Quantifying Uncertainty of Unsupported Linear Queries for Private Query Release

Brett Mullins, Gerome Miklau, and Daniel Sheldon
University of Massachusetts, Amherst
{bmullins,miklau,sheldon}@cs.umass.edu

Abstract

We propose SMEB, a novel mechanism for quantifying uncertainty in private linear query release. SMEB computes an unbiased estimate of an upper bound on expected query error for general linear queries, given a set of initial measured queries. We present preliminary results evaluating SMEB empirically against a baseline error bound and show that the bound obtained from SMEB is tighter than the baseline across a range of privacy budgets.

1 Introduction

A fundamental and practically relevant problem in differential privacy is private query release: given a target workload of queries, release noisy answers to the target workload that approximate the true query answers while satisfying differential privacy. A common approach to this problem for linear queries is to answer a smaller set of queries and derive answers to the target queries. Query release mechanisms of this sort yield unbiased answers and have simple output distributions, if the measured queries support the target queries i.e. the target queries can be expressed as a linear combination of measured queries. Beyond obtaining point estimates of query answers, analysts can reason about the uncertainty of the derived answers by, for instance, bounding expected error. Moreover, such bounds are often obtainable without additional expendature of the privacy budget.

In the case where the target queries are unsupported by the measured queries, no bounds on expected error are known. Such bounds, however, can be useful in practice to economize the privacy budget even at the expense of additional budget. For instance, if an analyst spends a proportion of her remaining budget to obtain a bound on expected error, then, given that the bound is sufficiently low, she can use the derived answers and save the remaining privacy budget for future queries.

Contributions. We propose a mechanism called SMEB which outputs an unbiased estimate of an upper bound on the expected error for arbitrary workloads of linear queries at the expense of a specified amount of the privacy budget, given an initial measured workload. We evaluate SMEB empirically against a baseline error bound and show that the bound obtained from SMEB is tighter than the baseline across a range of privacy budgets.

2 Preliminaries

Let $\mathcal{X}$ be a data domain with d categorical attributes such that $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_d$ and attribute X_i has $n_i < \infty$ categories. A database X is a multi-set of records from $\mathcal{X}$. Let $\Pi_{i=1}^{d} n_i = n$ be the size of the domain. We represent dataset X by the n-length data vector x, which contains the counts in which each element of $\mathcal{X}$ occurs in X. Throughout this paper, we often abuse notation by referring to the data vector x as the database.

Linear queries are a rich class of queries which can express common data aggregations such as histograms, marginals, and data cubes. A linear query is a vector $w \in \mathbb{R}^n$, and the answer to w on

Preprint. Under review.

database x is a linear combination of x given by $w^T x$. A workload W is a collection of m linear queries arranged row-wise in an $m \times n$ matrix. The answer to workload W on database x is given by Wx. Depending on the context, we refer to a workload as either a matrix or a set of row vectors. We can relate two workloads if the queries in the former can be expressed as linear combinations of the queries in the latter. For workloads W, W', W supports W' if there exists real-valued weight matrix V of appropriate dimension such that $W' = VW$. If W does not support W', we say that W' is unsupported by W.

2.1 Differential Privacy

Differential privacy is a mathematical criterion of privacy that bounds the effect of any individual in the dataset on the output of a mechanism by adding noise to the computation.

Definition 1. (Differential Privacy; [1, 2]) Let $\mathcal{M} : \mathcal{X} \to \mathcal{Y}$ be a randomized mechanism. For any neighboring datasets x, x' that differ by at most one record, denoted $x \sim x'$, and all measureable subsets $S \subseteq \mathcal{Y}$:

- if $\Pr(\mathcal{M}(x) \in S) \leq \exp(\epsilon) \cdot \Pr(\mathcal{M}(x') \in S) + \delta$, then $\mathcal{M}$ satisfies (ϵ, δ)-approximate differential privacy, denoted (ϵ, δ)-DP;
- if $D_\gamma(\mathcal{M}(x)||\mathcal{M}(x')) \leq \rho\gamma$ for all $\gamma \in (1, \infty)$ where D_γ is the γ-Renyi diveregence between distributions $\mathcal{M}(x), \mathcal{M}(x')$, then $\mathcal{M}$ satisfies ρ-zCDP.

