof of the "transfer theorem" underlying adaptive data analysis: that any mechanism for answering adaptively chosen statistical queries that is differentially private and sample-accurate is also accurate out-of-sample. Our new proof is elementary and gives structural insights that we expect will be useful elsewhere. We show: 1) that differential privacy ensures that the expectation of any query on the *conditional distribution* on datasets induced by the transcript of the interaction is close to its expectation on the data distribution, and 2) sample accuracy on its own ensures that any query answer produced by the mechanism is close to the expectation of the query on the conditional distribution. This second claim follows from a thought experiment in which we imagine that the dataset is resampled from the conditional distribution after the mechanism has committed to its answers. The transfer theorem then follows by summing these two bounds, and in particular, avoids the "monitor argument" used to derive high probability bounds in prior work.

An upshot of our new proof technique is that the concrete bounds we obtain are substantially better than the best previously known bounds, even though the improvements are in the constants, rather than the asymptotics (which are known to be tight). As we show, our new bounds outperform the naive "sample-splitting" baseline at dramatically smaller dataset sizes compared to the previous state of the art, bringing techniques from this literature closer to practicality.

2012 ACM Subject Classification Theory of computation → Sample complexity and generalization bounds

Keywords and phrases Differential Privacy, Adaptive Data Analysis, Transfer Theorem

Digital Object Identifier 10.4230/LIPIcs.ITCS.2020.31

Related Version arXiv version available at `https://arxiv.org/abs/1909.03577`.

Funding *Christopher Jung*: Supported in part by NSF grant AF-1763307.
Katrina Ligett: Supported in part by Israel Science Foundation (ISF) grant #1044/16, the United States Air Force and DARPA under contracts FA8750-16-C-0022 and FA8750-19-2-0222, and the Federmann Cyber Security Center in conjunction with the Israel national cyber directorate.

Seth Neel: Supported in part by an NSF Graduate Research Fellowship.
Aaron Roth: Supported in part by NSF grant AF-1763314, the United States Air Force and DARPA under Contract No FA8750-16-C-0022, and a grant from the Sloan Foundation.
Moshe Shenfeld: Supported in part by Israel Science Foundation (ISF) grant #1044/16, the United States Air Force and DARPA under contracts FA8750-16-C-0022 and FA8750-19-2-0222, and the Federmann Cyber Security Center in conjunction with the Israel national cyber directorate. Any opinions,findings and conclusions or recommendations expressed in this materialare those of the author(s) and do not necessarily reflect the views ofthe United States Air Force and DARPA.

Acknowledgements We thank Adam Smith for helpful conversations at an early stage of this work, and Daniel Roy for helpful feedback on the presentation of the result.

1 Introduction

Many data analysis pipelines are *adaptive*: the choice of which analysis to run next depends on the outcome of previous analyses. Common examples include variable selection for regression problems and hyper-parameter optimization in large-scale machine learning problems: in both cases, common practice involves repeatedly evaluating a series of models on the same dataset. Unfortunately, this kind of adaptive re-use of data invalidates many traditional methods of avoiding over-fitting and false discovery, and has been blamed in part for the recent flood of non-reproducible findings in the empirical sciences [14].

There is a simple way around this problem: don't re-use data. This idea suggests a baseline called *data splitting*: to perform k analyses on a dataset, randomly partition the dataset into k disjoint parts, and perform each analysis on a fresh part. The standard "holdout method" is the special case of $k = 2$. Unfortunately, this natural baseline makes poor use of data: in particular, the data requirements of this method grow *linearly* with the number of analyses k to be performed.

A recent literature starting with Dwork et al. [6] shows how to give a significant asymptotic improvement over this baseline via a connection to differential privacy: rather than computing and reporting exact sample quantities, perturb these quantities with noise. This line of work established a powerful *transfer theorem*, that informally says that any analysis that is simultaneously differentially private and accurate *in-sample* will also be accurate *out-of-sample*. The best analysis of this technique shows that for a broad class of analyses and a target accuracy goal, the data requirements grow only with $\sqrt{k}$ – a quadratic improvement over the baseline [1]. Moreover, it is known that in the worst case, this cannot be improved asymptotically [15, 23]. Unfortunately, thus far this literature has had little impact on practice. One major reason for this is that although the more sophisticated techniques from this literature give asymptotic improvements over the sample-splitting baseline, the concrete bounds do not actually improve on the baseline until the dataset is enormous. This remains true even after optimizing the constants that arise from the arguments of Dwork et al. [6] or Bassily et al. [1], and appears to be a fundamental limitation of their proof techniques [20]. In this paper, we give a new proof of the transfer theorem connecting differential privacy and in-sample accuracy to out-of-sample accuracy. Our proof is based on a simple insight that arises from imagining a "resampling" experiment, and in particular yields an improved concrete bound that beats the sample-splitting baseline at dramatically smaller data set sizes n compared to prior work. In fact, at reasonable dataset sizes, the magnitude of the improvement arising from our new theorem is significantly larger than the improvement between the bounds of Bassily et al. [1] and Dwork et al. [6]: see Figure 1.

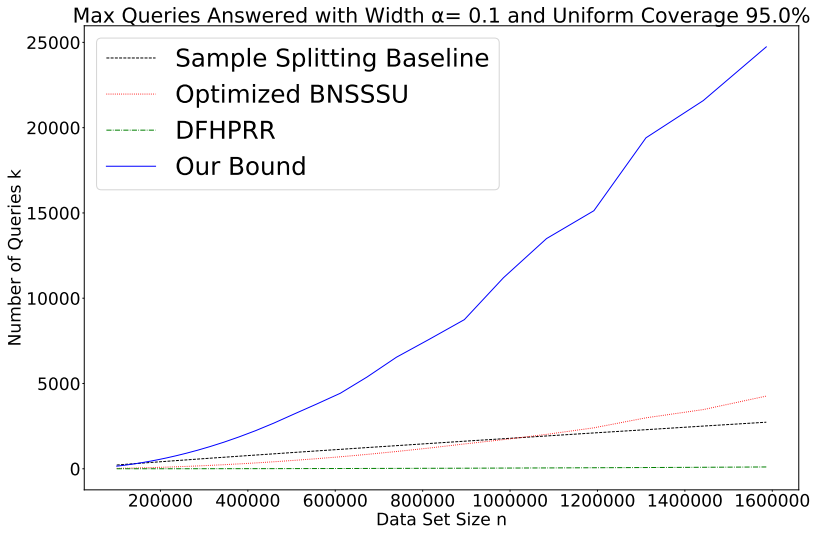

Figure 1 A comparison of the number of adaptive linear queries that can be answered using the Gaussian mechanism as analyzed by our transfer theorem (Theorem 9), the numerically optimized variant of the bound from Bassily et al. [1] (Optimized BNSSSU) as derived in [20], and the original transfer theorem from Dwork et al. [6] (DFHPRR). We plot for each dataset size n, the number of queries k that can be answered while guaranteeing confidence intervals around the answer that have width $\alpha = 0.1$ and uniform coverage probability $1 - \beta = 0.95$. We compare with the naive sample splitting baseline that simply splits the dataset into k pieces and answers each query with the empirical answer on a fresh piece.

1.1 Proof Techniques

Prior Work

Consider an unknown data distribution $\mathcal{P}$ over a data-domain $\mathcal{X}$, and a dataset $S \sim \mathcal{P}^n$ consisting of n i.i.d. draws from $\mathcal{P}$. It is a folklore observation (attributed to Frank McSherry) that if a predicate $q : \mathcal{X} \to [0, 1]$ is selected by an ϵ-differentially private algorithm M acting on S, then it will generalize *in expectation* (or *have low bias*) in the sense that $|\mathbb{E}_{q \sim M(S)}[\mathbb{E}_{x \sim \mathcal{P}}[q(x)] - \frac{1}{n}\sum_{x \in S} q(x)]| \approx \epsilon$. But bounds on bias are not enough to yield tight confidence intervals, and so prior work has focused on strengthening the above observation into a high probability bound. For small ϵ, the optimal bound has the asymptotic form: $\Pr_{q \sim M(S)}[|\mathbb{E}_{x \sim \mathcal{P}}[q(x)] - \frac{1}{n}\sum_{x \in S} q(x)| \geq \epsilon] \leq e^{-O(\epsilon^2 n)}$ [1]. Note that this bound does not refer to the *estimated answers* supplied to the data analyst: it says only that a differentially private data analyst is unlikely to be able to find a query whose average value on the dataset differs substantially from its expectation. Pairing this with a simultaneous high probability bound on the *in-sample accuracy* of a mechanism – that it supplies answers a such that with high probability the empirical error is small: $\Pr_{a \sim M(S)}[|a - \frac{1}{n}\sum_{x \in S} q(x)| \geq \alpha] \leq \beta$ – yields a bound on out-of-sample accuracy via the triangle inequality.

