Purifying Approximate Differential Privacy with Randomized Post-processing

Yingyu Lin[*], **Erchi Wang**[*], **Yi-An Ma, Yu-Xiang Wang**
University of California, San Diego
{yil208, erw011, yianma, yuxiangw}@ucsd.edu

Abstract

We propose a framework to convert (ε, δ)-approximate Differential Privacy (DP) mechanisms into $(\varepsilon', 0)$-pure DP mechanisms under certain conditions, a process we call "purification." This algorithmic technique leverages randomized post-processing with calibrated noise to eliminate the δ parameter while achieving near-optimal privacy-utility tradeoff for pure DP. It enables a new design strategy for pure DP algorithms: first run an approximate DP algorithm with certain conditions, and then purify. This approach allows one to leverage techniques such as strong composition and propose-test-release that require $\delta > 0$ in designing pure-DP methods with $\delta = 0$. We apply this framework in various settings, including Differentially Private Empirical Risk Minimization (DP-ERM), stability-based release, and query release tasks. To the best of our knowledge, this is the first work with a statistically and computationally efficient reduction from approximate DP to pure DP. Finally, we illustrate the use of this reduction for proving lower bounds under approximate DP constraints with explicit dependence in δ, avoiding the sophisticated fingerprinting code construction.

1 Introduction

Differential privacy (DP), in its original form [DMNS06, Definition 1], has only one privacy parameter ε. Over the two decades of research in DP, many have advocated that DP is too stringent to be practical and have proposed several relaxations. Among them, the most popular is arguably the *approximate DP* [DKM+06], which introduces a second parameter δ.

Definition 1 (Differential privacy [DMNS06, DR+14]) *A mechanism $\mathcal{M}$ satisfies (ε, δ)-differential privacy if, for all neighboring datasets $D \simeq D'$ (datasets differing in at most one entry) and for any measurable set $S \subseteq$ Range$(\mathcal{M})$, it holds that:*

$$\mathbb{P}[\mathcal{M}(D) \in S] \leq e^{\varepsilon}\mathbb{P}[\mathcal{M}(D') \in S] + \delta.$$

When $\delta = 0$, the definition is now fondly referred to as ε-*pure* DP. Choosing $\delta > 0$ significantly weakens the protection, as it could leave any event with probability smaller than δ completely unprotected.

Two common reasons why DP researchers adopt this relaxation are: **1. Utility gain:** approximate DP is perceived to be more practical, as it allows for larger utility; **2. Flexible algorithm design:** Many algorithmic tools (such as advanced composition and Propose-Test-Release) support approximate DP but not pure DP, which enables more flexible (and often more efficient) algorithm design when $\delta > 0$ is permitted. For these two reasons, it is widely believed that the relaxation is a *necessary evil*.

We argue that the claim of "worse utility for pure-DP" is oftentimes a *myth*. In some applications of DP, a pure-DP mechanism has both stronger utility and stronger privacy. For example, in low-dimensional private histogram release, the Laplace mechanism enjoys smaller variance in almost all

[*] Equal Contribution. Alphabetical order.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

regimes except when δ is too large to be meaningful (see Figure 1). In other cases, the poor utility is sometimes not caused by an information-theoretic barrier, but rather due to the second issue — it is often much harder to design optimal pure-DP mechanisms.

In the problem of privately releasing k linear queries of the dataset in $\{0,1\}^d$, it may appear that pure-DP mechanisms are much worse if we only consider composition-based methods. The composition of the Gaussian mechanism enjoys an expected worst-case error of $\mathcal{O}_p(\frac{\sqrt{k\log(1/\delta)}}{n\varepsilon})$. On the other hand, if we wish to achieve pure DP, the composition of the Laplace mechanism's error bound becomes $\mathcal{O}_p(\frac{k}{n\varepsilon})$. However, the bound can be improved to $\mathcal{O}_p(\frac{\sqrt{kd}}{n\varepsilon})$ when we use a more advanced pure-DP algorithm known as the K-norm mechanism [HT10] that avoids composition.

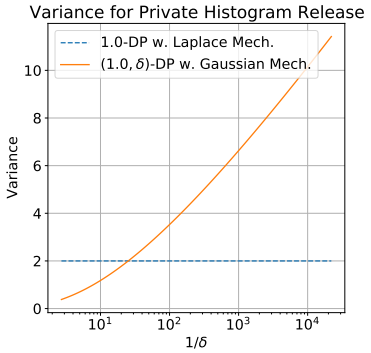

Figure 1: Per-coordinate variance of private histogram release under pure-DP and approximate DP with the same ε parameter. The dashed line is given by Laplace mechanism with pure 1.0-DP, regardless of δ. The solid line is generated using the analytic Gaussian mechanism in [BW18].

In the problem of private empirical risk minimization (DP-ERM) [BST14], the optimal excess empirical risk of $\mathbf{\Theta}\left(\frac{\sqrt{d\log(1/\delta)}}{n\varepsilon}\right)$ under approximate DP is achieved by the noisy-stochastic gradient descent (NoisySGD) mechanism and its full-batch gradient counterpart. These methods are natural and are by far the strongest approaches for differentially private deep learning as well [ACG+16, DBH+22]. The optimal rate of $\mathbf{\Theta}(\frac{d}{n\varepsilon})$ for pure DP, on the contrary, cannot be achieved by a Laplace noise version of NoisySGD, as advanced composition is not available for pure DP. Instead, it requires an exponential mechanism that demands a delicate method to deal with the mixing rate and sampling error of certain Markov chain Monte Carlo (MCMC) sampler [LMW+24].

The issue with the lack of algorithmic tools for pure DP becomes more severe in data-adaptive DP mechanisms. For example, all Propose-Test-Release (PTR) style methods [DL09] involve privately testing certain properties of the input dataset, which inevitably incurs a small failure probability that requires choosing $\delta > 0$. While smooth-sensitivity-based methods [NRS07] can achieve pure DP, they require adding heavy-tailed noise, which causes the utility to deteriorate exponentially as the dimension d increases.

1.1 Summary of the Results

In this paper, we develop a new algorithmic technique called "purification" that takes any (ε,δ)-approximate DP mechanism and converts it into an $(\varepsilon+\varepsilon')$-pure DP mechanism, while still enjoying similar utility guarantees of the original algorithm.

The contributions of our new technique for achieving pure DP are as follows.

1. It simplifies the design of the near-optimal pure DP mechanism by allowing the use of tools reserved for approximate DP.
2. It enables an $O(\sqrt{k})$-type composition without compromising the DP guarantee with $\delta > 0$, where k is the number of composition.
3. It shows that PTR-like mechanisms with pure DP are possible! To the best of our knowledge, our method is the only pure DP method for such purposes.
4. In DP-ERM and private selection problems, we show that, up to a logarithmic factor, the resulting pure DP mechanism enjoys an error rate that matches the optimal rate for pure DP, i.e., replacing the $\log(1/\delta)$ term with the dimension d or $\log(|\mathrm{OutputSpace}|)$.

Technical summary. The main idea of the "purification" technique is to use a randomized post-processing approach to "smooth" out the δ part of (ε,δ)-DP. To accomplish this, we proceed as follows. We leverage an equivalent definition of (ε,δ)-DP that interprets δ as the total variation distance from a pair of hypothetical distributions that are ε-indistinguishable. Next, we develop a method to convert the total variation distance to the ∞-Wasserstein distance. Finally, we leverage

Table 1: Summary of applications of our purification technique for constructing new pure-DP mechanisms from existing approximate DP mechanisms. The resulting pure-DP mechanisms are either *information-theoretically optimal or match the results of the best-known pure-DP mechanisms for the task* (we include more discussion in Appendix G). DP-ERM (SC) refers to DP-ERM with a strongly convex objective function, where λ denotes the strong convexity parameter of the individual loss function. DP-ERM (ℓ_1) refers to DP-ERM with an ℓ_1 constraint. $\lambda_{\min}$ in linear regression is the smallest eigenvalue of the sample covariance matrix, $X^T X/n$. All results are presented up to a logarithmic factor (in n and other parameters, but not in $1/\delta$. We note that in the table $\log(1/\delta)$ is proportional to d, the effective dimension.)

Problem	(ε, δ)-DP Mechanism	Utility (before purification)	Utility (after)
DP-ERM	Noisy-SGD [BST14]	$\frac{\sqrt{d\log(1/\delta)}L\|\mathcal{C}\|_2}{n\varepsilon}$	$\frac{dL\|\mathcal{C}\|_2}{n\varepsilon}$ (Thm 3)
DP-ERM (SC)	Noisy-SGD [BST14]	$\frac{d\log(1/\delta)L^2}{\lambda n^2\varepsilon^2}$	$\frac{d^2L^2}{\lambda n^2\varepsilon^2}$ (Thm 3)
DP-ERM (ℓ_1)	DP-Frank-Wolfe [TTZ14]	$\frac{(\log d)^{2/3}(\log(1/\delta))^{1/3}}{(n\varepsilon)^{2/3}}$	$\sqrt{\frac{\log d}{n\varepsilon}}$ (Thm 4)
Bounding Δ_{local}	PTR-type [KNRS13, DRE+20]	$\frac{d^{1/2}(\Delta_{\text{local}}+\log(1/\delta)/\varepsilon)}{\varepsilon}$	$\frac{d^{1/2}\Delta_{\text{local}}}{\varepsilon}+\frac{d^{3/2}}{\varepsilon^2}$ (Thm 5)
Mode Release	Distance to Instability [TS13]	$\frac{\log(1/\delta)}{\varepsilon}$	$\frac{\log\|\mathcal{X}\|}{\varepsilon}$ (Thm 6)
Linear Regression	AdaSSP [Wan18]	$\min\{\frac{\sqrt{d\log(1/\delta)}}{n\varepsilon}, \frac{d\log(1/\delta)}{\lambda_{\min}n^2\varepsilon^2}\}$	$\min\{\frac{d}{n\varepsilon}, \frac{d^2}{\lambda_{\min}n^2\varepsilon^2}\}$ (Thm 7)
Query Release	MWEM [HLM12]	$\frac{(\log k)^{1/2}(d\log(1/\delta))^{1/4}}{\sqrt{n\varepsilon}}$	$\frac{(\log k)^{1/3}d^{1/3}}{(n\varepsilon)^{1/3}}$ (Thm 8)

the Approximate Sample Perturbation (ASAP) technique from [LMW+24] that achieves pure DP by adding Laplace noise proportional to the ∞-Wasserstein distance. A challenge arises because the output distribution of a generic (ε, δ)-DP mechanism is not guaranteed to be supported on the entire output space, which may invalidate a tight TV distance to W_∞ conversion. We address this by interlacing a very small uniform distribution over a constraint set. Another challenge is that, unlike in [LMW+24], where the ε-indistinguishable distributions correspond to pure DP mechanisms on neighboring datasets, here the hypothetical distribution may depend on both neighboring datasets rather than just one. This prevents the direct application of the standard DP analysis of the Laplace mechanism and the composition theorem. To address this, we formulate our analysis in terms of the indistinguishability of general distributions rather than DP-specific language. We use a Laplace perturbation lemma and the weak triangle inequality, leading to a clean and effective analysis. Moreover, we show how the dimension-reduction technique can be used so the purification technique can be applied to discrete outputs and to sparse outputs, which ensures only logarithmic dependence in the output-space cardinality or dimension.

1.2 Related Work

The idea of using randomized post-processing to enhance privacy guarantees has been explored in prior work [FMTT18, MV22, LMW+24]. [FMTT18] focus on Rényi differential privacy, while [MV22, LMW+24] study the implementation of the exponential mechanism, specifically perturbing Markov chain Monte Carlo (MCMC) samples to obtain pure DP guarantees. Our work builds on [LMW+24], where we generalize the domain assumption of [LMW+24, Lemma 8] and extend the approach to more general upstream approximate DP mechanisms, such as DP-SGD [ACG+16, DBH+22]. We also discuss the connection to the classical statistics literature on randomized post-processing given by [B+51, Theorem 10] in Appendix B.

Previous works have investigated the purification of approximate differential privacy into pure DP, but limited to settings with a finite output space. A straightforward uniform mixing method can purify approximate DP mechanisms with finite output spaces, as summarized in [HC22], and we further discuss it in Appendix C. [BGH+23] first transform an approximate DP mechanism into a replicable one, and then apply a pure DP selection procedure to obtain a pure DP mechanism[2], at the cost of increased sample complexity.