While (ϵ, δ)-DP is a more common notion, it is often more convenient to work with zCDP. There exists a conversion from zCDP to (ϵ, δ)-DP.

Proposition 1 (zCDP to DP Conversion; [3])**.** *If mechanism $\mathcal{M}$ satisfies ρ-zCDP, then $\mathcal{M}$ satisfies (ϵ, δ)-DP for any $\epsilon > 0$ and $\delta = \min_{\alpha > 1} \frac{\exp((\alpha-1)(\alpha\rho-\epsilon))}{\alpha-1} \left(1 - \frac{1}{\alpha}\right)^\alpha$.*

Next, we introduce two building block mechanisms. An important quantity in analyzing the privacy of a mechanism is sensitivity. The L_p sensitivity of a function $f : \mathcal{X} \to \mathbb{R}$ is given by $\Delta_p(f) = \max_{x \sim x'} \|f(x) - f(x')\|_p$. If f is clear from the context, we write Δ_p.

Proposition 2 (zCDP of Gaussian mechanism; [2])**.** *Let W be an $m \times n$ workload. Given dataset x, the Gaussian mechanism adds i.i.d. Gaussian noise to Wx with scale parameter σ i.e., $\mathcal{M}(x) = Wx + \sigma\Delta_2(W)\mathcal{N}(0, \mathbb{I})$, where $\mathbb{I}$ is the $m \times m$ identity matrix. Then the Gaussian Mechanism satisfies $\frac{1}{2\sigma^2}$-zCDP.*

Proposition 3 (zCDP of exponential mechanism; [4, 5])**.** *Let $\epsilon > 0$ and $\text{Score}_r : \mathcal{X} \to \mathbb{R}$ be a quality score of candidate $r \in \mathcal{R}$ for dataset x. Then the exponential mechanism outputs a candidate $r \in \mathcal{R}$ according to the following distribution: $\Pr(\mathcal{M}(x) = r) \propto \exp\left(\frac{\epsilon}{2\Delta_1}\text{Score}_r(x)\right)$. The exponential mechanism satisfies $\frac{\epsilon^2}{8}$-zCDP.*

2.2 Private Query Release

A general recipe for private query release is *select-measure-reconstruct*. To obtain answers to a target workload Q, mechanisms following this recipe select both a workload M and a sufficient amount of noise to satisfy a given level of privacy. Then the mechanism privately measures Mx with the specified noise and derives answers to the target workload Q from noisy answers $\tilde{y}$.

The Matrix Mechanism is a well-known instance of the above recipe [6–8]. Given a target workload Q, the Matrix Mechanism selects both a strategy workload M supporting Q that minimizes expected workload error and Gaussian mechanism scaling parameter σ and measures $\tilde{y} = Mx + \sigma\Delta_2(M)\mathcal{N}(0, \mathbb{I})$. From these noisy answers, the mechanism reconstructs answers to Q by inferring a synthetic data vector $M^+\tilde{y}$, where M^+ is the Moore-Penrose pseudoinverse of M, and computing $QM^+\tilde{y}$. Expected workload error is given by $Err_{\tilde{y}}(Q) = \mathbb{E}_{\tilde{y}}[\|Qx - QM^+\tilde{y}\|_2]$.

In this paper, we restrict our attention to query release mechanisms that conform to the above recipe and use the pseudoinverse of the measured workload for the reconstruct step. Observe that using the inferred synthetic data vector, we can reconstruct answers not just to supported queries but to arbitrary linear queries. However, there are no known bounds on the expected workload error for unsupported queries. For example, suppose M is the workload of all two-way marginals and Q is a single three-way marginal workload. While Q is not supported by M, answers to Q can be inferred

by $QM^+\tilde{y}$, but bounds on the expected error of these answers are not known. See Appendix B for a description of workloads used thoughtout this paper.