Dwork et al. [6] proved their high probability bound via a direct computation on the *moments* of empirical query values, but this technique was unable to achieve the optimal rate. Bassily et al. [1] proved a bound with the optimal rate by introducing the ingenious *monitor technique*. This important technique has subsequently found other uses [24, 19, 13], but is a heavy hammer that seems unavoidably to yield large constant overhead, even after numeric optimization [20].

Our Approach

We take a fundamentally different approach by directly providing high probability bounds on the out-of-sample accuracy $|a - \mathbb{E}_{x\sim\mathcal{P}}[q(x)]|$ of mechanisms that are both differentially private and accurate in-sample. Our elementary approach is motivated by the following thought experiment: in actuality, the dataset S is fixed before any interaction with M begins. However, imagine that after the entire interaction with M is complete, the dataset S is *resampled* from the conditional distribution $\mathcal{Q}$ on datasets *conditioned* on the output of M. This thought experiment doesn't alter the joint distribution on datasets and outputs, and so any in-sample accuracy guarantees that M has continue to hold under this hypothetical re-sampling experiment. But because the empirical value of the queries on the re-sampled dataset are likely to be close to their expected value over the conditional distribution $\mathcal{Q}$, the only way the mechanism can promise to be sample-accurate with high probability is if it provides answers that are *close to their expected value over the conditional distribution with high probability.*

This focuses attention on the conditional distribution on datasets induced by differentially private transcripts. But it is not hard to show that a consequence of differential privacy is that the conditional expectation of any query must be close to its expectation over the data distribution with high probability. In contrast to prior work, this argument directly leverages high-probability in-sample accuracy guarantees of a private mechanism to derive high-probability out-of-sample guarantees, without the need for additional machinery like the monitor argument of Bassily et al. [1].

1.2 Further Related Work

The study of "adaptive data analysis" was initiated by Dwork et al. [6, 5] who provided upper bounds via a connection to differential privacy, and Hardt and Ullman [15] who provided lower bounds via a connection to fingerprinting codes. The upper bounds were subsequently strengthened by Bassily et al. [1], and the lower bounds by Steinke and Ullman [23] to be (essentially) matching, asymptotically. The upper bounds were optimized by Rogers et al. [20], which we use in our comparisons. Subsequent work proved transfer theorems related to other quantities like description length bounds [4] and compression schemes [3], and expanded the types of analyses whose generalization properties we could reason about via a connection to a quantity called approximate max information [4, 21]. Feldman and Steinke [11, 12] give improved methods that could guarantee out-of-sample accuracy bounds that depended on query variance. Neel and Roth [17] extend the transfer theorems from this literature to the related problem of adaptive data gathering, which was identified by Nie et al. [18]. Ligett and Shenfeld [16] give an algorithmic stability notion they call *local statistical stability* (also defined with respect to a conditional data distribution) that they show asymptotically characterizes the ability of mechanisms to offer high probability out-of-sample generalization guarantees for linear queries. A related line of work initiated by Russo and Zou [22] and extended by Xu and Raginsky [25] starts with weaker assumptions on the mechanism (mutual information bounds), and derives weaker conclusions (bounds on bias, rather than high probability generalization guarantees).

A more recent line of work aims at mitigating the fact that the worst-case bounds deriving from transfer theorems do not give non-trivial guarantees on reasonably sized datasets. Zrnic and Hardt [26] show that better bounds can be derived under the assumption that the data analyst is restricted in various ways to not be fully adaptive. Feldman et al. [10] show that overfitting by a classifier because of test-set re-use is mitigated in multi-label prediction

problems, compared to binary prediction problems. Rogers et al. [20] give a method for certifying the correctness of heuristically guessed confidence intervals, which they show often out-perform the theoretical guarantees by orders of magnitude.

Finally, Elder [9, 8] proposes a Bayesian reformulation of the adaptive data analysis problem. In the model of [9], the data distribution $\mathcal{P}$ is assumed to itself be drawn from a prior that is commonly known to the data analyst and mechanism. In contrast, we work in the standard adversarial setting originally introduced by Dwork et al. [6] in which the mechanism must offer guarantees for worst case data distributions and analysts, and focus our attention on conditional distributions purely as a proof technique.

2 Preliminaries

Let $\mathcal{X}$ be an abstract data domain, and let $\mathcal{P}$ be an arbitrary distribution over $\mathcal{X}$. A dataset of size n is a collection of n data records: $S = \{S_i\}_{i=1}^n \in \mathcal{X}^n$. We study datasets sampled *i.i.d.* from $\mathcal{P}$: $S \sim \mathcal{P}^n$. We will write S to denote the random variable and $\boldsymbol{x}$ for realizations of this random variable. A linear query is a function $q : \mathcal{X}^* \to [0, 1]$ that takes the following empirical average form when acting on a data set $S \in \mathcal{X}^n$:

$$q(S) = \frac{1}{n}\sum_{i=1}^{n} q(S_i).$$

We will be interested in estimating the expectations of linear queries over $\mathcal{P}$. Abusing notation, given a distribution $\mathcal{D}$ over datasets, we write $q(\mathcal{D})$ to denote the expectation of q over datasets drawn from $\mathcal{D}$, and write $S_i \sim S$ to denote a datapoint sampled uniformly at random from a dataset S. Note that for linear queries we have:

$$q(\mathcal{D}) = \mathop{\mathbb{E}}_{S\sim\mathcal{D}}[q(S)] = \mathop{\mathbb{E}}_{S\sim\mathcal{D},S_i\sim S}[q(S_i)]$$

We note that for linear queries, when the dataset distribution $\mathcal{D} = \mathcal{P}^n$, we have $q(\mathcal{P}^n) = \mathbb{E}_{x\sim\mathcal{P}}[q(x)]$, which we write as $q(\mathcal{P})$ when the notation is clear from context. However, the more general definition will be useful because we will need to evaluate the expectation of q over other (non-product) distributions over datasets in our arguments, and we will generalize beyond linear queries in Appendices A.1 and A.2.

Given a family of queries Q, a statistical estimator is a (possibly stateful) randomized algorithm $M : \mathcal{X}^n \times Q^* \to \mathbb{R}^*$ parameterized by a dataset S that interactively takes as input a stream of queries $q_i \in Q$, and provides answers $a_i \in \mathbb{R}$. An analyst is an arbitrary randomized algorithm $\mathcal{A} : \mathbb{R}^* \to Q^*$ that generates a stream of queries and receives a stream of answers (which can inform the next queries it generates). When an analyst interacts with a statistical estimator, they generate a transcript of their interaction $\pi \in \mathbf{\Pi}$ where $\mathbf{\Pi} = (Q \times \mathbb{R})^*$ is the space of all transcripts. Throughout we write Π to denote the transcript's random variable and π for its realizations.

The interaction is summarized in Algorithm 1, and we write $\text{Interact}(M, \mathcal{A}; S)$ to refer to it. When M and $\mathcal{A}$ are clear from context, we will abbreviate this notation and write simply $I(S)$. When we refer to an indexed query q_j, this is implicitly a function of the transcript π. Given a transcript $\pi \in \mathbf{\Pi}$, write $\mathcal{Q}_\pi$ to denote the conditional distribution on datasets conditional on $\Pi = \pi$: $\mathcal{Q}_\pi = (\mathcal{P}^n)|\text{Interact}(M, \mathcal{A}; S) = \pi$. Note that $\mathcal{Q}_\pi$ will no longer generally be a product distribution. We will be interested in evaluating uniform accuracy bounds, which control the worst-case error over all queries:

Algorithm 1 Interact$(M, \mathcal{A}; S)$: An Analyst Interacting with a Statistical Estimator to Generate a Transcript.

Input: A statistical estimator M, an analyst $\mathcal{A}$, and a dataset $S \in \mathcal{X}^n$.
for $t = 1$ *to* k **do**
  The analyst generates a query $q_t \leftarrow \mathcal{A}(a_1, \ldots, a_{t-1})$ and sends it to the statistical estimator;
  The statistical estimator generates an answer $a_t \leftarrow M(S; q_t)$;
return $\Pi = ((q_1, a_1), \ldots, (q_k, a_k))$.