2 Algorithm: Converting Approximate DP to Pure DP

In this section, we propose the purification algorithm (Algorithm 1) which converts (ε, δ)-approximate DP mechanisms with continuous output spaces into ε'-pure DP mechanisms under certain conditions. The algorithm consists of two steps: (1) mixing the approximate DP output with a uniform distribution

[2] Presented in [BGH+23, Algorithm 1], the algorithm can achieve a pure DP guarantee by replacing the approximate DP selection step with a pure DP alternative.

(Line 3), and (2) adding Laplace noise calibrated to $\delta^{\frac{1}{d}}$ (Line 4.) Intuitively, the first step is to enforce a bound on the ∞-Wasserstein distance (Definition 5), which can be loosely interpreted as a randomized analogue of ℓ_1 sensitivity. The second step adds Laplace noise proportional to this Wasserstein bound to guarantee pure DP. This step is based on techniques from [SWC17, LMW+24], and can be viewed as the generalization of the Laplace mechanism. The privacy and utility guarantee of Algorithm 1 are provided in Theorem 1.

Algorithm 1: $\mathcal{A}_{\text{pure}}(x_{\text{apx}}, \Theta, \varepsilon', \delta, \omega)$: Purification of Approximate Differential Privacy

1 **Input:** Privacy parameters ε, δ, additional privacy budget ε', an output x_{apx} of an (ε, δ)-DP algorithm $\mathcal{M}$ satisfying Assumption 1, an ℓ_q ball Θ as Assumption 1, the mixture level ω

2 Set $\Delta \leftarrow 2d^{1-\frac{1}{q}} R \left(\frac{\delta}{2\omega}\right)^{\frac{1}{d}}$ ▷ Lemma 6.

3 With probability $1-\omega$, set $x \leftarrow x_{\text{apx}}$; otherwise, sample $x \sim \text{Unif}(\Theta)$. ▷ Uniform Mixing

4 $x_{\text{pure}} \leftarrow x + \text{Laplace}^{\otimes \text{d}}\left(2\Delta/\varepsilon'\right)$

5 **Output:** x_{pure}

Notations. Let $\mathcal{X}$ be the space of data points, $\mathcal{X}^* := \cup_{n=0}^{\infty} \mathcal{X}^n$ be the space of the data set. For a vector $v = (v_1, \ldots, v_d) \in \mathbb{R}^d$ and $q \geq 1$, we define its ℓ_q-norm as $\|v\|_q := \left(\sum_{i=1}^d |v_i|^q\right)^{1/q}$. For a set $S \subseteq \mathbb{R}^d$, we denote its ℓ_q-norm diameter $\text{Diam}_q(S) := \sup_{x,y \in S} \|x-y\|_q$. For a (randomized) function $\mathcal{M} : \mathcal{X} \to \mathbb{R}^d$, we denote its range as $\text{Range}(\mathcal{M}) = \{\mathcal{M}(D) : D \in \mathcal{X}^*\}$.

Assumption 1. The (ε, δ)-DP algorithm $\mathcal{M}$ satisfies $\text{Diam}_q\left(\text{Range}(\mathcal{M})\right) \leq R$. Specifically, it lies in an ℓ_q ball of radius R, denoted by Θ.

Theorem 1 *Define x_{pure}, x_{apx} as in Algorithm 1 under Assumption 1. The output of Algorithm 1 satisfies $(\varepsilon + \varepsilon')$-DP with utility guarantee*

$$\mathbb{E}\left\|x_{pure} - x_{apx}\right\|_1 \leq \omega R + \frac{4Rd}{\varepsilon'}\left(\frac{\delta}{2\omega}\right)^{\frac{1}{d}}.$$

The detailed proofs are deferred to Appendix E and Appendix F. For clarity, Theorem 1 presents the utility guarantee in the ℓ_1 norm. Extensions to general ℓ_p norms follow directly from bounding the expected ℓ_q norm of the Laplace noise (see Equation (5)).

Remark 2 *When applying Algorithm 1 to various settings as shown in Table 1, the utility bounds either match the known information-theoretic lower bounds for pure DP or the best-known pure-DP mechanisms for the task. By the parameter setting given in Line 2 of Algorithm 1, the $\log(1/\delta)$ factor in the utility bounds can be replaced by d, omitting the logarithmic factors. In Section 4.2, we further show how dimension-reduction techniques can be used when applying purification to settings with sparsity conditions.*

Remark 3 *Parameter choices of Algorithm 1 for different settings are provided in later sections. For example, parameters for the purified DP-SGD are given in Corollary 4. Note that in Algorithm 1, we only require the range of $\mathcal{M}$ to be a* subset *of the ℓ_q-ball Θ, where Θ can be selected as ℓ_1 balls ($q = 1$), ℓ_2 balls ($q = 2$), or hypercubes ($q = \infty$), which admits simple $\mathcal{O}(d)$-runtime uniform sampling oracles.*

Corollary 4 (Parameters of Algorithm 1 for DP-SGD) *Let $\mathcal{M} : \mathcal{X}^* \to \Theta$ be an (ε, δ)-DP mechanism, where $\Theta \subset \mathbb{R}^d$ is an ℓ_2 ball with ℓ_2-diameter C. Let x_{apx} be the output of $\mathcal{M}$. Set the mixture level parameter as $\omega = \frac{1}{n^2}$, and set $\delta = \frac{2\omega}{(16Cdn^2)^d}$. Then, $\mathcal{A}_{pure}(x_{apx}, \Theta, \varepsilon' = \varepsilon, \delta, \omega)$ satisfies 2ε-DP guarantee with utility bound $\mathbb{E}[\|x_{pure} - x_{apx}\|_2] \leq \frac{1}{n^2\varepsilon} + \frac{C}{n^2}$.*

The purification algorithm can be applied to finite output spaces. The key idea is to embed the elements in the finite output space into a hypercube using the binary representation. Given a finite output space $\mathcal{Y} = \{1, 2, 3, \ldots, 2^d\} \doteq [|2^d|]$, we first map each element to its binary representation in $\{0,1\}^d$. Then, we apply Algorithm 1 on the cube $[0,1]^d$ in the Euclidean Space $\mathbb{R}^d$. The procedure is outlined in Algorithm 2 with DP and utility guarantees provided in Theorem 2, and the proof is deferred to Appendix F.

Algorithm 2: $\mathcal{A}_{\text{pure-discrete}}(\varepsilon, \delta, u_{\text{apx}}, \mathcal{Y})$: Binary Embedding Purification for Finite Spaces

Input: privacy parameters ε, δ, binary representation mapping $\mathsf{BinMap} : [2^d] \to \{0,1\}^d$, output u_{apx} from (ε, δ)-DP mechanism $\widetilde{M} : \mathcal{X}^* \to \mathcal{Y} = [2^d]$,
$z_{\text{bin}} \leftarrow \mathsf{BinMap}(u_{\text{apx}})$ ▷ Binary embedding
$z_{\text{pure}} \leftarrow \mathcal{A}_{\text{pure}}(z_{\text{bin}}, \Theta = [0,1]^d, \varepsilon' = \varepsilon, \delta, \omega = 2^{-d})$ ▷ Purify the embedding by Algorithm 1
$z_{\text{round}} \leftarrow \mathsf{Round}_{\{0,1\}^d}(z_{\text{pure}})$ ▷ $\mathsf{Round}_{\{0,1\}^d}(\boldsymbol{x}) = (\mathbf{1}(x_i \geq 0.5))_{i=1}^d$
$u_{\text{pure}} \leftarrow \mathsf{BinMap}^{-1}(z_{\text{round}})$ ▷ Decode index back to decimal integer index
Output: u_{pure}

Theorem 2 *If $\delta < \frac{\varepsilon^d}{(2d)^{3d}}$, then Algorithm 2 satisfies $(2\varepsilon, 0)$-pure DP with utility guarantee $\mathbb{P}[u_{apx} = u_{pure}] > 1 - 2^{-d} - \frac{d}{2}e^{-d}$.*

3 Technical Lemma: from TV distance to ∞-Wasserstein distance

In this section, we present a technical lemma that proofs the uniform mixing step in Algorithm 1 (Line 3) can enforce an ∞-Wasserstein distance bound, as mentioned in Section 2. This ∞-Wasserstein bound is a key step in our analysis, enabling the subsequent addition of Laplace noise calibrated to this bound to ensure pure DP. See Figure 4 for a summary of the privacy analysis. We define the ∞-Wasserstein distance below. Additional discussion can be found in Appendix A.

Definition 5 *The ∞-Wasserstein distance between distributions μ and ν is on a separable Banach space $(\Theta, \|\cdot\|_q)$ is defined as*

$$W_\infty^{\ell_q}(\mu,\nu) \doteq \inf_{\gamma \in \Gamma_c(\mu,\nu)} \operatorname{ess\,sup}_{(x,y)\sim\gamma} \|x-y\|_q = \inf_{\gamma \in \Gamma_c(\mu,\nu)} \{\alpha \mid \mathbb{P}_{(x,y)\sim\gamma}[\|x-y\|_q \leq \alpha] = 1\},$$

where $\Gamma_c(\mu,\nu)$ is the set of all couplings of μ and ν. The expression $\operatorname{ess\,sup}_{(x,y)\sim\gamma}$ denotes the essential supremum with respect to the measure γ.

By the equivalent characterization of approximate DP (Lemma 14), we can derive a TV distance bound between the output distributions of the (ε, δ)-DP mechanism on neighboring datasets and a pair of distributions that are ε-indistinguishable. Our goal is to translate this TV distance bound into a W_∞ distance bound. In general, the total variation distance bound does not imply a bound for the W_∞ distance. However, when the domain is bounded, we have the following result.

Lemma 6 (Converting d_{TV} to W_∞) *Let $q \geq 1$, and let $\Theta \subseteq \mathbb{R}^d$ be a convex set with ℓ_q-norm diameter R and containing an ℓ_q-ball of radius r. Let μ and ν be two probability measures on Θ. Suppose ν is the sum of two measures, $\nu = \nu_0 + \nu_1$, where ν_1 is absolutely continuous with respect to the Lebesgue measure and has density lower-bounded by a constant $p_{\min}$ over Θ, while ν_0 is an arbitrary measure. Define the W_∞ distance with respect to ℓ_q (Definition 5.) The following holds.*

$$\text{If } d_{\text{TV}}(\mu,\nu) < p_{\min} \cdot \text{Vol}(\mathbb{B}^d_{\ell_q}(1)) \cdot \left(\tfrac{r}{4R}\right)^d \cdot \Delta^d, \text{ then } W_\infty^{\ell_q}(\mu,\nu) \leq \Delta, \tag{1}$$

where $\text{Vol}(\mathbb{B}^d_{\ell_q}(1)) = \frac{2^d}{\Gamma(1+\frac{d}{q})} \prod_{i=1}^d \frac{\Gamma(1+\frac{1}{q})}{\Gamma(1+\frac{i}{q})}$ is the Lebesgue measure of the ℓ_q-norm unit ball, with Γ being the Gamma function. E.g., $\text{Vol}(\mathbb{B}^d_{\ell_2}(1)) = \frac{\pi^{d/2}}{\Gamma(\frac{d}{2}+1)}$, and $\text{Vol}(\mathbb{B}^d_{\ell_1}(1)) = \frac{2^d}{\Gamma(d+1)}$. In particular, if the domain Θ is an ℓ_q-ball, then the term $(\frac{r}{4R})$ Eq. (1) can be improved to $\frac{1}{4}$.

This result builds on [LMW+24] while generalizing its domain assumption from the ℓ_2-balls to more general convex sets. In the proof provided in Appendix F, instead of relying on [LMW+24, Lemma 24], which applies only to ℓ_2-balls, we construct a convex hull that extends to more general convex sets. We provide the proof sketch as follows.

Proof sketch of Lemma 6 To prove Lemma 6, we use an equivalent, non-coupling-based definition of the infinity-Wasserstein distance:

Lemma 7 ([GS84], Proposition 5) *Define μ, ν and W_∞ as Definition 5. Then,*

$$W_\infty^{\ell_q}(\mu,\nu) = \inf\{\alpha > 0 : \mu(U) \leq \nu(U^\alpha), \text{ for all open subsets } U \subset \Theta\},$$

where the α-expansion of U is denoted by $U^\alpha := \{x \in \Theta : \|x - U\|_q \leq \alpha\}$.