3 Select-Measure-Estimate Bound (SMEB) Mechanism

To bound the expected workload error of the target workload, we introduce the Select-Measure-Estimate Bound (SMEB) mechanism, presented in Algorithm 1. SMEB estimates an upper bound $\hat{\beta}_Q$ on workload error by spending a specified amount of the privacy budget.

Example 1. *Suppose an analyst working under differential privacy wants to answer all two-dimensional range queries (Q) for a given dataset and has already measured k queries from the workload (M). For answers to be useful to the analyst, the expected error for any given query is to be at most η. By running SMEB, the analyst spends a proportion of her remaining privacy budget to estimate a bound on workload error $\hat{\beta}_q$ for $q \in Q$. If $\hat{\beta}_q < \eta$ for all q, then the analyst can use the derived answers to Q and save her remaining privacy budget.*

SMEB utilizes the property that the target workload Q can be additively decomposed into two workloads Q_M, Q_C, the first of which is supported by the measured workload M, and each can be bounded independently. For Q_M, we derive a closed-form, data independent upper bound on $Err_{\tilde{y}}(Q_M)$ that is a corollary to a corresponding result from the Matrix Mechanism [7]. Since Q_C is not supported by M in general, however, we introduce the candidate workload C, which supports Q_C for any additive decomposition of Q. In practice, we choose C to be the smallest collection of marginals such that C supports both M, Q. For instance, with Example 1, we choose C as all two-way marginals.

Algorithm 1: Select-Measure-Estimate Bound (SMEB) Mechanism

Data: Data vector x, target workload Q, candidate workload C, measured workload M, measured workload answers $\tilde{y}$, privacy budget ρ

Result: Bound estimates $\hat{\beta}_Q$

$\epsilon \leftarrow 2\sqrt{\rho}$

$\sigma \leftarrow \sqrt{1/\rho}$

select $c^* \in C$ using the exponential mechanism with budget ϵ and score function $\text{Score}_W = \|Wx - WM^+\tilde{y}\|_2$

measure c^* using Gaussian noise: $\tilde{a} \leftarrow {c^*}^T x + \sigma\Delta_2(c^*)\mathcal{N}(0, \mathbb{I})$

estimate $\hat{\beta}_c \leftarrow \left\|\tilde{a} - {c^*}^T M^+\tilde{y}\right\|_2 + \frac{2\Delta_1}{\epsilon}\log(|C|)$

$Q_M \leftarrow U^*M$ where:

$$U^*, V^* = \arg\min_{U,V} \quad \|V\|_{1,1}$$
$$\text{s.t.} \quad Q = UM + VC$$

estimate $\hat{\beta}_Q \leftarrow \sqrt{\Delta_2(Q_M)\sigma}\,\|Q_M M^+\|_F + \hat{\beta}_c\,\|V^*\|_{1,1}$

To bound Q_C, we run the exponential mechanism over the queries in C with workload error as the score function to privately select the query $c^* \in C$ with approximately the highest workload error. By privately measuring c^* with Gaussian noise, we compute an unbiased estimate $\hat{\beta}_C$ of an upper bound on $Err_{\tilde{y}}(c)$ for all $c \in C$, similar to a technique used in [8]. By decomposing queries in Q_C into linear combinations of queries in C and applying the $\hat{\beta}_C$ bound, we obtain a bound on $Err_{\tilde{y}}(Q_C)$. Summing the bounds for Q_C and Q_M yields a bound on $Err_{\tilde{y}}(Q)$. Note that the additive decomposition above is arbitrary. As a heuristic, SMEB uses the decomposition that minimizes the weight on the Q_C bound, since this bound uses worst-case estimates of the workload error from queries in C. The resulting bound $\hat{\beta}_Q$ is unbiased, since it is a linear transformation of an unbiased estimate. Note that SMEB can be run without additional expendature of the privacy budget for $Q' \subseteq Q$ such as bounding error for individual queries in Q. We utilize this fact in Section 4.