▶ **Definition 1** (Accuracy). *M satisfies (α, β)-sample accuracy if for every data analyst $\mathcal{A}$ and every data distribution $\mathcal{P}$,*

$$\Pr_{S\sim\mathcal{P}^n, \Pi\sim Interact(M,\mathcal{A};S)}[\max_j |q_j(S) - a_j| \geq \alpha] \leq \beta.$$

We say M satisfies (α, β)-distributional accuracy if for every data analyst $\mathcal{A}$ and every data distribution $\mathcal{P}$,

$$\Pr_{S\sim\mathcal{P}^n, \Pi\sim Interact(M,\mathcal{A};S)}[\max_j |q_j(\mathcal{P}^n) - a_j| \geq \alpha] \leq \beta.$$

We will be interested in interactions I that satisfy *differential privacy.*

▶ **Definition 2** (Differential Privacy [7]). *Two datasets $S, S' \in \mathcal{X}^n$ are neighbors if they differ in at most one coordinate. An interaction Interact$(M, \cdot; \cdot)$ satisfies (ϵ, δ)-differential privacy if for all data analysts $\mathcal{A}$, pairs of neighboring datasets $S, S' \in \mathcal{X}^n$, and for all events $E \subseteq \mathbf{\Pi}$:*

$$\Pr_{\Pi\sim Interact(M,\mathcal{A};S)}[\Pi \in E] \leq e^\epsilon \cdot \Pr_{\Pi\sim Interact(M,\mathcal{A};S')}[\Pi \in E] + \delta.$$

If Interact$(M, \cdot; \cdot)$ satisfies (ϵ, δ)-differential privacy, we will also say that M satisfies (ϵ, δ)-differential privacy.

We introduce a novel quantity that will be crucial to our argument: it captures the effect of the transcript on the change in the expectation of a query contained in the transcript.

▶ **Definition 3.** *An interaction Interact$(M, \mathcal{A}; \cdot)$ is called (ϵ, δ)-posterior stable if for every data distribution $\mathcal{P}$:*

$$\Pr_{S\sim\mathcal{P}^n, \Pi\sim Interact(M,\mathcal{A};S)}[\max_j |q_j(\mathcal{P}^n) - q_j(\mathcal{Q}_\Pi)| \geq \epsilon] \leq \delta.$$

3 An Elementary Proof of the Transfer Theorem

3.1 A General Transfer Theorem

In this section we prove a general transfer theorem for sample accurate mechanisms with low posterior stability. In Section 3.2 we prove that differentially private mechanisms have low posterior stability.

▶ **Theorem 4** (General Transfer Theorem). *Suppose that Interact$(M, \mathcal{A}; \cdot)$ is an (α, β)-sample accurate, (ϵ, δ)-posterior stable interaction. Then for every $c > 0$ it also satisfies:*

$$\Pr_{S\sim\mathcal{P}^n, \Pi\sim Interact(M,\mathcal{A};S)}[\max_j |a_j - q_j(\mathcal{P})| > \alpha + c + \epsilon] \leq \frac{\beta}{c} + \delta$$

i.e. it is (α', β')-distributionally accurate for $\alpha' = \alpha + c + \epsilon$ and $\beta' = \frac{\beta}{c} + \delta$.

The theorem follows easily from a change in perspective driven by an elementary observation. Imagine that *after* the interaction is run and results in a transcript π, the dataset S is *resampled* from its conditional distribution $\mathcal{Q}_\pi$. This does not change the joint distribution on datasets and transcripts. This simple claim is formalized below: its elementary proof appears in Appendix B.

▶ **Lemma 5** (Resampling Lemma). *Let $E \subseteq \mathcal{X}^n \times \mathbf{\Pi}$ be any event. Then:*

$$\Pr_{S\sim\mathcal{P}^n,\Pi\sim I(S)}[(S,\Pi)\in E] = \Pr_{S\sim\mathcal{P}^n,\Pi\sim I(S),S'\sim\mathcal{Q}_\Pi}[(S',\Pi)\in E]$$

The change in perspective suggested by the resampling lemma makes it easy to see why the following must be true: any sample-accurate mechanism must in fact be accurate with respect to the conditional distribution it induces. This is because if it can first commit to answers, and guarantee that they are sample-accurate *after* the dataset is resampled from the conditional, the answers it committed to must have been close to the conditional means, because it is likely that the empirical answers on the resampled dataset will be. This argument is generic and does not use differential privacy.

▶ **Lemma 6.** *Suppose that M is (α,β)-sample accurate. Then for every $c>0$ it also satisfies:*

$$\Pr_{S\sim\mathcal{P}^n,\,\Pi\sim Interact(M,\mathcal{A};S)}[\max_j |a_j - q_j(\mathcal{Q}_\Pi)| > \alpha + c] \le \frac{\beta}{c}$$

Proof. Denote by $j^*(\pi) = \arg\max_j |a_j - q_j(\mathcal{Q}_\pi)|$. Given $\alpha \ge 0$ and $c>0$, and expanding the definition of $q_{j^*(\Pi)}(\mathcal{Q}_\Pi)$ we get:

$$\begin{aligned}
&\Pr_{S\sim\mathcal{P}^n,\Pi\sim I(S)}\left[a_{j^*(\Pi)} - q_{j^*(\Pi)}(\mathcal{Q}_\Pi) > \alpha + c\right] \\
&\quad= \Pr_{S\sim\mathcal{P}^n,\Pi\sim I(S)}\left[\mathop{\mathbb{E}}_{S'\sim\mathcal{Q}_\Pi}\left[a_{j^*(\Pi)} - q_{j^*(\Pi)}(S') - \alpha\right] > c\right] \\
&\quad\le \Pr_{S\sim\mathcal{P}^n,\Pi\sim I(S)}\left[\mathop{\mathbb{E}}_{S'\sim\mathcal{Q}_\Pi}\left[\max\left\{a_{j^*(\Pi)} - q_{j^*(\Pi)}(S') - \alpha, 0\right\}\right] > c\right] \\
&\quad\overset{(1)}{\le} \frac{1}{c}\mathop{\mathbb{E}}_{S\sim\mathcal{P}^n,\Pi\sim I(S)}\left[\mathop{\mathbb{E}}_{S'\sim\mathcal{Q}_\Pi}\left[\max\left\{a_{j^*(\Pi)} - q_{j^*(\Pi)}(S') - \alpha, 0\right\}\right]\right] \\
&\quad\overset{(2)}{\le} \frac{1}{c}\mathop{\mathbb{E}}_{S\sim\mathcal{P}^n,\Pi\sim I(S)}\left[\Pr_{S'\sim\mathcal{Q}_\Pi}\left[a_{j^*(\Pi)} - q_{j^*(\Pi)}(S') - \alpha > 0\right]\right] \\
&\quad= \frac{1}{c}\Pr_{S\sim\mathcal{P}^n,\Pi\sim I(S),S'\sim\mathcal{Q}_\Pi}\left[a_{j^*(\Pi)} - q_{j^*(\Pi)}(S') > \alpha\right] \\
&\quad\overset{(3)}{=} \frac{1}{c}\Pr_{S\sim\mathcal{P}^n,\Pi\sim I(S)}\left[a_{j^*(\Pi)} - q_{j^*(\Pi)}(S) > \alpha\right]
\end{aligned}$$

Here, inequality (1) follows from Markov's inequality, inequality (2) follows from the fact that $a_{j^*(\Pi)} - q_{j^*(\Pi)}(S') - \alpha \le 1$, and equality 3 follows from the Resampling Lemma (Lemma 5). Repeating this argument for $q_{j^*(\Pi)}(Q_\Pi) - a_{j^*(\Pi)}$ yields a symmetric bound, so by combining the two with the guarantee of (α,β)-sample accuracy we get,

$$\begin{aligned}
\Pr_{S\sim\mathcal{P}^n,\Pi\sim I(S)}\left[\left|a_{j^*(\Pi)} - q_{j^*(\Pi)}(Q_\Pi)\right| > \alpha + c\right] &\le \frac{1}{c}\Pr_{S\sim\mathcal{P}^n,\Pi\sim I(S)}\left[\left|a_{j^*(\Pi)} - q_{j^*(\Pi)}(S)\right| > \alpha\right] \\
&\le \frac{\beta}{c}
\end{aligned}$$

◀

Because sample accuracy implies accuracy with respect to the conditional distribution, together with a bound on posterior stability, the transfer theorem follows immediately:

Proof of Theorem 4. By the triangle inequality:

$$\max_j |a_j - q_j(\mathcal{P})| \leq \max_i |a_i - q_i(\mathcal{Q}_\Pi)| + \max_l |q_l(\mathcal{Q}_\Pi) - q_l(\mathcal{P})|.$$

Lemma 6 bounds the first term by $\alpha + c$ with probability $1 - \frac{\beta}{c}$ over Π, and the definition of posterior stability bounds the second term by ϵ with probability $1 - \delta$ over Π, which concludes the proof. ◀

3.2 A Transfer Theorem for Differential Privacy

In this section we prove a transfer theorem for differentially private mechanisms by demonstrating that they have low posterior stability and applying our general transfer theorem.