This definition provides a geometric interpretation of $W_\infty^{\ell_q}(\mu,\nu)$ by comparing the measure of a set U to the measure of its α-expansion U^α.

We summarize the key idea of the proof. Suppose $TV(\mu,\nu) \le \xi$. By the definition of total variation distance, this implies $\mu(U) \le \nu(U) + \xi$ for any measurable set U. To prove that $W_\infty^{\ell_q}(\mu,\nu) \le \Delta$ (for the Δ given in the lemma), it suffices, by Lemma 7, to show that $\mu(U) \le \nu(U^\Delta)$ for any open set U.

Given that $\mu(U) \le \nu(U) + \xi$, our goal thus reduces to proving $\nu(U) + \xi \le \nu(U^\Delta)$. Rewriting this inequality, it suffices to prove:

$$\nu(U^\Delta \setminus U) \ge \xi.$$

To establish this, we show that the "expansion band" $U^\Delta \setminus U$ must contain sufficient mass. We use the convexity of Θ to argue that this band must contain a small ℓ_q ball of a specific radius (related to Δ). We then use the minimum density $p_{\min}$ and the Lebesgue measure of this ℓ_q ball to lower-bound its ν-measure. The value of Δ in the lemma statement is chosen precisely so that this lower bound (and thus $\nu(U^\Delta \setminus U)$) is at least ξ, which completes the sketch. ■

The conversion lemma requires one of the distributions to satisfy a minimum density condition (the "$p_{\min}$".) This motivates the uniform mixing step in Algorithm 1, which ensures that the mixed distribution meets this requirement. Combining the TV bound implied by (ε,δ)-DP, the effect of uniform mixing, and the conversion lemma from TV distance to W_∞ distance, we obtain a bound on the ∞-Wasserstein distance after the uniform mixing step in Algorithm 1.

Example 8 (Tightness of the Conversion) *Let $\tilde{\Delta} \in (0,1)$. Consider the probability distributions $\mu = \tilde{\Delta}^d \delta_{\mathbf{0}}^{\mathrm{Dirac}} + (1-\tilde{\Delta}^d)\mathrm{Unif}\left(\mathbb{B}_{\ell_q}^d(1) \setminus \mathbb{B}_{\ell_q}^d(\tilde{\Delta})\right)$, and $\nu = \mathrm{Unif}\left(\mathbb{B}_{\ell_q}^d(1)\right)$, where $\delta_{\mathbf{0}}^{\mathrm{Dirac}}$ is the Dirac measure at $\mathbf{0}$, and $\mathrm{Unif}(S)$ denotes the uniform distribution over the set S. Then, by Definition 12 and Lemma 7, we have $d_{\mathrm{TV}}(\mu,\nu) = \tilde{\Delta}^d$, and $W_\infty^{\ell_q}(\mu,\nu) = \tilde{\Delta}$.*

This shows that the bound in Lemma 6 is tight up to a constant. In this case, we have $p_{\min} = (\mathrm{Vol}(\mathbb{B}_{\ell_q}^d(1)))^{-1}$ and $R = 2r$, so Lemma 6 gives the bound $W_\infty^{\ell_q} \le 8\tilde{\Delta}$, which matches the exact value $W_\infty^{\ell_q} = \tilde{\Delta}$ up to a constant.

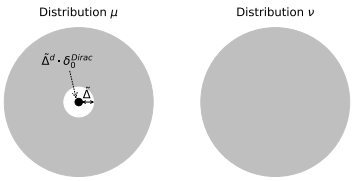

Figure 2: Illustration of Example 8 on the metric space $(\mathbb{R}^2, \|\cdot\|_2)$

4 Empirical Risk Minimization with Pure Differential Privacy

In this section, we apply our purification technique to develop pure differentially private algorithms for the Differential Private Empirical Risk Minimization (DP-ERM) problem, which has been extensively studied by the differential privacy community [CMS11, BST14, KJ16, WYX17, FKT20, KLL21, GHSGT23].

We consider the *convex* formulation of the DP-ERM problem, where the objective is to design a differentially private algorithm that minimizes the empirical risk $\mathcal{L}(\theta; D) = \frac{1}{n}\sum_{i=1}^n f(\theta; x_i)$ given a dataset $D = \{x_1, \ldots, x_n\} \subseteq \mathcal{X}^n$, a convex feasible set $\mathcal{C} \subseteq \mathbb{R}^d$, and a convex loss function $f : \mathcal{C} \times \mathcal{X} \to \mathbb{R}$. Algorithm performance is measured by the expected excess empirical risk $\mathbb{E}_{\mathcal{A}}[\mathcal{L}(\theta)] - \mathcal{L}^*$, where $\mathcal{L}^* = \min_{\theta \in \mathcal{C}} \mathcal{L}(\theta)$.

4.1 Purified DP Stochastic Gradient Descent

The Differential Private Stochastic Gradient Descent (DP-SGD) mechanisms [BST14, ACG+16] are the most popular algorithms for DP-ERM. These mechanisms are inherently iterative and heavily rely on (1) advanced privacy accounting/composition techniques [BS16, Mir17, DRS22] and (2) amplification by subsampling for Gaussian mechanisms [BBG18, BDRS18, WBK19, ZW19, KJH20].

Either of these two techniques results in (ε, δ)-DP guarantee with $\delta > 0$. In contrast, directly using the Laplace mechanism to release gradients fails to achieve a competitive utility rate.

While the exponential mechanism [MT07, BST14] achieves optimal utility rates under $(\varepsilon, 0)$-pure DP, this comes at the expense of increased computational complexity. Specifically, [BST14, Algorithm 2] implements the exponential mechanism via a random walk over the grid points of a cube that contains $\mathcal{C}$, ensuring convergence in terms of max-divergence. This approach constitutes a zero-order method that does not leverage gradient information. To design a fast, pure DP algorithm with nearly optimal utility, we propose a pure DP-SGD method that transforms the output of an (ε, δ)-DP SGD algorithm into a $(\varepsilon, 0)$-pure DP solution using Algorithm 1. Implementation details are provided in Algorithm 4 in Appendix H. Theoretical guarantees on utility, privacy, and computational efficiency are stated in Theorem 3.

Theorem 3 (Utility, privacy, and runtime for purified DP-SGD) *Let $\mathcal{C} \subset \mathbb{R}^d$ be a convex set with ℓ_2 diameter C, and suppose that $f(\cdot; x)$ is L-Lipschitz for every $x \in \mathcal{X}$. Set the parameters as specified in Corollary 4. Algorithm 4 guarantees 2ε-pure differential privacy. Furthermore, using $\tilde{\mathcal{O}}(n^2\varepsilon^{3/2}d^{-1})$ incremental gradient calls, the resulting output θ_{pure} satisfies:*

1. *If $f(\cdot; x)$ is convex for every $x \in \mathcal{X}$, then $\mathbb{E}_{\mathcal{A}}\left[\mathcal{L}(\theta_{\text{pure}})\right] - \mathcal{L}^* \leq \tilde{\mathcal{O}}\left(CLd/n\varepsilon\right)$.*
2. *If $f(\cdot; x)$ is λ-strongly convex for every $x \in \mathcal{X}$, then $\mathbb{E}_{\mathcal{A}}\left[\mathcal{L}\left(\theta_{\text{pure}}\right)\right] - \mathcal{L}^* \leq \tilde{\mathcal{O}}\left(d^2L^2/n^2\lambda\varepsilon^2\right)$.*

The total runtime of Algorithm 4 is $\tilde{\mathcal{O}}(n^2\varepsilon^{3/2} + d)$, where each incremental gradient computation incurs a cost of $\mathcal{O}(d)$ gradient operations, and the purification (Algorithm 1) executes in $\mathcal{O}(d)$ time. Table 2 presents a comparison of utility and computational efficiency. Notably, the purified DP-SGD attains a near-optimal utility rate, matching that of the exponential mechanism [BST14], while substantially reducing computation complexity by improving the dependence on both n and d.

Table 2: Comparison of utility and computational complexity in the ℓ_2 Lipschitz and convex setting under ε-pure differential privacy. For simplicity, we assume the data domain $\mathcal{C}$ has ℓ_2 diameter 1 and the Lipschitz constant $L = 1$. All results are stated up to logarithmic factors.

Mechanism	Reference	Utility	Runtime
Laplace Noisy GD	Lemma 19	$d^{1/2}/(n\varepsilon)^{1/2}$	$n^2\varepsilon$
Purified DP-SGD (Algo. 4)	Theorem 3	$d/n\varepsilon$	$n^2\varepsilon^{3/2} + d$
Exponential mechanism	[BST14, Theorem 3.4]	$d/n\varepsilon$	$d^4n^3 \vee d^3n^4\varepsilon$

4.2 Purified DP Frank-Wolfe Algorithm

As noted in Section 2, applying Algorithm 1 yields the relation $\log(1/\delta) \sim d\log(\varepsilon/\Delta)$, introducing an extra $\tilde{\mathcal{O}}(d)$ factor due to the Laplace noise scale Δ/ε. While optimal in low dimensions, this factor can degrade utility for algorithms with dimension-independent convergence, such as the Frank-Wolfe algorithm and its DP variants [FW+56, TTZ14, AFKT21, BGN21]. To address this, we integrate dimension-reduction and sparse recovery techniques (Algorithm 8) into our purification method (Algorithm 1). Applied to the (ε, δ)-DP Frank-Wolfe algorithm from [TGTZ15, Algorithm 2], which uses the exponential mechanism and advanced composition, our approach preserves dimension-independent convergence rates.

To obtain a pure-DP estimator from the approximate DP output θ_{FW}, we first apply dimension reduction, exploiting the problem's sparsity to project θ_{FW} into a lower-dimensional space. We then apply the purification Algorithm 1 in this space and recover the estimate in the original ambient space. The full procedure is given in Algorithm 7, with the proof of Theorem 4 deferred to Appendix I.3. Our method achieves pure differential privacy while matching the best known utility rate $\tilde{\mathcal{O}}(n^{-1/2}\varepsilon^{-1/2})$, as in [AFKT21, Theorem 6].

Theorem 4 *Let the domain $\mathcal{C}$ be an ℓ_1-ball centered at $\mathbf{0}$. Let ε be the pure differential privacy parameter. Assume that the function $f(\cdot; \mathrm{x})$ is convex, L_1-Lipschitz, and β-smooth with respect to the ℓ_1 norm for all $\mathrm{x} \in \mathcal{X}$. Algorithm 7 satisfies 2ε-pure differential privacy and achieves the following utility bound:*

$$\mathbb{E}_{\mathcal{A}}\left[\mathcal{L}(\theta_{\text{pure}})\right] - \mathcal{L}^* \leq \tilde{\mathcal{O}}\left(\frac{L_1^{1/2}\beta^{1/2}\|\mathcal{C}\|_1^{3/2}}{(n\varepsilon)^{1/2}}\right).$$

The runtime is $\mathcal{O}(dn^{3/2})$, plus a single call to a LASSO solver.

The upper bound is tight with respect to n and ε. We provide a lower bound in Lemma 30, which is derived using a packing argument. We defer proof details to Appendix I.4.

5 Pure DP Data-dependent Mechanisms

This section shows that our purification technique offers a systematic approach for designing data-dependent pure DP mechanisms. We present three examples: (1) the Propose-Test-Release (PTR) mechanism and its variant with privately released local sensitivity [DL09, KNRS13, DRE+20]; (2) stable value release methods, such as frequent item identification [TS13, Vad17]; and (3) a linear regression algorithm using adaptive perturbation of sufficient statistics [Wan18].

5.1 Propose-Test-Release with Pure Differential Privacy

Given a dataset $D \in \mathcal{X}^*$ and a query function $q : \mathcal{X} \to \mathbb{R}$, the Propose-Test-Release (PTR) framework proceeds in three steps: (1) Propose an upper bound b on $\Delta^q_{\text{Local}}(D)$, the local sensitivity of $q(D)$ (as defined in Definition 33); (2) Privately test whether D is sufficiently distant from any dataset that violates this bound; (3) If the test succeeds, assume the sensitivity is bounded by b, and use a differentially private mechanism, such as the Laplace mechanism with scale parameter b/ε, to release a slightly perturbed query response. However, due to the failure probability of the testing step, PTR provides only approximate differential privacy. By applying the purification technique outlined in Algorithm 10 in Appendix J.1, PTR can be transformed into a pure DP mechanism.