Let us demonstrate the formal properties of SMEB. Theorem 1 shows that SMEB satisfies differential privacy, and Theorem 2 shows that $\hat{\beta}_Q$ is an unbiased estimate of a valid upper bound β_Q on the expected workload error of Q. Proofs of these results are in Appendix A.

Theorem 1. *Algorithm 1 satifies ρ-zCDP.*

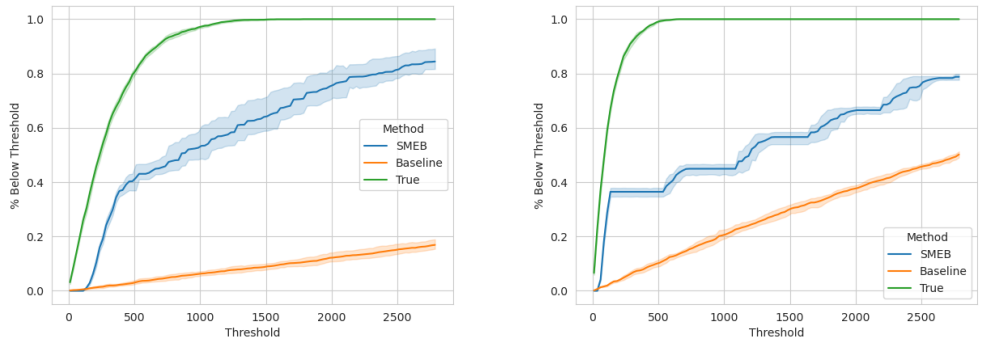
Theorem 2. *Let $x, Q, C, M, c^*, \epsilon, \tilde{y}, \tilde{a}$ be defined as in Algorithm 1 and suppose $Q = Q_M + Q_C$, C supports Q, M, and M supports Q_M. Define β_C, β_Q as follows:*

$$\beta_c = \mathbb{E}_{c^*,\tilde{a},\tilde{y}}\left[\left\|\tilde{a} - {c^*}^T M^+ \tilde{y}\right\|_2\right] + \frac{2\Delta_1}{\epsilon}\log(|C|); \beta_Q = \sqrt{\text{Var}(b_1)}\left\|Q_M M^+\right\|_F + \beta_c \|V\|_{1,1}$$

where $\|\cdot\|_F$ is the Frobenius norm and $\|\cdot\|_{1,1}$ is sum of the absolute value over entries in the matrix. Then $Q_C = VC$ and $Err_{\tilde{y}}(Q) \leq \beta_Q$. Moreover, $\hat{\beta}_Q$ is an unbiased estimator of β_Q where $\hat{\beta}_Q$ is obtained by replacing random values of β_Q with observed values.

4 Experiments

To evaluate if SMEB outputs bounds that are sufficiently tight to be informative, we compare the per-query error bounds output by SMEB with a baseline bound. The baseline is obtained by measuring the workload error of each query in the target workload using Gaussian noise and taking the absolute value. We compare these methods on the Titanic dataset with seven attributes, 1304 records, and domain size 30618 [9]. The target workload is all two-dimensional range queries, the measured workload is five randomly chosen two-dimensional range queries, and the candidate workload is all two-way marginals.



(a) $\epsilon = 0.1$ (b) $\epsilon = 0.32$

Figure 1: Average % of per-query error bounds under threshold as a function of threshold over five trials. The privacy budget is divided equally among the five measurement steps and the bound estimation and satisfies (ϵ, δ)-DP for ϵ as given and $\delta = 1 \times 10^{-9}$.