We here show that differentially private mechanisms have low posterior stability for linear queries. In the Appendix we extend this argument to low-sensitivity and optimization queries.

▶ **Lemma 7.** *If M is (ϵ, δ)-differentially private, then for any data distribution $\mathcal{P}$, any analyst $\mathcal{A}$, and any constant $c > 0$:*

$$\Pr_{S \sim \mathcal{P}^n, \Pi \sim Interact(M, \mathcal{A}; S)} \left[\max_j |q_j(\mathcal{Q}_\Pi) - q_j(\mathcal{P})| > (e^\epsilon - 1) + 2c \right] \leq \frac{\delta}{c}$$

i.e. it is (ϵ', δ')-posterior stable for every data analyst $\mathcal{A}$, where $\epsilon' = e^\epsilon - 1 + 2c$ and $\delta' = \frac{\delta}{c}$.

Proof. Given a transcript $\pi \in \mathbf{\Pi}$, let $j^*(\pi) \in \arg\max_j |q_j(\mathcal{Q}_\pi) - q_j(\mathcal{P})|$. Define for an $\alpha > 0$:

$$\mathbf{\Pi}_\alpha = \left\{ \pi \in \mathbf{\Pi} | \, q_{j^*(\pi)}(\mathcal{Q}_\pi) - q_{j^*(\pi)}(\mathcal{P}) > \alpha \right\}$$

$$\mathcal{X}^+(\pi) = \left\{ x \in \mathcal{X} | \, \Pr_{S \sim \mathcal{Q}_\pi, S_i \sim S} [S_i = x] > \Pr_{S_i \sim \mathcal{P}} [S_i = x] \right\}$$

$$B^+_\alpha = \bigcup_{\pi \in \mathbf{\Pi}_\alpha} \left(\mathcal{X}^+(\pi) \times \{\pi\} \right)$$

$$\mathbf{\Pi}^+_\alpha(x) = \left\{ \pi \in \mathbf{\Pi} | \, (x, \pi) \in B^+_\alpha \right\}$$

Fix any α. Suppose that $\Pr\left[\left| q_{j^*(\Pi)}(\mathcal{Q}_\Pi) - q_{j^*(\Pi)}(\mathcal{P}) \right| > \alpha \right] > \frac{\delta}{c}$. We must have that either $\Pr\left[q_{j^*(\Pi)}(\mathcal{Q}_\Pi) - q_{j^*(\Pi)}(\mathcal{P}) > \alpha \right] > \frac{\delta}{2c}$ or $\Pr\left[q_{j^*(\Pi)}(\mathcal{P}) - q_{j^*(\Pi)}(\mathcal{Q}_\Pi) > \alpha \right] > \frac{\delta}{2c}$. Without loss of generality, assume

$$\Pr\left[q_{j^*(\Pi)}(\mathcal{Q}_\Pi) - q_{j^*(\Pi)}(\mathcal{P}) > \alpha \right] = \Pr\left[\Pi \in \mathbf{\Pi}_\alpha \right] > \frac{\delta}{2c} \tag{1}$$

Let S_i be the random variable obtained by first sampling $S \sim \mathcal{P}^n$ and then sampling $S_i \in S$ uniformly at random. We compare the probability measure of B^+_α under the joint distribution on S_i and Π with its corresponding measure under the product distribution of S_i and Π:

$$
\begin{aligned}
&\Pr_{(S_i,\Pi)}\left[(S_i,\Pi)\in B_\alpha^+\right]-\Pr_{S_i\otimes\Pi}\left[(S_i,\Pi)\in B_\alpha^+\right]\\
&\quad=\sum_{\pi\in\mathbf{\Pi}_\alpha}\Pr[\Pi=\pi]\sum_{x\in\mathcal{X}^+(\pi)}\left(\Pr[S_i=x|\Pi=\pi]-\Pr[S_i=x]\right)\\
&\quad\geq\sum_{\pi\in\mathbf{\Pi}_\alpha}\Pr[\Pi=\pi]\sum_{x\in\mathcal{X}^+(\pi)}q_{j^*(\pi)}(x)\left(\Pr[S_i=x|\Pi=\pi]-\Pr[S_i=x]\right)\\
&\quad\geq\sum_{\pi\in\mathbf{\Pi}_\alpha}\Pr[\Pi=\pi]\sum_{x\in\mathcal{X}}q_{j^*(\pi)}(x)\left(\Pr[S_i=x|\Pi=\pi]-\Pr[S_i=x]\right)\\
&\quad=\sum_{\pi\in\mathbf{\Pi}_\alpha}\Pr[\Pi=\pi]\left(q_{j^*(\pi)}(\mathcal{Q}_\pi)-q_{j^*(\pi)}(\mathcal{P})\right)\\
&\quad>\alpha\cdot\Pr\left[\Pi\in\mathbf{\Pi}_\alpha\right]
\end{aligned}
$$

On the other hand, using the definition of (ϵ,δ)-differential privacy (See Lemma 21 for the elementary derivation of the first inequality):

$$
\begin{aligned}
&\Pr_{(S_i,\Pi)}\left[(S_i,\Pi)\in B_\alpha^+\right]-\Pr_{S_i\otimes\Pi}\left[(S_i,\Pi)\in B_\alpha^+\right]\\
&\quad=\sum_{x\in\mathcal{X}}\Pr[S_i=x]\left(\Pr\left[\Pi\in\mathbf{\Pi}_\alpha^+(x)|S_i=x\right]-\Pr\left[\Pi\in\mathbf{\Pi}_\alpha^+(x)\right]\right)\\
&\quad\leq\sum_{x\in\mathcal{X}}\Pr[S_i=x]\left((e^\epsilon-1)\Pr\left[\Pi\in\mathbf{\Pi}_\alpha^+(x)\right]+\delta\right)\\
&\quad=(e^\epsilon-1)\Pr_{S_i\otimes\Pi}\left[(S_i,\Pi)\in B_\alpha^+\right]+\delta\\
&\quad\leq(e^\epsilon-1)\Pr\left[\Pi\in\mathbf{\Pi}_\alpha\right]+\delta\\
&\quad<(e^\epsilon-1)\Pr\left[\Pi\in\mathbf{\Pi}_\alpha\right]+2c\Pr\left[\Pi\in\mathbf{\Pi}_\alpha\right]\quad\text{(by Equation (1))}\\
&\quad=\left((e^\epsilon-1)+2c\right)\cdot\Pr\left[\Pi\in\mathbf{\Pi}_\alpha\right]
\end{aligned}
$$

This is a contradiction for $\alpha\geq(e^\epsilon-1)+2c$. ◀

▶ Remark 8. Note

1. Since differential privacy is closed under post processing, this claim can be generalized beyond queries contained in the transcript to any query generated as function of the transcript.
2. In the case of $(\epsilon,0)$-differential privacy, choosing $c=0$, the claim holds for every query with probability 1.

Combined with our general transfer theorem (Theorem 4), this directly yields a transfer theorem for differential privacy:

▶ **Theorem 9** (Transfer Theorem for (ϵ,δ)-Differential Privacy)**.** *Suppose that M is (ϵ,δ)-differentially private and (α,β)-sample accurate for linear queries. Then for every analyst $\mathcal{A}$ and $c,d>0$ it also satisfies:*

$$
\Pr_{S\sim\mathcal{P}^n,\,\Pi\sim Interact(M,\mathcal{A};S)}[\max_j|a_j-q_j(\mathcal{P})|>\alpha+(e^\epsilon-1)+c+2d]\leq\frac{\beta}{c}+\frac{\delta}{d}
$$

i.e. it is (α',β')-distributionally accurate for $\alpha'=\alpha+(e^\epsilon-1)+c+2d$ and $\beta'=\frac{\beta}{c}+\frac{\delta}{d}$.

▶ Remark 10. As we will see in Section 4, the Gaussian mechanism (and many other differentially private mechanisms) has a sample accuracy bound that depends only on the square root of the log of both $1/\beta$ and $1/\delta$. Thus, despite the Markov-like term $\beta' = \frac{\beta}{c} + \frac{\delta}{d}$ in the above transfer theorem, together with the sample accuracy bounds of the Gaussian mechanism, it yields Chernoff-like concentration.