Next, we examine a variant of the PTR framework, presented in Algorithm 11, where the output space of the query function has dimension d, and the local sensitivity $\Delta^q_{\text{Local}}(D)$ is assumed to have bounded global sensitivity. In this approach, Algorithm 11 first privately constructs a high-probability upper bound on the local sensitivity. The query is then released with additive noise proportional to this bound, and the purification technique is applied to ensure a pure DP release. This method enables an adaptive utility upper bound that depends on the local sensitivity, as established in Theorem 5. The proof is provided in Appendix J.2.

Theorem 5 *Algorithm 11 satisfies 3ε-DP. Moreover, the output q_{pure} from Algorithm 11 satisfies*

$$\mathbb{E}[\|q_{\text{pure}} - q(D)\|_2] \leq \tilde{\mathcal{O}}\left(\frac{d^{1/2}\Delta^q_{\text{Local}}(D)}{\varepsilon} + \frac{d^{3/2}}{\varepsilon^2}\right).$$

5.2 Pure DP Mode Release

We address the problem of privately releasing the most frequent item, commonly referred to as mode release, argmax release, or voting [DR+14]. Our approach follows the distance-to-instability algorithm, $\mathcal{A}_{\text{dist}}$, introduced in [TS13] and further outlined in Section 3.3 of [Vad17]. Let $\mathcal{X}$ denote a finite data universe, and let $D \in \mathcal{X}^n$ be a dataset. We define the mode function $f : \mathcal{X}^n \to \mathcal{X}$, where $f(D)$ returns the most frequently occurring element in the dataset D.

Theorem 6 *Let $D \in \mathcal{X}^n$ be a dataset. Algorithm 12 satisfies 2ε-DP and runs in time $\mathcal{O}(n + \log|\mathcal{X}|)$. Furthermore, if the gap between the frequencies of the two most frequent items exceeds $\Omega\left(\log|\mathcal{X}|\log(\log|\mathcal{X}|/\varepsilon)/\varepsilon\right)$, Algorithm 12 returns the mode of D with probability at least $1 - \mathcal{O}(1/|\mathcal{X}|)$.*

The proof is provided in Appendix J.3. Consider the standard Laplace histogram approach, which releases the element with the highest noisy count. To ensure that the correct mode is returned with high probability, the gap between the largest and second-largest counts must exceed $\boldsymbol{\Theta}\left(\log|\mathcal{X}|/\varepsilon\right)$ [Vad17, Proposition 3.4]. Our result matches this bound up to a $\log\log|\mathcal{X}|$ factor.

5.3 Pure DP Linear Regression

We conclude this section by presenting a pure DP algorithm for the linear regression problem. Given a fixed design matrix $X \in \mathcal{X} \subset \mathbb{R}^{n\times d}$ and a response variable $Y \in \mathcal{Y} \subset \mathbb{R}^n$, we assume the existence of $\theta^* \in \Theta$ such that $Y = X\theta^*$. The non-private ordinary least squares estimator $(X^\top X)^{-1}X^\top Y$, requires computing the two sufficient statistics: $X^\top X$ and $X^\top Y$. These can be privatized using Sufficient Statistics Perturbation (SSP) [VS09, FGWC16] or its adaptive variant, AdaSSP [Wan18], to achieve improved utility.

To facilitate the purification procedure, we first localize the output of AdaSSP by deriving a high-probability upper bound. This bound is then used to clip the output of AdaSSP, after which the purification technique is applied. Implementation details are provided in Algorithm 13, and the corresponding utility guarantee is stated in Theorem 7. The proof is deferred to Appendix J.4.

Theorem 7 *Assume $X^\top X$ is positive definite and $\|\mathcal{Y}\|_2 \lesssim \|\mathcal{X}\|_2\|\theta^*\|_2$. Then, with high probability, the output θ_{pure} of Algorithm 13 satisfies:*

$$\text{MSE}(\theta_{\text{pure}}) \leq \tilde{\mathcal{O}}\left(\frac{d\|\mathcal{X}\|_2^2\|\theta^*\|_2^2}{n\varepsilon} \wedge \frac{d^2\|\mathcal{X}\|_2^4\|\theta^*\|_2^2}{\varepsilon^2 n^2 \lambda_{\min}}\right), \tag{2}$$

Here, $\lambda_{\min}$ denotes the normalized minimum eigenvalue, defined as $\lambda_{\min}(X^\top X/n)$, and MSE *denotes the mean squared error, given by* $\text{MSE}(\theta) = \frac{1}{2n}\|Y - X\theta\|_2^2$.

We note that [AD20] also proposes an ε-DP mechanism for linear regression using the approximate inverse sensitivity mechanism, which has an excess mean squared error of $\mathcal{O}\left(dL^2\text{tr}(X^\top X/n)/n^2\varepsilon^2\right)$, where L is the Lipschitz constant of the individual mean squared error loss function. We make two comparisons: (1) compared to the utility bound in [AD20, Proposition 4.3], Theorem7 avoids dependence on the Lipschitz constant, which in turn relies on the potentially unbounded quantity $\|\Theta\|_2$; (2) both algorithms exhibit a computational complexity of $\mathcal{O}(nd^2 + d^3)$. Nonetheless, the inverse sensitivity mechanism entails further computational overhead due to the rejection sampling procedure ([AD20, Algorithm 3]).

6 Pure DP Query Release

We study the private query release problem, where the data universe is defined as $\mathcal{X} = \{0,1\}^d$, and the dataset is represented as a histogram $D \in \mathbb{N}^{|\mathcal{X}|}$, satisfying $\|D\|_1 = n$. We consider linear query functions $q : \mathcal{X} \to [0,1]$, where $q \in Q$, and the workload Q consists of K distinct queries. For convenience, we define $Q(D) := \left(\frac{1}{n}\sum_{i=1}^n q_1(d_i), \ldots, \frac{1}{n}\sum_{i=1}^n q_K(d_i)\right)$. Our goal is to release a privatized version of $Q(D)$ that minimizes the ℓ_∞ error.

Our algorithm, described in Algorithm 16, proceeds in three steps: (1) the multiplicative weights exponential mechanism [HLM12] is used to release a synthetic dataset; (2) subsampling is applied to this synthetic dataset to further reduce its size; (3) the subsampled dataset is encoded into a binary representation and then passed through the purification procedure (Algorithm 2). The subsampling step is necessary to balance the additional error introduced by purification, which depends on the size of the output space. The privacy and utility guarantees are formalized in Theorem 8, with the proof deferred to Appendix K.2.

Theorem 8 *Algorithm 16 satisfies 2ε-DP, and the output D_{pure} of the following utility guarantee:*

$$\mathbb{E}[\|Q(D) - Q(D_{\text{pure}})\|_\infty] \leq \tilde{\mathcal{O}}\left(\frac{d^{1/3}}{n^{1/3}\varepsilon^{1/3}}\right).$$

Moreover, the runtime of the algorithm is $\tilde{\mathcal{O}}(nK + |\mathcal{X}| + \varepsilon^{2/3}n^{2/3}d^{1/3}|\mathcal{X}|K)$.

We emphasize that our algorithmic framework (Algorithm 16) is flexible. The upstream approximate differential privacy query release algorithm—currently instantiated using MWEM [HLM12]—can be replaced with other (ε, δ)-DP algorithms, such as the Private Multiplicative Weights method [HR10] or the Gaussian mechanism. Our approach achieves the tightest known utility upper bound for pure DP, as provided by the Small Database mechanism [BLR13], while offering improved runtime.

7 Purification as a Tool for Proving Lower Bounds

In this section, we demonstrate that the purification technique can be leveraged to establish utility lower bounds for (ε, δ)-differentially private algorithms. Lower bounds for pure differential privacy mechanisms are often more straightforward to derive, commonly relying on packing arguments [HT10, BBKN14]. In contrast, establishing lower bounds for approximate DP mechanisms typically requires more intricate constructions, such as those based on fingerprinting codes [BUV14, DSS+15, BSU17, LLL21, SU16].

We begin by providing the intuition behind how the purification technique can be used to prove lower bounds for approximate DP. Suppose there exists an (ε, δ)-DP mechanism with δ within appropriate

range such that $\log(1/\delta) \sim \tilde{\mathcal{O}}(d)$ and an error bound of $\mathcal{O}\left(d^c \log^{1-c}(1/\delta)/n\varepsilon\right)$ for some constant $c \in (0,1)$, and we assume the output space is a bounded set in $\mathbb{R}^d$. Then, Theorem 1 and Remark 2 imply that it is possible to construct a 2ε-DP mechanism with an error bound of $\tilde{\mathcal{O}}\left(d/n\varepsilon\right)$, where $\tilde{\mathcal{O}}$ omits logarithmic factors. This provides a new approach to proving lower bounds for (ε, δ)-DP algorithms via a *contrapositive argument*:

If no $\mathcal{O}(\varepsilon)$-pure-DP mechanism exists with an error bound $\tilde{\mathcal{O}}\left(\frac{d}{n\varepsilon}\right)$, then no (ε, δ)-DP mechanism can achieve an error bound of $\mathcal{O}\left(d^c \log^{1-c}(1/\delta)/n\varepsilon\right)$ for all appropriate δ, for any $c \in (0,1)$.

We apply this approach to derive the lower bound for (ε, δ)-DP mean estimation task. Specifically, let the data universe be $\mathcal{D} := \{-1/\sqrt{d}, 1/\sqrt{d}\}^d$, and consider a dataset $D = \{x_1, \ldots, x_n\} \subseteq \{-1/\sqrt{d}, 1/\sqrt{d}\}^d$. The objective is to privately release the column-wise empirical mean $\bar{D} = \frac{1}{n}\sum_{i=1}^n x_i \in [-1/\sqrt{d}, 1/\sqrt{d}]^d$. For pure DP we have the following lower bound:

Lemma 9 (Lemma 5.1 in [BST14], simplified) *For any ε-pure DP mechanism $\mathcal{M}$, there exists a dataset $D \in \{-1/\sqrt{d}, 1/\sqrt{d}\}^{n\times d}$ such that with probability at least $1/2$ over the randomness of $\mathcal{M}$, we have $\|\mathcal{M}(D) - \bar{D}\|_2 = \Omega\left(\frac{d}{n\varepsilon}\right)$.*

By applying the purification technique with an appropriately chosen δ and a contrapositive argument based on Lemma 9, we obtain the following (ε, δ)-differential privacy lower bound, as stated in Theorem 9. We defer proof to Appendix L.

Theorem 9 *Let $\varepsilon \le \mathcal{O}(1)$, and $\delta \in \left(\exp(-4d\log(d)\log^2(nd)),\ 4n^{-d}\log^{-2d}(8d)\right)$. For any (ε, δ)-DP mechanism $\mathcal{M}$, there exist a dataset $D \in \{-1/\sqrt{d}, 1/\sqrt{d}\}^{n\times d}$ such that with probability at least $1/4$ over the randomness of $\mathcal{M}$, we have*

$$\|\mathcal{M}(D) - \bar{D}\|_2 = \tilde{\Omega}\left(\frac{\sqrt{d\log(1/\delta)}}{\varepsilon n}\right).$$

Here, $\tilde{\Omega}(\cdot)$ hides all polylogarithmic factors, except those with respect to δ.

Remark 10 *For mean estimation under (ε, δ)-DP with exponentially small δ, we can also establish a $\tilde{\Omega}(d)$ lower bound via a packing argument, as detailed in Appendix L.3.*

Additional examples illustrating the use of the purification trick to derive lower bounds is provided in Appendix L.2, which includes bounds for one-way marginal release and private selection.

8 Conclusion and Limitation

In this paper, we introduced a novel purification framework that systematically converts approximate DP mechanisms into pure DP mechanisms via randomized post-processing. Our approach bridges the flexibility of approximate DP and the stronger privacy guarantees of pure DP. Our purification technique has broad applicability across several fundamental DP problems. In particular, we propose a faster pure DP algorithm for the DP-ERM problem that achieves near-optimal utility rates for pure DP. We also show that our method can be applied to design pure-DP PTR algorithms. A limitation of our approach is that when applying this continuous purification framework to mechanisms with a finite output space, the utility bound incurs an additional $\log\log|\mathcal{Y}|$ factor compared to prior work, where $|\mathcal{Y}|$ is the output space cardinality. Nonetheless, to our knowledge, this is the first systematic purification result applicable to continuous domains. Future work includes extending our framework to more adaptive settings and refining its applicability to high-dimensional problems.