The results in Figure 1 show that the bound obtained from SMEB is tighter than the baseline for $\epsilon \leq 0.32$. At an error tolerance of 500 for $\epsilon = 0.32$, approx. 38% of queries had a SMEB bound under the threshold, while only 11% had a baseline bound under the threshold. The green line represents the non-private error of deriving answers to the given query and serves as an upper bound on performance in this experiment. This suggests that SMEB is better able to utilize a small privacy budget than the baseline. Since the SMEB bound is stochastic, $\hat{\beta}_q$ may not be a valid upper bound for $q \in Q$. For $\epsilon = 0.1$, 76.6% of bounds were valid, and, for $\epsilon = 0.32$, 80.2% of bounds were valid.

5 Limitations and Future Work

SMEB shares many of the scalability limitations of the Matrix Mechanism. Both mechanisms require workloads to be explicitly represented and operations such as the pseudoinverse to be computed. Moreover, for large data domains, even materializing the data vector can be computationally taxing. Approaches to improve the scalability of the Matrix Mechanism such as HDMM [10, 11] and ResidualPlanner [12] utilize implicit Kronecker product representations of the workload. Future work will be to utilize implicit workload representations to improve the scalability of SMEB.

References

[1] Cynthia Dwork, Frank McSherry Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *TCC*, pages 265–284, 2006. doi: 10.29012/jpc.v7i3.405.

[2] Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference*, pages 635–658. Springer, 2016. doi: 10.1007/978-3-662-53641-4_24.

[3] Clément L. Canonne, Gautam Kamath, and Thomas Steinke. The discrete gaussian for differential privacy. In *NeurIPS*, 2020. URL `https://proceedings.neurips.cc/paper/2020/hash/b53b3a3d6ab90ce0268229151c9bde11-Abstract.html`.

[4] Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *FOCS*, 2007. doi: 10.1145/2090236.2090254.

[5] Mark Cesar and Ryan Rogers. Bounding, concentrating, and truncating: Unifying privacy loss composition for data analytics. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, volume 132 of *Proceedings of Machine Learning Research*, pages 421–457, 2021. URL `https://proceedings.mlr.press/v132/cesar21a.html`.

[6] Chao Li and Gerome Miklau. An adaptive mechanism for accurate query answering under differential privacy. *PVLDB*, 5(6):514–525, 2012.

[7] Chao Li, Gerome Miklau, Michael Hay, Andrew McGregor, and Vibhor Rastogi. The matrix mechanism: optimizing linear counting queries under differential privacy. *The VLDB Journal*, 24(6):757–781, 2015. doi: 10.1007/s00778-015-0398-x.

[8] Ryan McKenna, Brett Mullins, Daniel Sheldon, and Gerome Miklau. Aim: An adaptive and iterative mechanism for differentially private synthetic data. *Proc. VLDB Endow.*, 15(11): 2599–2612, Jul 2022. ISSN 2150-8097. doi: 10.14778/3551793.3551817. URL `https://doi.org/10.14778/3551793.3551817`.

[9] Thomas Cason Frank E. Harrell Jr. Encyclopedia titanica.

[10] Ryan McKenna, Gerome Miklau, Michael Hay, and Ashwin Machanavajjhala. Optimizing error of high-dimensional statistical queries under differential privacy. *Proceedings of the VLDB Endowment*, 11(10):1206–1219, 2018. doi: 10.14778/3231751.3231769.

[11] Ryan McKenna, Gerome Miklau, Michael Hay, and Ashwin Machanavajjhala. Hdmm: Optimizing error of high-dimensional statistical queries under differential privacy. *arXiv preprint arXiv:2106.12118*, 2021.

[12] Yingtai Xiao, Guanlin He, Danfeng Zhang, and Daniel Kifer. An optimal and scalable matrix mechanism for noisy marginals under convex loss functions. *arXiv preprint arXiv:2305.08175*, 2023.

[13] Adi Ben-Israel and Thomas NE Greville. *Generalized inverses: theory and applications*, volume 15. Springer, 2003.