Our technique extends easily to reason about arbitrary low sensitivity queries and minimization queries. See Appendix A.1 and A.2 for more details.

4 Applications: The Gaussian Mechanism

We now apply our new transfer theorem to derive the concrete bounds that we plotted in Figure 1. The Gaussian mechanism is extremely simple and has only a single parameter σ: for each query q_i that arrives, the Gaussian mechanism returns the answer $a_i \sim \mathcal{N}(q_i(S), \sigma^2)$ where $\mathcal{N}(q_i(S), \sigma^2)$ denotes the Gaussian distribution with mean $q_i(S)$ and standard deviation σ. First, we recall the differential privacy properties of the Gaussian mechanism.

▶ **Theorem 11** ([2]). *When used to answer k linear queries, for every $0 < \delta < 1$, the Gaussian mechanism with parameter σ satisfies (ϵ, δ)-differential privacy for:*

$$\epsilon = \frac{k}{2n^2\sigma^2} + \sqrt{2\frac{k}{n^2\sigma^2}\log\left(\sqrt{\pi \cdot \frac{k}{2n^2\sigma^2}}/\delta\right)}$$

It is also easy to see that the sample-accuracy of the Gaussian mechanism is characterized by the CDF of the Gaussian distribution:

▶ **Lemma 12.** *For any $0 < \beta < 1$, the Gaussian mechanism with parameter σ is (α_G, β)-sample accurate for:*

$$\alpha_G = \sqrt{2}\sigma \cdot \mathrm{erfc}^{-1}\left(2 - 2\left(1 - \frac{\beta}{2}\right)^{1/k}\right) < \sqrt{2}\sigma \cdot \mathrm{erfc}^{-1}\left(\frac{\beta}{k}\right) < \sqrt{2}\sigma\sqrt{\log\left(\frac{\sqrt{2}k}{\pi\beta}\right)}.$$

Above, $\mathrm{erfc}(x) = 1 - \mathrm{erf}(x)$ is the complementary error function.

Proof. For a query q_j, write $a_j = q_j(S) + Z_j$ where $Z_j \sim \mathcal{N}(0, \sigma^2)$. The sample error is $\max_j |a_j - q_j(S)| = \max_j |Z_j|$. We have that $\Pr[\max_j |Z_j| \geq \alpha] \leq \Pr[\max_j Z_j \geq \alpha] + \Pr[\min_j Z_j \leq -\alpha]$. α_G is the value that solves the equation $\Pr[\max_j Z_j \geq \alpha] = \Pr[\min_j Z_j \leq -\alpha] = \beta/2$ ◀

With these quantities in hand, we can now apply Theorem 9 to derive distributional accuracy bounds for the Gaussian mechanism:

▶ **Theorem 13.** *Fix a desired confidence parameter $0 < \beta < 1$. When σ is set optimally, the Gaussian mechanism can be used to answer k linear queries while satisfying (α, β)-distributional accuracy, where α is the solution to the following unconstrained minimization problem:*

$$\alpha = \min_{\sigma,\delta>0}\left\{\sqrt{2}\sigma \cdot \mathrm{erfc}^{-1}\left(\frac{\delta}{k}\right) + e^{\frac{k}{2n^2\sigma^2} + \sqrt{2\frac{k}{n^2\sigma^2}\log\left(\sqrt{\pi\cdot\frac{k}{2n^2\sigma^2}}/\delta\right)}} - 1 + 6\left(\frac{\delta}{\beta}\right)\right\}$$

Proof. Using Theorem 9 and fixing $\beta' = \delta$ and $c = d$, we have that an (α', β')-sample accurate, (ϵ, δ)-differentially private mechanism is (α, β)-distributionally accurate for $\alpha = \alpha' + (e^\epsilon - 1) + 3c$ and $\beta = \frac{2\delta}{c}$ where c can be an arbitrary parameter. For any fixed value of β, we can take $c = \frac{2\delta}{\beta}$, and see that we obtain (α, β)-distributional accuracy where $\alpha = \alpha' + (e^\epsilon - 1) + 6\,(\delta/\beta)$. The theorem then follows from plugging in the privacy bound from Theorem 11, the sample accuracy bound from Theorem 12, and optimizing over the free variables σ and δ. ◀

5 Discussion

We have given a new proof of the transfer theorem for differential privacy that has several appealing properties. Besides being simpler than previous arguments, it achieves substantially better concrete bounds than previous transfer theorems, and uncovers new structural insights about the role of differential privacy and sample accuracy. In particular, sample accuracy serves to guarantee that the reported answers are close to their conditional means, and differential privacy serves to guarantee that the conditional means are close to their true answers. This focuses attention on the conditional data distribution as a key quantity of interest, which we expect will be fruitful in future work. In particular, it may shed light on what makes certain data analysts overfit less than worst-case bounds would suggest: because they choose queries whose conditional means are closer to the prior than the worst-case query.

There seems to be one remaining place to look for improvement in our transfer theorem: Lemmas 6 and 7 both exhibit a Markov-like tradeoff between a parameter c and β and δ respectively. Although the dependence on β and δ in our ultimate bounds is only root-logarithmic, it would still yield an improvement if this Markov-like dependence could be replaced with a Chernoff-like dependence. It *is* possible to do this for the β parameter: we give an alternative (and even simpler) proof of the transfer theorem for $(\epsilon, 0)$-differential privacy which shows that conditional distributions induced by private mechanisms exhibit Chernoff-like concentration, in Appendix D. But the only way we know to extend this argument to (ϵ, δ)-differential privacy requires dividing δ by a factor of n, which yields a final theorem that is inferior to Theorem 9.

References

1. Raef Bassily, Kobbi Nissim, Adam Smith, Thomas Steinke, Uri Stemmer, and Jonathan Ullman. Algorithmic stability for adaptive data analysis. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 1046–1059. ACM, 2016.
2. Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference*, pages 635–658. Springer, 2016.
3. Rachel Cummings, Katrina Ligett, Kobbi Nissim, Aaron Roth, and Zhiwei Steven Wu. Adaptive learning with robust generalization guarantees. In *Conference on Learning Theory*, pages 772–814, 2016.
4. Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toni Pitassi, Omer Reingold, and Aaron Roth. Generalization in adaptive data analysis and holdout reuse. In *Advances in Neural Information Processing Systems*, pages 2350–2358, 2015.
5. Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349(6248):636–638, 2015.
6. Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Leon Roth. Preserving statistical validity in adaptive data analysis. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 117–126. ACM, 2015.

7 Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

8 Sam Elder. Bayesian adaptive data analysis guarantees from subgaussianity. *arXiv preprint*, 2016. arXiv:1611.00065.

9 Sam Elder. Challenges in bayesian adaptive data analysis. *arXiv preprint*, 2016. arXiv:1604.02492.

10 Vitaly Feldman, Roy Frostig, and Moritz Hardt. The advantages of multiple classes for reducing overfitting from test set reuse. In *International Conference on Machine Learning*, pages 1892–1900, 2019.

11 Vitaly Feldman and Thomas Steinke. Generalization for Adaptively-chosen Estimators via Stable Median. In *Conference on Learning Theory*, pages 728–757, 2017.

12 Vitaly Feldman and Thomas Steinke. Calibrating Noise to Variance in Adaptive Data Analysis. In *Conference On Learning Theory*, pages 535–544, 2018.

13 Vitaly Feldman and Jan Vondrak. Generalization bounds for uniformly stable algorithms. In *Advances in Neural Information Processing Systems*, pages 9747–9757, 2018.

14 Andrew Gelman and Eric Loken. The Statistical Crisis in Science. *American Scientist*, 102(6):460, 2014.

15 Moritz Hardt and Jonathan Ullman. Preventing false discovery in interactive data analysis is hard. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pages 454–463. IEEE, 2014.

16 Katrina Ligett and Moshe Shenfeld. A necessary and sufficient stability notion for adaptive generalization. *arXiv preprint*, 2019. arXiv:1906.00930.

17 Seth Neel and Aaron Roth. Mitigating Bias in Adaptive Data Gathering via Differential Privacy. In *International Conference on Machine Learning (ICML)*, 2018.

18 Xinkun Nie, Xiaoying Tian, Jonathan Taylor, and James Zou. Why Adaptively Collected Data Have Negative Bias and How to Correct for It. In *International Conference on Artificial Intelligence and Statistics*, pages 1261–1269, 2018.

19 Kobbi Nissim and Uri Stemmer. Concentration Bounds for High Sensitivity Functions Through Differential Privacy. *Journal of Privacy and Confidentiality*, 9(1), 2019.