Acknowledgments

The research is partially supported by NSF Awards #2048091, CCF-2112665 (TILOS), DARPA AIE program, the U.S. Department of Energy, Office of Science, and CDC-RFA-FT-23-0069 from the CDC's Center for Forecasting and Outbreak Analytics. We thank helpful discussion with Abrahdeep Thakurta, who pointed us towards proving the packing lower bound pure-DP DP-ERM under the L1 constraints (Lemma 30). We also thank anonymous reviewers from TPDP'25 for the reference to the *folklore* that mixing a uniform distribution alone suffices for "purifying" approximate DP in the discrete output space regime, as well as the anonymous reviewer from Neurips 2025 for a simple purification approach for Euclidean spaces that extends this discrete-set purification folklore (Appendix C and D). Finally, we thank Xin Lyu for pointing out a simple alternative proof of Theorem 9 based on a packing argument (Remark 10).

References

[ACG+16] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

[AD20] Hilal Asi and John C Duchi. Instance-optimality in differential privacy via approximate inverse sensitivity mechanisms. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 14106–14117. Curran Associates, Inc., 2020.

[AFKT21] Hilal Asi, Vitaly Feldman, Tomer Koren, and Kunal Talwar. Private stochastic convex optimization: Optimal rates in l1 geometry. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 393–403. PMLR, 18–24 Jul 2021.

[B+51] David Blackwell et al. Comparison of experiments. In *Proceedings of the second Berkeley symposium on mathematical statistics and probability*, volume 1, page 26, 1951.

[BBG18] Borja Balle, Gilles Barthe, and Marco Gaboardi. Privacy amplification by subsampling: Tight analyses via couplings and divergences. *Advances in neural information processing systems*, 31, 2018.

[BBKN14] Amos Beimel, Hai Brenner, Shiva Prasad Kasiviswanathan, and Kobbi Nissim. Bounds on the sample complexity for private learning and private data release. *Machine learning*, 94:401–437, 2014.

[BDRS18] Mark Bun, Cynthia Dwork, Guy N Rothblum, and Thomas Steinke. Composable and versatile privacy via truncated cdp. In *ACM Symposium on the Theory of Computing*, 2018.

[BGH+23] Mark Bun, Marco Gaboardi, Max Hopkins, Russell Impagliazzo, Rex Lei, Toniann Pitassi, Satchit Sivakumar, and Jessica Sorrell. Stability is stable: Connections between replicability, privacy, and adaptive generalization. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 520–527, 2023.

[BGN21] Raef Bassily, Cristóbal Guzmán, and Anupama Nandi. Non-euclidean differentially private stochastic convex optimization. In *Conference on Learning Theory*, pages 474–499. PMLR, 2021.

[BLR13] Avrim Blum, Katrina Ligett, and Aaron Roth. A learning theory approach to noninteractive database privacy. *Journal of the ACM (JACM)*, 60(2):1–25, 2013.

[BS16] Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of cryptography conference*, pages 635–658. Springer, 2016.

[BST14] Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *55th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2014*, pages 464–473. IEEE Computer Society, 2014.

[BSU17] Mark Bun, Thomas Steinke, and Jonathan Ullman. Make up your mind: The price of online queries in differential privacy. In *Proceedings of the twenty-eighth annual ACM-SIAM symposium on discrete algorithms*, pages 1306–1325. SIAM, 2017.

[BUV14] Mark Bun, Jonathan Ullman, and Salil Vadhan. Fingerprinting codes and the price of approximate differential privacy. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 1–10, 2014.

[BW18] Borja Balle and Yu-Xiang Wang. Improving the gaussian mechanism for differential privacy: Analytical calibration and optimal denoising. In *International conference on machine learning*, pages 394–403. PMLR, 2018.

[CHS14] Kamalika Chaudhuri, Daniel Hsu, and Shuang Song. The large margin mechanism for differentially private maximization. *Advances in Neural Information Processing Systems*, 27, 2014.

[CMS11] Kamalika Chaudhuri, Claire Monteleoni, and Anand D. Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(29):1069–1109, 2011.

[CT05] Emmanuel J Candes and Terence Tao. Decoding by linear programming. *IEEE transactions on information theory*, 51(12):4203–4215, 2005.

[DBH+22] Soham De, Leonard Berrada, Jamie Hayes, Samuel L Smith, and Borja Balle. Unlocking high-accuracy differentially private image classification through scale. *arXiv preprint arXiv:2204.13650*, 2022.

[DKM+06] Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *Advances in Cryptology-EUROCRYPT*, pages 486–503. Springer, 2006.

[DL09] Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 371–380, 2009.

[DMNS06] 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.

[DR+14] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[DRE+20] Chris Decarolis, Mukul Ram, Seyed Esmaeili, Yu-Xiang Wang, and Furong Huang. An end-to-end differentially private latent Dirichlet allocation using a spectral algorithm. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 2421–2431. PMLR, 13–18 Jul 2020.

[DRS22] Jinshuo Dong, Aaron Roth, and Weijie J Su. Gaussian differential privacy. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(1):3–37, 2022.

[DRV10] Cynthia Dwork, Guy N Rothblum, and Salil Vadhan. Boosting and differential privacy. In *2010 IEEE 51st annual symposium on foundations of computer science*, pages 51–60. IEEE, 2010.

[DSS+15] Cynthia Dwork, Adam Smith, Thomas Steinke, Jonathan Ullman, and Salil Vadhan. Robust traceability from trace amounts. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pages 650–669. IEEE, 2015.

[DTTZ14] Cynthia Dwork, Kunal Talwar, Abhradeep Thakurta, and Li Zhang. Analyze gauss: optimal bounds for privacy-preserving principal component analysis. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 11–20, 2014.

[FGWC16] James Foulds, Joseph Geumlek, Max Welling, and Kamalika Chaudhuri. On the theory and practice of privacy-preserving bayesian data analysis. *arXiv preprint arXiv:1603.07294*, 2016.

[FKT20] Vitaly Feldman, Tomer Koren, and Kunal Talwar. Private stochastic convex optimization: optimal rates in linear time. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2020, page 439–449. Association for Computing Machinery, 2020.

[FMTT18] Vitaly Feldman, Ilya Mironov, Kunal Talwar, and Abhradeep Thakurta. Privacy amplification by iteration. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 521–532. IEEE, 2018.

[FTS17] Kazuto Fukuchi, Quang Khai Tran, and Jun Sakuma. Differentially private empirical risk minimization with input perturbation. In *Discovery Science: 20th International Conference, DS 2017, Kyoto, Japan, October 15–17, 2017, Proceedings 20*, pages 82–90. Springer, 2017.

[FW+56] Marguerite Frank, Philip Wolfe, et al. An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2):95–110, 1956.

[GG23] Guillaume Garrigos and Robert M Gower. Handbook of convergence theorems for (stochastic) gradient methods. *arXiv preprint arXiv:2301.11235*, 2023.

[GHSGT23] Arun Ganesh, Mahdi Haghifam, Thomas Steinke, and Abhradeep Guha Thakurta. Faster differentially private convex optimization via second-order methods. In *Advances in Neural Information Processing Systems*, volume 36, pages 79426–79438, 2023.

[GLL22] Sivakanth Gopi, Yin Tat Lee, and Daogao Liu. Private convex optimization via exponential mechanism. In *Conference on Learning Theory*, pages 1948–1989. PMLR, 2022.

[GS84] Clark R Givens and Rae Michael Shortt. A class of wasserstein metrics for probability distributions. *Michigan Mathematical Journal*, 31(2):231–240, 1984.

[GTU22] Arun Ganesh, Abhradeep Thakurta, and Jalaj Upadhyay. Differentially private sampling from rashomon sets, and the universality of langevin diffusion for convex optimization. *arXiv preprint arXiv:2204.01585*, 2022.

[HC22] Yiyang Huang and Clément L Canonne. Lemmas of differential privacy. *arXiv preprint arXiv:2211.11189*, 2022.

[HLM12] Moritz Hardt, Katrina Ligett, and Frank McSherry. A simple and practical algorithm for differentially private data release. *Advances in neural information processing systems*, 25, 2012.

[HR10] Moritz Hardt and Guy N Rothblum. A multiplicative weights mechanism for privacy-preserving data analysis. In *2010 IEEE 51st annual symposium on foundations of computer science*, pages 61–70. IEEE, 2010.

[HT10] Moritz Hardt and Kunal Talwar. On the geometry of differential privacy. In *Proceedings of the forty-second ACM symposium on Theory of computing*, pages 705–714, 2010.

[INS+19] Roger Iyengar, Joseph P. Near, Dawn Song, Om Thakkar, Abhradeep Thakurta, and Lun Wang. Towards practical differentially private convex optimization. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 299–316, 2019.

[KJ16] Shiva Prasad Kasiviswanathan and Hongxia Jin. Efficient private empirical risk minimization for high-dimensional learning. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, page 488–497. JMLR.org, 2016.

[KJH20] Antti Koskela, Joonas Jälkö, and Antti Honkela. Computing tight differential privacy guarantees using fft. In *International Conference on Artificial Intelligence and Statistics*, pages 2560–2569. PMLR, 2020.

[KLL21] Janardhan Kulkarni, Yin Tat Lee, and Daogao Liu. Private non-smooth erm and sco in subquadratic steps. In *Advances in Neural Information Processing Systems*, volume 34, pages 4053–4064, 2021.

[KLT24] Guy Kornowski, Daogao Liu, and Kunal Talwar. Improved sample complexity for private nonsmooth nonconvex optimization. *arXiv preprint arXiv:2410.05880*, 2024.

[KNRS13] Shiva Prasad Kasiviswanathan, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. Analyzing graphs with node differential privacy. In *Theory of Cryptography: 10th Theory of Cryptography Conference, TCC 2013, Tokyo, Japan, March 3-6, 2013. Proceedings*, pages 457–476. Springer, 2013.

[LJSB12] Simon Lacoste-Julien, Mark Schmidt, and Francis Bach. A simpler approach to obtaining an o (1/t) convergence rate for the projected stochastic subgradient method. *arXiv preprint arXiv:1212.2002*, 2012.

[LLL21] Yin Tat Lee, Daogao Liu, and Zhou Lu. The power of sampling: Dimension-free risk bounds in private erm. *arXiv preprint arXiv:2105.13637*, 2021.

[LMW+24] Yingyu Lin, Yian Ma, Yu-Xiang Wang, Rachel Emily Redberg, and Zhiqi Bu. Tractable MCMC for private learning with pure and gaussian differential privacy. In *The Twelfth International Conference on Learning Representations*, 2024.

[LSS14] Adeline Langlois, Damien Stehlé, and Ron Steinfeld. Gghlite: More efficient multilinear maps from ideal lattices. In *Annual international conference on the theory and applications of cryptographic techniques*, pages 239–256. Springer, 2014.

[MBST22] Paul Mangold, Aurélien Bellet, Joseph Salmon, and Marc Tommasi. Differentially private coordinate descent for composite empirical risk minimization. In *International Conference on Machine Learning*, pages 14948–14978. PMLR, 2022.

[Mir17] Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th computer security foundations symposium (CSF)*, pages 263–275. IEEE, 2017.

[MMR12] Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Multiple source adaptation and the rényi divergence. *arXiv preprint arXiv:1205.2628*, 2012.

[MT07] Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pages 94–103. IEEE, 2007.

[MV22] Oren Mangoubi and Nisheeth Vishnoi. Sampling from log-concave distributions with infinity-distance guarantees. *Advances in Neural Information Processing Systems*, 35:12633–12646, 2022.

[NRS07] Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. Smooth sensitivity and sampling in private data analysis. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pages 75–84, 2007.

[NS21] Aleksandar Nikolov and Thomas Steinke. Open problem - optimal query release for pure differential privacy. https://differentialprivacy.org/open-problem-optimal-query-release/, 2021.

[RBHT12] Benjamin IP Rubinstein, Peter L Bartlett, Ling Huang, and Nina Taft. Learning in a large function space: Privacy-preserving mechanisms for svm learning. *Journal of Privacy and Confidentiality*, 4(1), 2012.

[RKW23] Rachel Redberg, Antti Koskela, and Yu-Xiang Wang. Improving the privacy and practicality of objective perturbation for differentially private linear learners. *Advances in Neural Information Processing Systems*, 36:13819–13853, 2023.

[SSTT21] Shuang Song, Thomas Steinke, Om Thakkar, and Abhradeep Thakurta. Evading the curse of dimensionality in unconstrained private glms. In *International Conference on Artificial Intelligence and Statistics*, pages 2638–2646. PMLR, 2021.