A Proofs

Let us introduce two properties of zCDP that are used in the proof of Theorem 1.

Proposition 4 (zCDP Properties [2, 5])**.** *zCDP satisfies two properties of differential privacy:*

1. *(Adaptive Composition) Let $\mathcal{M}_1 : \mathcal{X} \to \mathcal{Y}_1$ satisfy ρ_1-zCDP and $\mathcal{M}_2 : \mathcal{X} \times \mathcal{Y}_1 \to \mathcal{Y}_2$ satisfy ρ_2-zCDP. Then $\mathcal{M} = \mathcal{M}_2(x, \mathcal{M}_1(x))$ satisfies $(\rho_1 + \rho_2)$-zCDP.*
2. *(Postprocessing) Let $\mathcal{M}_1 : \mathcal{X} \to \mathcal{Y}$ satisfy ρ-zCDP and $f : \mathcal{Y} \to \mathcal{Z}$ be a randomized algorithm. Then $\mathcal{M} : \mathcal{X} \to \mathcal{Z} = f \circ \mathcal{M}_1$ satisfies ρ-zCDP.*

Proof of Theorem 1.

Proof. Recall that the exponential mechanism satisfies $\frac{\epsilon^2}{8}$-zCDP and the Gaussian mechanism satisfies $\frac{1}{2\sigma^2}$-zCDP. Then $\frac{(2\sqrt{\rho})^2}{8} = \frac{\rho}{2}$ and $\frac{1}{2(\sqrt{1/\rho})^2} = \frac{\rho}{2}$. By adaptive composition and postprocessing, SMEB satisfies $\frac{\rho}{2} + \frac{\rho}{2} = \rho$-zCDP. □

Let us prove Theorem 2 piecemeal over the following results.

Lemma 1. *(Supported Workload Bound) Let M support W and $\tilde{y} = Mx + \tilde{b}$ where $\tilde{b} \sim \mathcal{N}(0, \sigma^2\mathbb{I})$. Then $\mathbb{E}_{\tilde{y}}[Wx - WM^+\tilde{Y}] = 0$ and*

$$Err_{\tilde{y}}(W) \leq \sigma \left\|WM^+\right\|_F$$

where $\|\cdot\|_F$ is the Frobenius norm.

Proof. Since M supports W, $W = UM$ for some U of appropriate dimension. Then

$$\begin{aligned} Wx - WM^+\tilde{y} &= W - WM^+(Mx + \tilde{b}) \\ &= UM - UMM^+Mx - WM^+\tilde{b} \\ &= -WM^+\tilde{b}. \end{aligned}$$

Observe that $UMM^+M = M$ by properties of the Moore-Penrose pseudoinverse [13]. Applying this result, we see that $WM^+\tilde{y}$ is an unbiased estimator for Wx:

$$\begin{aligned} \mathbb{E}_{\tilde{y}} Wx - WM^+\tilde{y} &= \mathbb{E}_{\tilde{b}}[-WM^+\tilde{b}] \\ &= 0. \end{aligned}$$

Let us now bound the workload error of W:

$$\begin{aligned} Err_{\tilde{y}}(W) &= \mathbb{E}_{\tilde{b}}\Big[\Big(\left\|WM^+\tilde{b}\right\|_2^2\Big)^{1/2}\Big] \\ &\leq \Big(\mathbb{E}_{\tilde{b}}\Big[\Big(\left\|WM^+\tilde{b}\right\|_2^2\Big)\Big]\Big)^{1/2} \\ &= \Big(\mathbb{E}_{\tilde{b}}\Big[\Big(\Sigma_{w\in W}|wM^+\tilde{b}|^2\Big)\Big]\Big)^{1/2} \\ &= \big(\Sigma_{w\in W}\mathrm{Var}\big(wM^+\tilde{b}\big)\big)^{1/2} \\ &= \Big(\mathrm{Var}(\tilde{b}_1)\Sigma_{w\in W}\left\|wM^+\right\|_2^2\Big)^{1/2} \\ &= \sigma\left\|WM^+\right\|_F. \end{aligned}$$