20 Ryan Rogers, Aaron Roth, Adam Smith, Nathan Srebro, Om Thakkar, and Blake Woodworth. Guaranteed Validity for Empirical Approaches to Adaptive Data Analysis. *arXiv preprint*, 2019. arXiv:1906.09231.

21 Ryan Rogers, Aaron Roth, Adam Smith, and Om Thakkar. Max-information, differential privacy, and post-selection hypothesis testing. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 487–494. IEEE, 2016.

22 Daniel Russo and James Zou. Controlling bias in adaptive data analysis using information theory. In *Artificial Intelligence and Statistics*, pages 1232–1240, 2016.

23 Thomas Steinke and Jonathan Ullman. Interactive fingerprinting codes and the hardness of preventing false discovery. In *Conference on Learning Theory*, pages 1588–1628, 2015.

24 Thomas Steinke and Jonathan Ullman. Subgaussian tail bounds via stability arguments. *arXiv preprint*, 2017. arXiv:1701.03493.

25 Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. In *Advances in Neural Information Processing Systems*, pages 2524–2533, 2017.

26 Tijana Zrnic and Moritz Hardt. Natural Analysts in Adaptive Data Analysis. In *International Conference on Machine Learning*, pages 7703–7711, 2019.

A Extensions

A.1 Low Sensitivity Queries

Our technique extends easily to reason about arbitrary *low sensitivity* queries. We only need to generalize our lemma about posterior stability.

▶ **Definition 14.** *A query $q : \mathcal{X}^n \to \mathbb{R}$ is called Δ-sensitive if for all pairs of neighbouring datasets $S, S' \in \mathcal{X}^n$: $|q(S) - q(S')| \leq \Delta$. Note that linear queries are $(1/n)$-sensitive.*

▶ **Lemma 15.** *If M is an (ϵ, δ)-differentially private mechanism for answering Δ-sensitive queries, then for any data distribution $\mathcal{P}$, analyst $\mathcal{A}$, and any constant $c > 0$:*

$$\Pr_{S\sim\mathcal{P}^n, \Pi\sim Interact(M,\mathcal{A};S))}\left[\max_j |q_j(\mathcal{Q}_\Pi) - q_j(\mathcal{P}^n)| > (e^\epsilon - 1 + 4c)n\Delta\right] \leq \frac{\delta}{c}$$

i.e. it is $(\epsilon', \frac{\delta}{c})$-posterior stable for every $\mathcal{A}$, where $\epsilon' = (e^\epsilon - 1 + 4c)n\Delta$.

Proof. We introduce a useful bit of notation: $\bar{q}\left(\boldsymbol{x}_{\leq i}\right) = \mathop{\mathbb{E}}_{S'\sim\mathcal{P}^{n-i}}\left[q\left(\left(\boldsymbol{x}_{\leq i}, S'\right)\right)\right]$. Notice that $\bar{q}\left(\boldsymbol{x}_{\leq 0}\right) = q\left(\mathcal{P}^n\right)$ and $\bar{q}\left(\boldsymbol{x}_{\leq n}\right) = q\left(\boldsymbol{x}\right)$. Given a transcript $\pi \in \mathbf{\Pi}$, let $j^*(\pi) \in \arg\max_j |q_j(\mathcal{Q}_\pi) - q_j(\mathcal{P}^n)|$. Denote for any $\alpha \geq 0$

$$\mathbf{\Pi}_\alpha = \left\{\pi \in \mathbf{\Pi} \,|\, q_{j^*(\pi)}\left(\mathcal{Q}_\pi\right) - q_{j^*(\pi)}\left(\mathcal{P}^n\right) > \alpha\right\}$$

and for any $z \in [0, 2\Delta]$ denote

$$\mathbf{\Pi}_{\alpha,z}\left(\boldsymbol{x}_{\leq i}\right) = \left\{\pi \in \mathbf{\Pi}_\alpha \,|\, \bar{q}_{j^*(\pi)}\left(\boldsymbol{x}_{\leq i}\right) - \bar{q}_{j^*(\pi)}\left(\boldsymbol{x}_{\leq i-1}\right) > z - \Delta\right\}$$

From the definition of differential privacy:

$$\begin{aligned}
& \mathop{\mathbb{E}}_{S\sim\mathcal{P}^n}\left[\sum_{\pi\in\mathbf{\Pi}_\alpha} \Pr_{\Pi\sim I(S)}\left[\Pi = \pi\right]\left(\bar{q}_{j^*(\pi)}\left(S_{\leq i}\right) - \bar{q}_{j^*(\pi)}\left(S_{\leq i-1}\right) + \Delta\right)\right] \\
= & \mathop{\mathbb{E}}_{S\sim\mathcal{P}^n}\left[\int_0^{2\Delta} \Pr_{\Pi\sim I(S)}\left[\Pi \in \mathbf{\Pi}_{\alpha,z}\left(S_{\leq i}\right)\right] dz\right] \\
\leq & \mathop{\mathbb{E}}_{S\sim\mathcal{P}^n, Y\sim\mathcal{P}}\left[\int_0^{2\Delta}\left(e^\epsilon \Pr_{\Pi\sim I(S^{i\leftarrow Y})}\left[\Pi \in \mathbf{\Pi}_{\alpha,z}\left(S_{\leq i}\right)\right] + \delta\right) dz\right] \\
= & \mathop{\mathbb{E}}_{S\sim\mathcal{P}^n, Y\sim\mathcal{P}}\left[e^\epsilon \sum_{\pi\in\mathbf{\Pi}_\alpha} \Pr_{\Pi\sim I(S^{i\leftarrow Y})}\left[\Pi = \pi\right]\left(\bar{q}_{j^*(\pi)}\left(S_{\leq i}\right) - \bar{q}_{j^*(\pi)}\left(S_{\leq i-1}\right) + \Delta\right) + 2\Delta\delta\right] \\
= & \mathop{\mathbb{E}}_{S\sim\mathcal{P}^n, Y\sim\mathcal{P}}\left[e^\epsilon \sum_{\pi\in\mathbf{\Pi}_\alpha} \Pr_{\Pi\sim I(S)}\left[\Pi = \pi\right]\left(\bar{q}_{j^*(\pi)}\left(S^{i\leftarrow Y}_{\leq i}\right) - \bar{q}_{j^*(\pi)}\left(S_{\leq i-1}\right) + \Delta\right) + 2\Delta\delta\right]
\end{aligned}$$

where $S^{i\leftarrow Y} = (S_1, \ldots, S_{i-1}, Y, S_{i+1}, \ldots, S_n)$, and the last equality follows from the observation that (S, Y) and $(S^{i\leftarrow Y}, S_i)$ are identically distributed. Since $Y \sim \mathcal{P}$, independently from Π, we get that $\mathbb{E}_{Y\sim\mathcal{P}}\left[\bar{q}_{j^*(\pi)}\left(S^{i\leftarrow Y}_{\leq i}\right)\right] = \bar{q}_{j^*(\pi)}\left(S_{\leq i-1}\right)$, so

$$\begin{aligned}
& \mathop{\mathbb{E}}_{S\sim\mathcal{P}^n}\left[\sum_{\pi\in\mathbf{\Pi}_\alpha} \Pr_{\Pi\sim I(S)}\left[\Pi = \pi\right]\left(\bar{q}_{j^*(\pi)}\left(S_{\leq i}\right) - \bar{q}_{j^*(\pi)}\left(S_{\leq i-1}\right) + \Delta\right)\right] \\
\leq & \mathop{\mathbb{E}}_{S\sim\mathcal{P}^n}\left[\left(e^\epsilon \Pr_{\Pi\sim I(S)}\left[\Pi \in \mathbf{\Pi}_\alpha\right] + 2\delta\right)\Delta\right] \\
= & \left(e^\epsilon \Pr\left[\Pi \in \mathbf{\Pi}_\alpha\right] + 2\delta\right)\Delta
\end{aligned}$$

Subtracting $\Delta \Pr[\Pi \in \Pi_\alpha]$ from both sides we get

$$\mathop{\mathbb{E}}_{S\sim\mathcal{P}^n}\left[\sum_{\pi\in\mathbf{\Pi}_\alpha}\Pr_{\Pi\sim I(S)}[\Pi=\pi]\left(\bar{q}_{j^*(\pi)}(S_{\le i})-\bar{q}_{j^*(\pi)}(S_{\le i-1})\right)\right] \le ((e^\epsilon-1)\Pr[\Pi\in\mathbf{\Pi}_\alpha]+2\delta)\Delta \tag{2}$$