[SU16] Thomas Steinke and Jonathan Ullman. Between pure and approximate differential privacy. *Journal of Privacy and Confidentiality*, 7(2), 2016.

[SWC17] Shuang Song, Yizhen Wang, and Kamalika Chaudhuri. Pufferfish privacy mechanisms for correlated data. In *Proceedings of the 2017 ACM International Conference on Management of Data*, pages 1291–1306, 2017.

[TGTZ15] Kunal Talwar, Abhradeep Guha Thakurta, and Li Zhang. Nearly optimal private lasso. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

[Tia24] Kevin Tian. Cs395t: Continuous algorithms, part ix, sparse recovery, 2024.

[TS13] Abhradeep Guha Thakurta and Adam Smith. Differentially private feature selection via stability arguments, and the robustness of the lasso. In *Conference on Learning Theory*, pages 819–850. PMLR, 2013.

[TTZ14] Kunal Talwar, Abhradeep Thakurta, and Li Zhang. Private empirical risk minimization beyond the worst case: The effect of the constraint set geometry. *arXiv preprint arXiv:1411.5417*, 2014.

[Vad17] Salil Vadhan. *The Complexity of Differential Privacy*, pages 347–450. Springer, Yehuda Lindell, ed., 2017.

[Vai89] Pravin M Vaidya. Speeding-up linear programming using fast matrix multiplication. In *30th annual symposium on foundations of computer science*, pages 332–337. IEEE Computer Society, 1989.

[VEH14] Tim Van Erven and Peter Harremos. Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

[VS09] Duy Vu and Aleksandra Slavkovic. Differential privacy for clinical trial data: Preliminary evaluations. In *2009 IEEE International Conference on Data Mining Workshops*, pages 138–143. IEEE, 2009.

[Wal74] Alastair J Walker. New fast method for generating discrete random numbers with arbitrary frequency distributions. *Electronics Letters*, 10(8):127–128, 1974.

[Wan18] Yu-Xiang Wang. Revisiting differentially private linear regression: optimal and adaptive prediction & estimation in unbounded domain. *arXiv preprint arXiv:1803.02596*, 2018.

[WBK19] Yu-Xiang Wang, Borja Balle, and Shiva Prasad Kasiviswanathan. Subsampled rényi differential privacy and analytical moments accountant. In *The 22nd international conference on artificial intelligence and statistics*, pages 1226–1235. PMLR, 2019.

[Wu16] Yihong Wu. Ece598: Information-theoretic methods in high-dimensional statistics, 2016.

[WYX17] Di Wang, Minwei Ye, and Jinhui Xu. Differentially private empirical risk minimization revisited: Faster and more general. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[WZ10] Larry Wasserman and Shuheng Zhou. A statistical framework for differential privacy. *Journal of the American Statistical Association*, 105(489):375–389, 2010.

[ZW19] Yuqing Zhu and Yu-Xiang Wang. Poission subsampled rényi differential privacy. In *International Conference on Machine Learning*, pages 7634–7642. PMLR, 2019.

Contents

A Preliminaries

This section provides definitions of max divergence, total variation distance, and ∞-Wasserstein distance, and the lemmas that will be used in the privacy analysis of Theorem 1.

Notations. Let $\mathcal{X}$ be the space of data points, $\mathcal{X}^* := \cup_{n=0}^{\infty}\mathcal{X}^n$ be the space of the data set. For an integer n, let $[n] = \{1, \ldots, n\}$. A subgradient of a convex function f at x, denoted $\partial f(x)$, is the set of vectors $\mathbf{g}$ such that $f(y) \geq f(x) + \langle \mathbf{g}, y - x\rangle$, for all y in the domain. For simplicity, we assume the functions are differentiable in this paper and consider the gradient ∇f. The operators $\cdot \vee \cdot$ and $\cdot \wedge \cdot$ denote the maximum and minimum of the two inputs, respectively. We use $\|\cdot\|_q$ to denote ℓ_q norm. For a set A, $\|A\|_q := \sup_{x\in A}\|A\|_q$ represents the ℓ_q radius of set A and $\mathrm{Diam}_q(A) := \sum_{x,y\in A}\|x - y\|_q$ represent the ℓ_q diameter of set A. For a finite set S, we denote its cardinality by $|S|$. Throughout this paper, we use $\mathbb{E}_{\mathcal{A}}[\cdot]$ to denote taking expectation over the randomness of the algorithm.

A.1 Definitions on Distributional Discrepancy

Definition 11 ([VEH14], Theorem 6; [DTTZ14], Definition 3.6) *The Rényi divergence of order ∞ (also known as Max Divergence) between two probability measures μ and ν on a measurable space $(\Theta, \mathcal{F})$ is defined as:*

$$D_\infty(\mu\|\nu) = \ln \sup_{S\in\mathcal{F},\ \mu(S)>0}\left[\frac{\mu(S)}{\nu(S)}\right].$$

In this paper, we say that μ and ν are ε-indistinguishable if $D_\infty(\mu\|\nu) \leq \varepsilon$, and $D_\infty(\nu\|\mu) \leq \varepsilon$.

A mechanism $\mathcal{M}$ is ε-differentially private if and only if for every two neighboring datasets D and D', we have $D_\infty(\mathcal{M}(D)\|\mathcal{M}(D')) \leq \varepsilon,$ and $D_\infty(\mathcal{M}(D')\|\mathcal{M}(D)) \leq \varepsilon$.

Definition 12 *The total variation (TV) distance between two probability measures μ and ν on a measurable space $(\Theta, \mathcal{F})$ is defined as:*

$$d_{\mathrm{TV}}(\mu, \nu) = \sup_{S\in\mathcal{F}}|\mu(S) - \nu(S)|.$$

The ∞-Wasserstein distance between distributions captures the largest discrepancy between samples with probability 1 under the optimal coupling. Unlike other Wasserstein distances that consider expected transport costs, it offers a worst-case perspective. Though the ∞-Wasserstein distance can be defined with a general metric, for clarity, we focus on its ℓ_q-norm version in this paper.

Definition 13 (Restatement of Definition 5) *The ∞-Wasserstein distance between distributions μ and ν on a separable Banach space $(\Theta, \|\cdot\|_q)$ is defined as*

$$W_\infty^{\ell_q}(\mu, \nu) \doteq \inf_{\gamma\in\Gamma_c(\mu,\nu)}\operatorname*{ess\,sup}_{(x,y)\sim\gamma}\|x - y\|_q = \inf_{\gamma\in\Gamma_c(\mu,\nu)}\{\alpha \mid \mathbb{P}_{(x,y)\sim\gamma}[\|x - y\|_q \leq \alpha] = 1\},$$

where $\Gamma_c(\mu, \nu)$ is the set of all couplings of μ and ν, i.e., the set of all joint probability distributions γ with marginals μ and ν respectively. The expression $\operatorname{ess\,sup}_{(x,y)\sim\gamma}$ denotes the essential supremum with respect to measure γ. By [GS84, Proposition 1], the infimum in this definition is attainable, i.e., there exists $\gamma^* \in \Gamma_c(\mu, \nu)$ such that $W_\infty^{\ell_q}(\mu, \nu) = \operatorname{ess\,sup}_{(x,y)\sim\gamma^*}\|x - y\|_q$.

For Laplace perturbation (Lemma 15), we require a bound on $W_\infty^{\ell_1}$, the Wasserstein distance defined by the ℓ_1-norm, i.e., for $q = 1$. A bound for $W_\infty^{\ell_1}$ can be derived from $W_\infty^{\ell_q}$ using the inequality $\|\cdot\|_1 \leq d^{1-\frac{1}{q}}\|\cdot\|_q$, which gives $W_\infty^{\ell_1}(\cdot,\cdot) \leq d^{1-\frac{1}{q}}W_\infty^{\ell_q}(\cdot,\cdot)$.

A.2 Lemmas for DP Analysis of Algorithm 1

We now introduce the three lemmas for the privacy analysis: the equivalence definition of (ε, δ)-DP (Lemma 14), the Laplace perturbation (Lemma 15), and the weak triangle inequality for ∞-Rényi divergence (Lemma 16).

Lemma 14 (Lemma 3.17 of [DR+14]) *A randomized mechanism $\mathcal{M}$ satisfies (ε, δ)-DP if and only if for all neighboring datasets $D \simeq D'$, there exist probability measures P, P' such that $d_{\mathrm{TV}}(\mathcal{M}(D), P) \leq \frac{\delta}{e^\varepsilon+1}$, $d_{\mathrm{TV}}(\mathcal{M}(D'), P') \leq \frac{\delta}{e^\varepsilon+1}$, $D_\infty(P \parallel P') \leq \varepsilon$, and $D_\infty(P' \parallel P) \leq \varepsilon$.*

Lemma 15 (Laplace perturbation, adapted from Theorem 3.2 of [SWC17]) *Let μ and ν be probability distributions on $\mathbb{R}^d$. Let $\mathrm{Laplace}^{\otimes \mathrm{d}}(b)$ denote the distribution of $\mathbf{z} \in \mathbb{R}^d$, where $z_i \overset{i.i.d.}{\sim} \mathrm{Laplace}(b)$. If $W_\infty^{\ell_1}(\mu, \nu) \leq \Delta$, then we have*

$$D_\infty\left(\mu * \mathrm{Laplace}^{\otimes \mathrm{d}}\left(\tfrac{\Delta}{\varepsilon}\right) \| \nu * \mathrm{Laplace}^{\otimes \mathrm{d}}\left(\tfrac{\Delta}{\varepsilon}\right)\right) \leq \varepsilon, \text{ and}$$
$$D_\infty\left(\nu * \mathrm{Laplace}^{\otimes \mathrm{d}}\left(\tfrac{\Delta}{\varepsilon}\right) \| \mu * \mathrm{Laplace}^{\otimes \mathrm{d}}\left(\tfrac{\Delta}{\varepsilon}\right)\right) \leq \varepsilon.$$

The proof is provided in Appendix F for completeness. The Laplace mechanism is a special case of Lemma 15 by setting μ and ν to Dirac distributions at $f(D)$ and $f(D')$, respectively. Lemma 15 can also be derived from the limit case of [FMTT18, Lemma 20] as the Rényi order approaches infinity.

Lemma 16 (Weak Triangle Inequality, adapted from [LSS14], Lemma 4.1) *Let μ, ν, π be probability measures on a measurable space $(\Theta, \mathcal{F})$. If $D_\infty(\mu\|\pi) < \infty$ and $D_\infty(\pi\|\nu) < \infty$, then $D_\infty(\mu\|\nu) \leq D_\infty(\mu\|\pi) + D_\infty(\pi\|\nu)$.*

The proof is deferred to Appendix F. This lemma generalizes [LSS14, Lemma 4.1], which focuses on discrete distributions. Additionally, Lemma 16 corresponds to the infinite Rényi order limit of [Mir17, Proposition 11] and [MMR12, lemma 12].

B Supplementary Discussion on Randomized Post-Processing

In this section,we clarify the distinction between our "purification" process and the randomized post-processing in [B+51, Theorem 10]. Define the following function

$$f_{\varepsilon,\delta}(\alpha) = \max\left\{0, 1-\delta-e^{\varepsilon}\alpha, e^{-\varepsilon}(1-\delta-\alpha)\right\}.$$

By [WZ10], a mechanism $\mathcal{M}$ is (ε, δ)-DP if and only if $\mathcal{M}$ is f_ε-DP.

[B+51, Theorem 10] [DRS22, Theorem 2.10] establish the existence of a (randomized) post-processing method that transforms a pair of distributions into another pair with a dominating trade-off function. Specifically, they show that if $T(P, Q) \leq T(P', Q')$, then there exists a randomized algorithm Proc such that $\mathrm{Proc}(P) = P'$ and $\mathrm{Proc}(Q) = Q'$, where T denotes the trade-off function [DRS22, Definition 2.1]. Their proof constructs a sequence of transformations and takes the limit. In contrast, our "purification" process provides a computationally efficient post-processing method for a different problem. Given that $f_{\varepsilon,\delta} \leq f_{(\varepsilon+\varepsilon'),0}$ (see Figure 3) and that $T(\mathcal{M}(D), \mathcal{M}(D')) \geq f_{\varepsilon,\delta}$ for all neighboring datasets $D \simeq D'$, we seek a randomized post-processing procedure $\mathcal{A}_{\text{pure}}$ such that $T(\mathcal{A}_{\text{pure}} \circ \mathcal{M}(D), \mathcal{A}_{\text{pure}} \circ \mathcal{M}(D')) \geq f_{(\varepsilon+\varepsilon'),0}$ while maintaining utility guarantee.