Note that the inequality follows from Jensen's inequality. □

Lemma 2. *Let C supports W. Then $W = VC$ and $Err_{\tilde{y}}(W) \leq \Sigma_{ij}|v_{ij}|Err_{\tilde{y}}(c_j)$.*

Proof. Since C supports W, $W = VC$. Let us bound the workload error of W:

$$\begin{aligned} Err_{\tilde{y}}(W) &= \mathbb{E}_{\tilde{y}}\Big[\left\|Wx - WM^+\tilde{y}\right\|_2\Big] \\ &= \mathbb{E}_{\tilde{y}}\Big[\left\|VCx - VCM^+\tilde{y}\right\|_2\Big] \\ &\leq \Sigma_i\mathbb{E}_{\tilde{y}}\Big[\left\|v_iCx - v_iCM^+\tilde{y}\right\|_2\Big] \\ &\leq \Sigma_{ij}\mathbb{E}_{\tilde{y}}\Big[\left\|v_{ij}c_j^Tx - v_{ij}c_j^TM^+\tilde{y}\right\|_2\Big] \\ &= \Sigma_{ij}|v_{ij}|Err_{\tilde{y}}(c_j). \end{aligned}$$

Note that both inequalities are applications of the Triangle inequality. □

Lemma 3. *(Workload Error Decomposition) Let $W = W_M + W_C$, C supports W, M, M support W_M, and $\tilde{y} = Mx + \tilde{b}$ where $\tilde{b} \sim \mathcal{N}(0, \sigma^2\mathbb{I})$. Then $W_C = VC$ and*

$$Err_{\tilde{y}}(W) \leq \sigma\left\|W_MM^+\right\|_F + \Sigma_{ij}|v_{ij}|Err_{\tilde{y}}(c_j).$$

Proof. Since C supports W, W_M and $W_C = W - W_M$, C supports W_C. Then $W_C = VC$. Observe that

$$\begin{aligned} Err_{\tilde{y}}(W) &\leq Err_{\tilde{y}}(W_M) + Err_{\tilde{y}}(W_C) \\ &\leq \sigma \left\| W_M M^+ \right\|_F + \Sigma_{ij} |v_{ij}| Err_{\tilde{y}}(c_j). \end{aligned}$$

The first inequality follows from the Triangle inequality and the second from Lemmas 1, 2. □

It remains to bound $Err_{\tilde{y}}(c)$ for $c \in C$. Let us first prove a helpful lemma.

Lemma 4. *Let $a, b \in \mathbb{R}^k, c = b + z, z \sim N(0, \sigma^2)^k$. Then*

$$\mathbb{E}_z[\|a - c\|_2] \geq \|a - b\|_2 .$$

Proof. Let $f(z) = \|a - b - z\|_2$. Then f is a convex function of z. By Jensen's inequality, $\mathbb{E}[f(z)] \geq f(\mathbb{E}[z])$. Since $\mathbb{E}[z] = \mathbf{0}$ and $f(\mathbf{0}) = \|a - b\|_2$, we obtain the desired result. □

Theorem 3. *(Candidate Workload Bound) Let $x, C, M, c^*, \epsilon, \tilde{y}, \tilde{a}$ be defined as in Algorithm 1. Define*

$$\beta_c = \mathbb{E}_{c^*, \tilde{a}, \tilde{y}} \left[\left\| \tilde{a} - {c^*}^T M^+ \tilde{y} \right\|_2 \right] + \frac{2\Delta_1}{\epsilon} \log(|C|).$$

Then, for all $c \in C$, $Err_{\tilde{y}}(c) \leq \beta_c$.