We now chooose $\alpha = (e^\epsilon - 1 + 4c)\, n\Delta$. Suppose that $\Pr\left[\left|q_{j^*(\Pi)}(\mathcal{Q}_\Pi) - q_{j^*(\Pi)}(\mathcal{P}^n)\right| > \alpha\right] > \frac{\delta}{c}$. We must have that either $\Pr\left[q_{j^*(\Pi)}(\mathcal{Q}_\Pi) - q_{j^*(\Pi)}(\mathcal{P}^n) > \alpha\right] > \frac{\delta}{2c}$ or $\Pr\left[q_{j^*(\Pi)}(\mathcal{P}^n) - q_{j^*(\Pi)}(\mathcal{Q}_\Pi) > \alpha\right] > \frac{\delta}{2c}$. Without loss of generality, assume

$$\Pr\left[q_{j^*(\Pi)}(\mathcal{Q}_\Pi) - q_{j^*(\Pi)}(\mathcal{P}^n) > \alpha\right] = \Pr[\Pi\in\mathbf{\Pi}_\alpha] > \frac{\delta}{2c} \tag{3}$$

But this leads to a contradiction, since

$$\begin{aligned}
&\Pr[\Pi\in\mathbf{\Pi}_\alpha]\,(e^\epsilon-1+4c)\,n\Delta \\
&< \sum_{\pi\in\mathbf{\Pi}_\alpha}\Pr[\Pi=\pi]\left(q_{j^*(\pi)}(\mathcal{Q}_\pi)-q_{j^*(\pi)}(\mathcal{P}^n)\right) \\
&= \mathop{\mathbb{E}}_{S\sim\mathcal{P}^n}\left[\sum_{\pi\in\mathbf{\Pi}_\alpha}\Pr_{\Pi\sim I(S)}[\Pi=\pi]\left(q_{j^*(\pi)}(S)-q_{j^*(\pi)}(\mathcal{P}^n)\right)\right] \\
&= \sum_{i=1}^{n}\mathop{\mathbb{E}}_{S\sim\mathcal{P}^n}\left[\sum_{\pi\in\mathbf{\Pi}_\alpha}\Pr_{\Pi\sim I(S)}[\Pi=\pi]\left(\bar{q}_{j^*(\pi)}(S_{\le i})-\bar{q}_{j^*(\pi)}(S_{\le i-1})\right)\right] \\
&\le ((e^\epsilon-1)\Pr[\Pi\in\mathbf{\Pi}_\alpha]+2\delta)\,n\Delta \quad \text{(by Equation (2))} \\
&< \Pr[\Pi\in\mathbf{\Pi}_\alpha]\,(e^\epsilon-1+4c)\,n\Delta \quad \text{(by Equation (3))}
\end{aligned}$$

◀

We can combine this Lemma with Lemma 6 (which holds for any query type) to get our transfer theorem:

▶ **Theorem 16** (Transfer Theorem for Low Sensitivity Queries). *Suppose that M is (ϵ,δ)-differentially private and (α,β)-sample accurate for Δ-sensitive queries. Then for every analyst $\mathcal{A}$, $c,d>0$ it also satisfies:*

$$\Pr_{S\sim\mathcal{P}^n,\,\Pi\sim Interact(M,\mathcal{A};S)}\left[\max_j |a_j - q_j(\mathcal{P}^n)| > \alpha + c + (e^\epsilon-1+4d)n\Delta\right] \le \frac{\beta}{c}+\frac{\delta}{d}$$

i.e. it is (α',β')-distributionally accurate for $\alpha' = \alpha + c + (e^\epsilon-1+4d)n\Delta$ and $\beta' = \frac{\beta}{c}+\frac{\delta}{d}$.

A.2 Minimization Queries

▶ **Definition 17.** *Minimization queries are specified by a loss function $L : \mathcal{X}^n \times \Theta \to [0,1]$ where Θ is generally known as the "parameter space". An answer to a minimization query L is a parameter $\theta\in\Theta$. We work with Δ-sensitive minimization queries: for all pairs of neighbouring datasets $S, S'\in\mathcal{X}^n$ and all $\theta\in\Theta$, $|L(S,\theta) - L(S',\theta)| \le \Delta$.*

A mechanism M is (α,β)-sample accurate for minimization queries if for every data analyst $\mathcal{A}$ and every dataset $S\in\mathcal{X}^n$:

$$\Pr_{\Pi\sim Interact(M,\mathcal{A};S)}\left[\max_j\left|L_j(S,\theta_j) - \min_{\theta\in\Theta}L_j(S,\theta)\right| \ge \alpha\right] \le \beta$$

We say that M satisfies (α,β)-distributional accuracy for minimization queries if for every data analyst $\mathcal{A}$ and every data distribution $\mathcal{P}$:

$$\Pr_{S\sim\mathcal{P}^n,\Pi\sim Interact(M,A;S)}\left[\max_j\left|\mathop{\mathbb{E}}_{S'\sim\mathcal{P}^n}\left[L_j(S',\theta_j) - \min_{\theta\in\Theta}L_j(S',\theta)\right]\right| \ge \alpha\right] \le \beta$$

▶ Remark 18. Note that

$$\mathop{\mathbb{E}}_{S'\sim\mathcal{P}^n}[L_j(S',\theta_j)] - \min_{\theta\in\Theta}\mathop{\mathbb{E}}_{S'\sim\mathcal{P}^n}[L_j(S',\theta)] \leq \mathop{\mathbb{E}}_{S'\sim\mathcal{P}^n}\left[L_j(S',\theta_j) - \min_{\theta\in\Theta} L_j(S',\theta)\right]$$

So as long as the RHS is bounded, the LHS is bounded too.

▶ Remark 19. For a given Δ-sensitive minimization query L_j and an answer θ_j, define:

$$q_j(S) := L_j(S,\theta_j) - \min_{\theta\in\Theta} L_j(S,\theta) \quad \text{and} \quad a_j := 0$$

Note several things:

1. If L_j is Δ-sensitive, then q_j is 2Δ-sensitive.
2. The mapping from a minimization query transcript $\pi = ((L_1,\theta_1),\ldots,(L_k,\theta_k))$ to the 2Δ-sensitive query transcript $\pi' = ((q_1,a_1),\ldots,(q_k,a_k))$ as defined above is a dataset-independent post-processing $\pi' = f(\pi)$.
3. π satisfies an (α,β)-accuracy guarantee if and only if π' does.

With the above observation, the transfer theorem for minimization queries immediately follows by Lemma 15 and Lemma 6.

▶ **Theorem 20** (Transfer Theorem for Minimization Queries). *Suppose that M is (ϵ,δ)-differentially private and (α,β)-sample accurate for Δ-sensitive minimization queries. Then for every analyst $\mathcal{A}$ and $c,d>0$ it also satisfies:*

$$\Pr\left[\max_j \left|\mathop{\mathbb{E}}_{S'\sim\mathcal{P}^n}\left[L_j(S',\theta_j) - \min_{\theta\in\Theta} L_j(S',\theta)\right]\right| > \alpha + c + 2(e^\epsilon - 1 + 4d)n\Delta\right] \leq \frac{\beta}{c} + \frac{\delta}{d}$$

i.e. it is (α',β')-distributionally accurate for $\alpha' = \alpha + c + 2(e^\epsilon - 1 + 4d)n\Delta$ and $\beta' = \frac{\beta}{c} + \frac{\delta}{d}$.

B Details from Section 3.1

Proof of Lemma 5. This follows from the expansion of the definition, and an application of Bayes Rule.

$$\begin{aligned}
&\Pr_{S\sim\mathcal{P}^n,\Pi\sim I(S),S'\sim\mathcal{Q}_\Pi}[(S',\Pi)\in E]\\
&= \sum_{\boldsymbol{x}}\sum_{\pi}\sum_{\boldsymbol{x}'} \Pr[S=\boldsymbol{x}]\Pr[\Pi=\pi|S=\boldsymbol{x}]\Pr_{S'\sim\mathcal{Q}_\pi}[S'=\boldsymbol{x}']\mathbb{1}[(\boldsymbol{x}',\pi)\in E]\\
&= \sum_{\pi}\sum_{\boldsymbol{x}'} \Pr[\Pi=\pi]\Pr_{S'\sim\mathcal{Q}_\pi}[S'=\boldsymbol{x}']\mathbb{1}[(\boldsymbol{x}',\pi)\in E]\\
&= \sum_{\pi}\sum_{\boldsymbol{x}'} \Pr[\Pi=\pi]\Pr[S=\boldsymbol{x}'|\Pi=\pi]\mathbb{1}[(\boldsymbol{x}',\pi)\in E]\\
&= \sum_{\pi}\sum_{\boldsymbol{x}'} \Pr[\Pi=\pi]\frac{\Pr[\Pi=\pi|S=\boldsymbol{x}']\cdot\Pr[S=\boldsymbol{x}']}{\Pr[\Pi=\pi]}\mathbb{1}[(\boldsymbol{x}',\pi)\in E]\\
&= \Pr_{S\sim\mathcal{P}^n,\Pi\sim I(S)}[(S,\Pi)\in E]
\end{aligned}$$