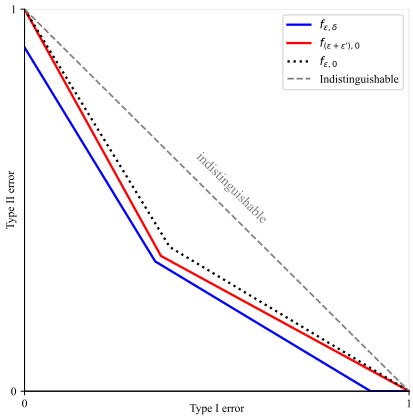

Figure 3: Trade-off functions for (ε, δ)-DP, $(\varepsilon, 0)$-DP, and $(\varepsilon + \varepsilon', 0)$-DP. Our method provides a solution to post-process the (ε, δ)-DP distribution pair (in blue) to the $(\varepsilon + \varepsilon', 0)$-DP pair (in red).

C Discussion: Purification on Finite Output Spaces

This section discusses a "folklore" method for converting approximate DP mechanisms into pure DP when the output space is *finite*, and explains why this method fails in the *continuous* case. Given an

(ε, δ)-DP mechanism $\mathcal{M} : \mathcal{X}^* \to \mathcal{Y}$ with finite output space $\mathcal{Y}$, one can construct a new mechanism by mixing $\mathcal{M}$ with uniform distribution over $\mathcal{Y}$: $\mathcal{A}_{\text{mix}}(\mathcal{M}) = (1-\omega)\mathcal{M} + \omega\text{Unif}(\mathcal{Y})$. The resulting mechanism $\mathcal{A}_{\text{mix}}(\mathcal{M})$ satisfies $(\varepsilon + \ln(1 + \frac{\delta|\mathcal{Y}|}{\omega}e^{-\varepsilon}))$-pure DP [HC22, Lemma 3.2]. However, this strategy does not yield pure DP when the output space is continuous, as the following counterexample shows.

Example 17 *Let $f : D \to [0,1]$ be a statistic computed on a dataset $D \in \mathcal{X}^*$. Consider the following mechanism: with probability δ, output the true value $f(D)$; with probability $1-\delta$, output a value uniformly from $[0,1]$. That is, $\mathcal{M}(D) = \delta \cdot \boldsymbol{\delta}^{\text{Dirac}}_{f(D)} + (1-\delta)\text{Unif}([0,1])$, where $\boldsymbol{\delta}^{\text{Dirac}}_{f(D)}$ denotes the Dirac measure at $f(D)$. Then $\mathcal{M}$ satisfies $(0,\delta)$-DP. Now, consider applying the uniform mixture strategy:*

$$\begin{aligned}\mathcal{A}_{\text{mix}}(\mathcal{M}(D)) &= (1-\omega)\mathcal{M}(D) + \omega\text{Unif}([0,1]) \\ &= (1-\omega)\delta \cdot \boldsymbol{\delta}^{\text{Dirac}}_{f(D)} + (1-\delta-\delta\omega)\text{Unif}([0,1]).\end{aligned}$$

This new mechanism satisfies $(0,(1-\omega)\delta)$-DP, but not pure DP unless $\omega = 1$, which is the trivial mixture.

D Discussion: Extending the Purification on Finite Output Spaces to the Euclidean Space

A natural approach for purification on Euclidean space is to quantize the continuous output space into a finite discrete set, and apply the folklore purification for finite sets. Specifically, to purify the (ε, δ)-DP mechanism $\mathcal{M} : \mathcal{X}^* \to \Theta \subseteq \mathbb{R}^d$, one could first cover the range Θ by a Δ-net, round the output of $\mathcal{M}$ to the nearest grid point, and then purify it using the finite range folklore method. We thank an anonymous NeurIPS 2025 reviewer for suggesting this approach. The main challenge of this approach is the explicit construction of the Δ-net and the efficient uniform sampling from it. When constructing the Δ-net by the cubes (as detailed below), this natural method achieves a similar utility bound as Algorithm 1, as the bottleneck of both bounds is the term $\left(\frac{\delta}{\omega}\right)^{\frac{1}{d}}$. The bound is different in the *non-dominating* terms: if Θ is an ℓ_∞ ball, this quantization-based method has tighter bound; if Θ is an ℓ_1 ball, it yields a worse bound.

Construction of Δ-net. We first note that a naive approach, sampling from an ℓ_1 ball (or ℓ_2 ball) and then rounding to the nearest grid point, does not yield a uniform distribution on the grid points. This non-uniformity arises from boundary effects. To illustrate, consider the ℓ_1 ball of radius R and take the point $x = (R, 0, \ldots, 0)$ on its surface. That the volume of the intersection:

$$\{z : \|z\|_1 \le R,\ \|z - x\|_1 \le r\},$$

is only a $\frac{1}{2^d}$ fraction of the ball $\{z : \|z - x\|_1 \le r\}$.

We now provide the derivation. Writing the intersection in coordinates:

$$\sum_{i=1}^{d} |z_i| \le R,\ |z_1 - R| + \sum_{i=2}^{d} |z_i| \le r,$$

we get $R - r \le z_1 \le R$, and for such z_1, the other coordinates must satisfy

$$\sum_{i=2}^{d} |z_i| \le \min\{R - z_1, r - R + z_1\}$$

Therefore, the intersection volume is $\int_{R-r\le z_1\le R} \int_{\sum_{i=2}^d |z_i| \le \min\{R-z_1, r-R+z_1\}} \mathbf{1} d(z_2, \ldots, z_d) dz_1$, $= \int_{R-r\le z_1\le R} \frac{2^{d-1}}{(d-1)!}(\min\{R - z_1, r - R + z_1\})^{d-1} dz_1 = \frac{2^{d-1}}{(d-1)!}(\int_{R-r\le z_1\le R-r/2}(r - R + z_1)^{d-1} dz_1 + \int_{R-r/2\le z_1\le R}(R - z_1)^{d-1}) dz_1 = \frac{2^{d-1}}{(d-1)!} \cdot 2\frac{1}{d}(\frac{r}{2})^d = \frac{1}{d!}r^d$, which is $\frac{1}{2^d}$ portion of the volume of the ball $\{z : \|z - x\|_1 \le r\}$.

We also note that the above intersection volume can be lower bounded for ℓ_2 norm, see Lemma 24 in [LMW+24].

Given these challenges with non-uniformity, we instead consider constructing the Δ-net by tiling a large cube $\Theta' \supseteq \Theta$ with smaller cubes.

Cube-Based Δ-net Construction and Analysis. Without loss of generality, assume $\Theta' = [0, R]^d$. Take $\Delta = 2R\left(\frac{\delta}{2\omega}\right)^{\frac{1}{d}} \log d$, and $K = 2\lfloor \frac{R}{\Delta} \rfloor$. Construct the Δ-net using cubes of the form $\prod_{k=1}^{d} \left[\frac{i_k R}{K}, \frac{(i_k+1)R}{K}\right]$. By the folklore purification, this algorithm satisfies $\left(\varepsilon + \ln(1 + \frac{\delta K^d}{\omega} e^{-\varepsilon})\right)$-pure DP. We have

$$\varepsilon + \ln(1 + \frac{\delta K^d}{\omega} e^{-\varepsilon}) \le \varepsilon + \frac{\delta K^d}{\omega} e^{-\varepsilon} \le \varepsilon + \frac{\delta \left(\frac{R}{R\left(\frac{\delta}{2\omega}\right)^{\frac{1}{d}} \log d}\right)^d}{\omega} e^{-\varepsilon} \le 2\varepsilon.$$

When $q = \infty$, this method achieves the utility bound of $\mathbb{E} \left\| x_{\text{pure}} - x_{\text{apx}} \right\|_\infty \le \omega R + \frac{4R \log d}{\varepsilon'} \left(\frac{\delta}{2\omega}\right)^{\frac{1}{d}}$.

This bound is tighter than the bound by directly applying Algorithm 1 in the dependence in d. For comparison, directly extending the proof of Theorem 1 (Appendix F.4) to the ℓ_∞ norm yields: $\mathbb{E} \left\| x_{\text{pure}} - x_{\text{apx}} \right\|_\infty \le \omega R + \frac{4Rd(\log d + 1)}{\varepsilon'} \left(\frac{\delta}{2\omega}\right)^{\frac{1}{d}}$, which is derived as follows:

$$\begin{aligned}
\mathbb{E} \left\| x_{\text{pure}} - x_{\text{apx}} \right\|_\infty &= \mathbb{P}(u > \omega) \mathbb{E}\left[\left\| x_{\text{pure}} - x_{\text{apx}} \right\|_\infty \mid u > \omega\right] + \mathbb{P}(u \le \omega) \mathbb{E}\left[\left\| x_{\text{pure}} - x_{\text{apx}} \right\|_\infty \mid u \le \omega\right] \\
&\le \mathbb{E}\left[\max_i |z_i|\right] + \omega R \qquad (z_i \sim \text{Laplace}\left(\tfrac{2\Delta}{\varepsilon'}\right)) \\
&= \left(\sum_{k=1}^{d} \frac{1}{k}\right) \frac{2\Delta}{\varepsilon'} + \omega R \\
&\le (\log d + 1) \frac{2\Delta}{\varepsilon'} + \omega R \\
&\le \frac{4Rd(\log d + 1)}{\varepsilon'} \left(\frac{\delta}{2\omega}\right)^{\frac{1}{d}} + \omega R.
\end{aligned}$$

When $q = 1$, applying the standard conversion $\|\cdot\|_1 \le d \|\cdot\|_\infty$, this method achieves the utility bound of $\mathbb{E} \left\| x_{\text{pure}} - x_{\text{apx}} \right\|_1 \le \omega d R + \frac{4Rd \log d}{\varepsilon'} \left(\frac{\delta}{2\omega}\right)^{\frac{1}{d}}$, which is worse than Theorem 1 by a factor of d.

E Privacy Analysis: Proof Sketch of Theorem 1

The proof sketch of Theorem 1 is illustrated in Figure 4. Let D and D' be neighboring datasets, and let $\mathcal{M}$ be an (ε, δ)-DP mechanism. We aim to show that $\mathcal{A}_{\text{pure}}(\mathcal{M}(D))$ and $\mathcal{A}_{\text{pure}}(\mathcal{M}(D'))$—the post-processed outputs of $\mathcal{M}(D)$ and $\mathcal{M}(D')$ after applying Algorithm 1—are ε-indistinguishable. By sketching the proof, we also provide the intuition of the design of Algorithm 1.

First, by the equivalent definition of approximate DP (Lemma 14), there exists a hypothetical ε-indistinguishable distribution pair P and P', such that $\mathcal{M}(D)$ and $\mathcal{M}(D')$ are $\mathcal{O}(\delta)$-close to P and P', respectively, in total variation distance. Note that P and P' can both depend on D and D', i.e., $P = P(D, D')$ and $P' = P'(D, D')$, rather than simply $P = P(D)$ and $P' = P'(D')$. This dependence complicates the direct application of standard DP analysis. To address this, we transition to a distributional perspective.

To show that $\mathcal{A}_{\text{pure}}(\mathcal{M}(D))$ and $\mathcal{A}_{\text{pure}}(\mathcal{M}(D'))$ are ε-indistinguishable distributions, by the weak triangle inequality of ∞-Rényi divergence (Lemma 16), it suffices to show that $\mathcal{A}_{\text{pure}}(\mathcal{M}(D))$ and $\mathcal{A}_{\text{pure}}(P)$ are ε-indistinguishable (in terms of ∞-Rényi divergence), and that $\mathcal{A}_{\text{pure}}(\mathcal{M}(D'))$ and $\mathcal{A}_{\text{pure}}(P')$ are ε-indistinguishable as well. Now the problem reduces to: given two distributions with total variation distance bound, how to post-process them (with randomness) to obtain the ∞-Rényi divergence bound?