Proof. From the guarantees of the exponential mechanism, for all $c \in C$,

$$\mathbb{E}_{c^*}[\text{Score}_{c^*}] \geq \text{Score}_c - \frac{2\Delta_1}{\epsilon} \log(|C|). \tag{1}$$

Recall that $\text{Score}_c = \left\| c^T x - c^T M^+ \tilde{y} \right\|_2$. Observe the following:

$$\begin{aligned} \left\| c^T x - c^T M^+ \tilde{y} \right\|_2 &\leq \mathbb{E}_{c^* | \tilde{y}} \left[\left\| {c^*}^T x - {c^*}^T M^+ \tilde{y} \right\|_2 \right] + \frac{2\Delta_1}{\epsilon} \log(|C|) \\ &\leq \mathbb{E}_{c^*, \tilde{a} | \tilde{y}} \left[\left\| \tilde{a} - {c^*}^T M^+ \tilde{y} \right\|_2 \right] + \frac{2\Delta_1}{\epsilon} \log(|C|). \end{aligned}$$

The first inequality follows from (1) and the second from Lemma 4. Taking expectations with respect to $\tilde{y}$ and applying iterated expectations to the RHS expectation, the desired result is obtained. □

We can now put the above results together to derive a bound to the target workload Q.

Theorem 4. *(Workload Error Bound) Let $Q = Q_M + Q_C$, C support Q, M, M support Q_M, and $x, C, M, c^*, \epsilon, \tilde{y}, \tilde{a}$ be defined as in Algorithm 1. Define*

$$\beta_Q = \sqrt{\text{Var}(b_1)} \left\| Q_M M^+ \right\|_F + \beta_c \|V\|_{1,1}$$

where $\|\cdot\|_{1,1}$ is sum of the absolute value over entries in the matrix. Then $Q_C = VC$ and $Err_{\tilde{y}}(Q) \leq \beta_Q$.

Finally, we show that the bound output by SMEB is unbiased.

Theorem 5. *$\hat{\beta}_Q$ is an unbiased estimator of β_Q.*

Proof. Since $\hat{\beta}_C$ is calculated from a single sample from the distribution of β_C, $\hat{\beta}_C$ is an unbiased estimator for β_C. Moreover, $\hat{\beta}_Q, \beta_Q$ is obtained from a linear transformation of $\hat{\beta}_C, \beta_C$, respectively. Since the choice of the linear combination does not depend on the value of $\hat{\beta}_C$, $\hat{\beta}_Q$ is an unbiased estimator for β_Q. □

B Workload Descriptions

Two broad classes of linear queries are marginal queries and range queries. The k-way marginal query for attributes $X_{i_1}, \ldots, X_{i_k}$ with values $d_{i_1}, \ldots, d_{i_k} \in \mathcal{X}_{i_1} \times \cdots \times \mathcal{X}_{i_k}$ counts the number of records such that $X_{i_j} = d_{i_j}, 1 \leq j \leq k$. Such a query represents one entry in the histogram over $X_{i_1}, \ldots, X_{i_k}$. For attributes $X_{i_1}, \ldots, X_{i_k}$, we refer to the collection of all k-way marginal queries as the k-way marginal workload for $X_{i_1}, \ldots, X_{i_k}$. The k-way marginal workload for $X_{i_1}, \ldots, X_{i_k}$ computes the full histogram over these attributes. When the context is clear, we refer to a marginal workload simply a marginal. Finally, the collection of all k-way marginals for data domain $\mathcal{X}$ is referred to as the k-way marginal workload. The workload of all k-way marginals captures correlations between k attributes.

Range queries are an extension of marginal queries in which attributes can each take any value between a lower and upper boundary rather than a particular value. The one-dimensional range query for attribute X_i with lower and upper boundaries d_i^l, d_i^h counts the number of records where X_i takes on values between d_i^l, d_i^h. As with marginal queries, we can extend to a k-dimensional range query, the workload of all k-dimensional range queries for a given collection of attributes, and the workload of all k-dimensional range workloads for a given data domain $\mathcal{X}$.