◀

C Details from Section 3.2

▶ **Lemma 21.** *If M is (ϵ,δ)-differentially private, then for any event E and datapoint x:*

$$\Pr_{S\sim\mathcal{P}^n,S_i\sim S,\Pi\sim I(S)}[\Pi\in E|S_i=x] \leq e^\epsilon \Pr_{S\sim\mathcal{P}^n,\Pi\sim I(S)}[\Pi\in E] + \delta$$

Proof. This follows from expanding the definitions.

$$\begin{aligned}
\Pr_{S\sim\mathcal{P}^n, S_i\sim S, \Pi\sim I(S)}[\Pi \in E | S_i = x] &= \frac{1}{n}\sum_{i=1}^{n} \Pr_{S\sim\mathcal{P}^n, \Pi\sim I(S)}[\Pi \in E | S_i = x] \\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{\boldsymbol{x}\in\mathcal{X}^n} \Pr_{S\sim\mathcal{P}^n}[S = \boldsymbol{x}] \cdot \Pr[\Pi \in E | S = (\boldsymbol{x}_{-i}, x)] \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\sum_{\boldsymbol{x}\in\mathcal{X}^n} \Pr_{S\sim\mathcal{P}^n}[S = \boldsymbol{x}] \cdot (e^{\epsilon}\Pr[\Pi \in E | S = \boldsymbol{x}] + \delta) \\
&= e^{\epsilon} \Pr_{S\sim\mathcal{P}^n, \Pi\sim I(S)}[\Pi \in E] + \delta
\end{aligned}$$

where the inequality follows from the definition of differential privacy. ◀

D An (even) Simpler and Better Proof for ϵ-Differential Privacy

In this section we give an *even simpler* proof of an *even better* transfer theorem for $(\epsilon, 0)$-differential privacy. Rather than using Markov's inequality as we did in the proof of Lemma 6, we can directly show that conditional distributions induced by differentially private mechanisms exhibit Chernoff-like concentration.

▶ **Lemma 22.** *If M is $(\epsilon, 0)$-differentially private, then for any data distribution $\mathcal{P}$, any transcript $\pi \in \mathbf{\Pi}$, any linear query q, and any $\eta > 0$:*

$$\Pr_{S\sim\mathcal{Q}_\pi}\left[|q(S) - q(\mathcal{P})| \geq (e^{\epsilon} - 1) + \sqrt{\frac{2\ln(2/\eta)}{n}}\right] \leq \eta$$

Proof. Define the random variables $V_i = q(S_i) - \mathbb{E}[q(S_i)|S_{<i}]$, and let $X_i = \frac{1}{n}\sum_{j=1}^{i} V_j$. Then the sequence $0 = X_0, X_1, \ldots, X_n$ forms a martingale and $|X_i - X_{i-1}| = \frac{1}{n}|V_i| \leq \frac{1}{n}$. We can therefore apply Azuma's inequality to conclude that:

$$\Pr\left[\left|\frac{1}{n}\sum_{i=1}^{n} q(S_i) - \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[q(S_i)|S_{<i}]\right| \geq t\right] \leq 2\exp\left(\frac{-t^2 n}{2}\right) \tag{4}$$

Now fix any realization $\boldsymbol{x}$, and consider each term: $\mathbb{E}[q(S_i)|S_{<i} = \boldsymbol{x}_{<i}]$. We have:

$$\begin{aligned}
\mathop{\mathbb{E}}_{S\sim\mathcal{Q}_\pi}[q(S_i)|S_{<i} = \boldsymbol{x}_{<i}] &= \sum_x q(x) \cdot \Pr_{S\sim\mathcal{P}^n}[S_i = x | \Pi = \pi, S_{<i} = \boldsymbol{x}_{<i}] \\
&= \sum_x q(x) \cdot \frac{\Pr_{S\sim\mathcal{P}^n}[\Pi = \pi | S_i = x, S_{<i} = \boldsymbol{x}_{<i}] \cdot \Pr_{S\sim\mathcal{P}^n}[S_i = x]}{\Pr[\Pi = \pi | S_{<i} = \boldsymbol{x}_{<i}]} \\
&\leq e^{\epsilon} \cdot \sum_x q(x) \cdot \Pr_{S\sim\mathcal{P}^n}[S_i = x] \\
&= e^{\epsilon} q(\mathcal{P})
\end{aligned}$$

where the inequality follows from the definition of $(\epsilon, 0)$-differential privacy. Symmetrically, we can show that $\mathbb{E}_{S\sim\mathcal{Q}_\pi}[q(S_i)|S_{<i} = \boldsymbol{x}_{<i}] \geq e^{-\epsilon} q(\mathcal{P})$. Therefore we have that:

$$e^{-\epsilon} q(\mathcal{P}) \leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[q(S_i)|S_{<i}] \leq e^{\epsilon} q(\mathcal{P}).$$

Combining this with Equation 4 gives us that for any $\eta > 0$, with probability $1 - \eta$ when $S \sim \mathcal{Q}_\pi$:

$$q(S) \leq e^{\epsilon} q(\mathcal{P}) + \sqrt{\frac{2\ln(2/\eta)}{n}} \quad \text{and} \quad q(S) \geq e^{-\epsilon} q(\mathcal{P}) - \sqrt{\frac{2\ln(2/\eta)}{n}}$$ ◀

A transfer theorem follows immediately from lemma 22.

▶ **Theorem 23.** *Suppose that M is $(\epsilon, 0)$-differentially private and (α, β)-sample accurate. Then for any $\eta > 0$ it is (α', β')-distributionally accurate for $\alpha' = \alpha + (e^\epsilon - 1) + \sqrt{\frac{2\ln(2/\eta)}{n}}$ and $\beta' = \beta + \eta$.*

Proof. For a given π, let $j^*(\pi) = \arg\max_j |a_j - q_j(\mathcal{P})|$. By the triangle inequality we have:

$$\begin{aligned}|a_{j^*(\Pi)} - q_{j^*(\Pi)}(\mathcal{P})| &\leq |a_{j^*(\Pi)} - q_{j^*(\Pi)}(S)| + |q_{j^*(\Pi)}(S) - q_{j^*(\Pi)}(\mathcal{P})| \\ &\leq \max_j |a_j - q_j(S)| + |q_{j^*(\Pi)}(S) - q_{j^*(\Pi)}(\mathcal{P})|\end{aligned}$$

By the definition of (α, β)-sample accuracy, we have that with probability $1 - \beta$, $\max_j |a_j - q_j(S)| \leq \alpha$. The Resampling Lemma (Lemma 5) gives us that:

$$\begin{aligned}&\Pr_{S\sim\mathcal{P}^n,\, \Pi\sim I(S)}\left[|q_{j^*(\Pi)}(S) - q_{j^*(\Pi)}(\mathcal{P})| \geq (e^\epsilon - 1) + \sqrt{\frac{2\ln(2/\eta)}{n}}\right] \\ &= \Pr_{S\sim\mathcal{P}^n,\, \Pi\sim I(S),\, S'\sim\mathcal{Q}_\Pi}\left[|q_{j^*(\Pi)}(S') - q_{j^*(\Pi)}(\mathcal{P})| \geq (e^\epsilon - 1) + \sqrt{\frac{2\ln(2/\eta)}{n}}\right] \\ &= \mathop{\mathbb{E}}_{S\sim\mathcal{P}^n,\, \Pi\sim I(S)}\left[\Pr_{S'\sim\mathcal{Q}_\Pi}\left[|q_{j^*(\Pi)}(S') - q_{j^*(\Pi)}(\mathcal{P})| \geq (e^\epsilon - 1) + \sqrt{\frac{2\ln(2/\eta)}{n}}\right]\right] \\ &\leq \eta\end{aligned}$$

Because Lemma 22 guarantees us that for *every* π,

$$\Pr_{S'\sim\mathcal{Q}_\pi}\left[|q_{j^*(\pi)}(S') - q_{j^*(\pi)}(\mathcal{P})| \geq (e^\epsilon - 1) + \sqrt{\frac{2\ln(2/\eta)}{n}}\right] \leq \eta.$$

The theorem then follows from a union bound. ◀