A natural way to establish the ∞-Renyi bound is via the Laplace mechanism, which perturbs *deterministic* variables with Laplace noise according to the ℓ_1-sensitivity. To generalize the Laplace mechanism to perturbing *random* variables, we develop the Laplace perturbation (Lemma 15), which achieves the ∞-Rényi bound by convolving Laplace noise according to the $W_\infty^{\ell_1}$ distance. The $W_\infty^{\ell_1}$ distance can be viewed as a randomized analog of the ℓ_1-distance. The remaining step is to derive a $W_\infty^{\ell_1}$ bound from the TV distance bound, and this motivates Lemma 6, a d_{TV} to $W_\infty^{\ell_q}$ conversion

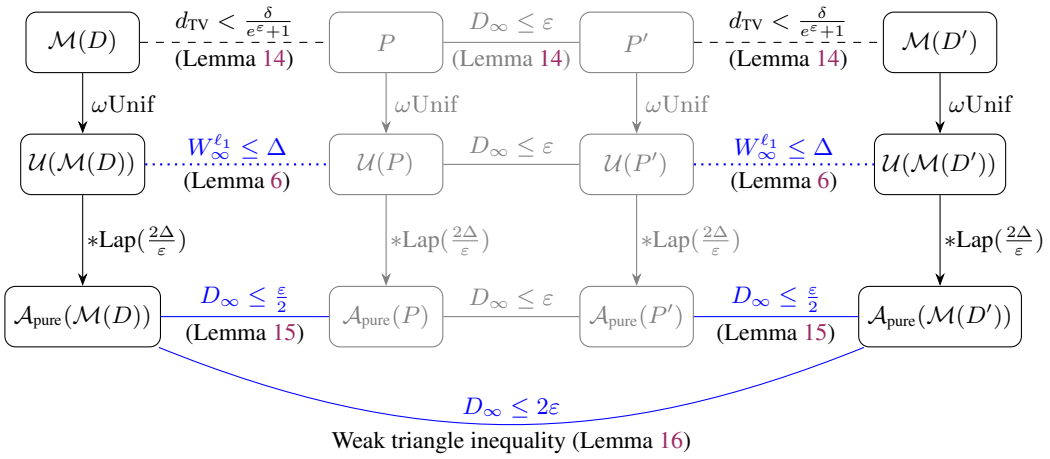

Figure 4: Flowchart illustrating the proof sketch of Theorem 1 and the intuition behind Algorithm 1. The notation $D_\infty \le \varepsilon$ is an abbreviation for the pair of inequalities $D_\infty(\mu\|\nu) \le \varepsilon$ and $D_\infty(\nu\|\mu) \le \varepsilon$, where μ and ν correspond to the two end nodes of the respective edges (e.g., P and P'). The symbol ωUnif represents a mixture with the uniform distribution (Algorithm 1, Line 3), where $\mathcal{U}(\cdot) = (1-\omega)\cdot + \omega\mathrm{Unif}(\Theta)$. The notation $*$Lap refers to the convolution with the Laplace distribution, as in Algorithm 1, Line 4.

lemma that generalizes [LMW+24, Lemma 8]. A small mixture of a uniform distribution ensures that the conditions of Lemma 6 hold. Therefore, Algorithm 1 consists of two key steps: mixture with the uniform distribution (Line 3), and the Laplace perturbation calibrated to δ (Line 4.) The formal proof of Theorem 1, along with proofs of the key lemmas, is deferred to Appendix F.

F Deferred Proofs in Section 2 and Appendix A

F.1 Proof of Lemma 6

Proof of Lemma 6 Denote $\mathrm{dist}(x,y) = \|x-y\|_q$. Denote $\mathcal{B}(c,r)$ the ball with center c, and ℓ_q-radius r. Assume $\mathcal{B}(c,r) \in \Theta$. Without loss of generality, assume $\Delta < R$.

Fix Δ and set $\xi = p_{\min} \cdot \mathrm{Vol}\left(\mathbb{B}^d_{\ell_q}(1)\right) \cdot \left(\frac{r}{4R}\right)^d \cdot \Delta^d$, so that $d_{\mathrm{TV}}(\mu,\nu) < \xi$.

To prove the result that $W_\infty \le \Delta$, we use the equivalent definition of W_∞ in Lemma 7. By this definition, to prove $W_\infty(\mu,\nu) \le \Delta$, it suffices to show that

$$\mu(A) \le \nu(A^\Delta), \text{ for all open set } A \subseteq \Theta,$$

where we re-define $A^\Delta := \{x \in \mathbb{R}^d \mid \mathrm{dist}(x,A) \le \Delta\}$. Note that by this definition, A^Δ might extend beyond Θ. However, we still have $\nu(A^\Delta) = \nu(A^\Delta \cap \Theta)$, since ν is supported on Θ.

Note that if $\nu(A^\Delta) = 1$, it is obvious that $\mu(A) \le \nu(A^\Delta)$. When A is an empty set, the proof is trivial. So we only consider nonempty open set $A \subseteq \Theta$ with $\nu(A^\Delta) < 1$.

Note that for an arbitrary open set $A \subseteq \Theta$, we have

$$\mu(A) \le \nu(A) + d_{TV}(\mu,\nu) < \nu(A) + \xi.$$

Thus to prove $\mu(A) \le \nu(A^\Delta)$, it suffices to prove $\nu(A) + \xi \le \nu(A^\Delta)$, i.e., $\nu(A^\Delta \setminus A) \ge \xi$.

To prove $\nu(A^\Delta \setminus A) \ge \xi$, we construct a ball $K \subset \mathbb{R}^d$ of radius $\frac{r\Delta}{4R}$ satisfying the properties:

Property 1 K is contained in the set $A^\Delta \setminus A$, i.e., $K \subseteq A^\Delta \setminus A$, and

Property 2 K is contained in Θ, i.e., $K \subseteq \Theta$. This guarantees that $\nu(K) = \nu(K \cap \Theta) \ge p_{\min} \cdot \mathrm{Vol}(K) = \xi$.

To construct the above ball K, we adopt the following strategy:

Step 1 First, select a point $y \in \Theta$ such that $\mathrm{dist}(y,A) = \Delta/2$.

Step 2 Then, construct the ball K with center $c_1 = \omega c + (1-\omega)y$, radius ωr, where $\omega = \frac{\Delta}{4R}$[3], i.e., $K = \mathcal{B}(c_1, \omega r)$. We will later show that this ball is contained in the convex hull of $\mathcal{B}(c, r) \cup \{y\}$. This construction is inspired by [MV22].

We first prove such point y in Step 1 exists, that is, the set $\Theta \cap \{x \in \mathbb{R}^d | \mathrm{dist}(x, A) = \Delta/2\}$ is nonempty. (Note that we only consider nonempty open set $A \subseteq \Theta$ with $\nu(A^\Delta) < 1$.)

We prove it by contradiction. If instead $\Theta \cap \{x \in \mathbb{R}^d | \mathrm{dist}(x, A) = \Delta/2\} = \emptyset$, then for all $x \in \Theta$, $\mathrm{dist}(x, A) \neq \Delta/2$. Due to the continuity of dist and the convexity of Θ, we know that only one of these two statements holds:

- $\mathrm{dist}(x, A) < \Delta/2$, for all $x \in \Theta$.
- $\mathrm{dist}(x, A) > \Delta/2$, for all $x \in \Theta$.

Since $\emptyset \neq A \subseteq \Theta$, there exist $x' \in A \in \Theta$, such that $\mathrm{dist}(x', A) = 0$. Therefore, the first statement holds. Thus $\Theta \subseteq A^{\Delta/2} \subseteq A^\Delta$, which contradicts $\nu(A^\Delta) < 1$.

Therefore, there exist $y \in \Theta$, such that $\mathrm{dist}(y, A) = \Delta/2$, making Step 1 valid.

Next, we show that the K we construct in Step 2 satisfies Property 1 and Property 2.

To prove $K \subseteq A^\Delta \setminus A$, let $x \in K = \mathcal{B}(c_1, \omega r)$. We show that $x \in A^\Delta$ and $x \notin A$. We have

$$\begin{aligned} \mathrm{dist}(x, y) &\leq \mathrm{dist}(x, c_1) + \mathrm{dist}(c_1, y) \\ &= \|x - c_1\|_q + \|\omega c + (1-\omega)y - y\|_q \\ &= \|x - c_1\|_q + \omega \|c - y\|_q \\ &\leq \omega r + \omega R \\ &= \frac{\Delta}{4R}(r + R) \\ &\leq \Delta/2 \end{aligned} \tag{3}$$

- ($x \notin A$). If $x \in A$, since A is an open set and $\mathrm{dist}(y, A) = \Delta/2$, we have that $\mathrm{dist}(x, y) > \Delta/2$, which contradicts to (3). Therefore $x \notin A$.
- ($x \in A^\Delta$). We have $\mathrm{dist}(x, A) \leq \mathrm{dist}(x, y) + \mathrm{dist}(y, A) \overset{(3)}{\leq} \Delta$, implying that $x \in A^\Delta$.

To prove $K \subseteq \Theta$, take any $x \in K = \mathcal{B}(c_1, \omega r)$, we show that $x \in \Theta$. Write $x = c_1 + \omega r \mathbf{v}$, where $\|\mathbf{v}\|_q \leq 1$. We have

$$x = \omega c + (1-\omega)y + \omega r \mathbf{v} = \omega(c + r\mathbf{v}) + (1-\omega)y.$$

Since $c + r\mathbf{v} \in \mathcal{B}(c, r) \subseteq \Theta$, $y \in \Theta$, and $0 < \omega = \frac{\Delta}{4R} < 1$, by the convexity of Θ, we have $x \in \Theta$, which completes the proof.

In particular, if Θ is an ℓ_q-ball centered at c, say $\Theta = c + \mathbb{B}^d_{\ell_q}(r)$, then the ω in Step 2 can be chosen as $\omega = \frac{\Delta}{4r}$, which similarly follows, since $\|c - x\|_p \leq r$ for any $x \in \Theta$. This improves the conversion by:

$$\text{If} \quad d_{\mathrm{TV}}(\mu, \nu) < p_{\min} \cdot \mathrm{Vol}(\mathbb{B}^d_{\ell_q}(1)) \cdot \left(\frac{1}{4}\right)^d \cdot \Delta^d, \quad \text{then} \quad W_\infty(\mu, \nu) \leq \Delta. \tag{4}$$

■

F.2 Proof of Lemma 15

Proof of Lemma 15 denote $P = \mu * \mathrm{Laplace}^{\otimes \mathrm{d}}\left(\frac{\Delta}{\varepsilon}\right)$, $Q = \nu * \mathrm{Laplace}^{\otimes \mathrm{d}}\left(\frac{\Delta}{\varepsilon}\right)$. Then, P and Q are absolutely continuous with respect to the Lebesgue measure. This is because for any Lebesgue zero measure set S, $P(S) = \int_x \mathbb{P}_{\mathrm{Laplace}}(S - x) \mathrm{d}\mu(x)$, where $S - x \doteq \{y | x + y \in S\}$, and $\mathbb{P}_{\mathrm{Laplace}}$ denotes the probability measure of $\mathrm{Laplace}^{\otimes \mathrm{d}}\left(\frac{\Delta}{\varepsilon}\right)$. Since $S - x$ is zero measure for all $x \in \mathbb{R}^d$, and $\mathbb{P}_{\mathrm{Laplace}}$ is absolutely continuous w.r.t. the Lebesgue measure, we have $\mathbb{P}_{\mathrm{Laplace}}(S - x)$ for all

[3]We note that the symbol ω in Section F.1 is distinct from the ω used in other parts of the paper. This distinction is an intentional abuse of notation for clarity within specific contexts.

$x \in \mathbb{R}^d$. Therefore, $P(S) = 0$. Thus P is absolutely continuous w.r.t. the Lebesgue measure, and Q similarly follows.

Denote p and q the probability density function of P and Q respectively. Since $W_\infty^{\ell_1}(\mu, \nu) \le \Delta$, by [GS84, Proposition 1], there exists $\gamma^* \in \Gamma_c(\mu, \nu)$, such that $\mathbb{P}_{(u,v)\sim\gamma^*}[\|u - v\|_1 > \Delta] = \gamma^*\{\|u - v\|_1 > \Delta\} = 0$. By the definition of the max divergence (Theorem 6 of [VEH14]), we have

$$
\begin{aligned}
e^{D_\infty(P\|Q)} &= \operatorname*{ess\,sup}_{x\sim P} \frac{p(x)}{q(x)} && \text{([VEH14, Theorem 6])} \\
&\le \operatorname*{ess\,sup}_{x\sim P} \frac{\int_{\mathbb{R}^d} e^{-\frac{\|x-u\|_1}{b}} \mathrm{d}\mu(u)}{\int_{\mathbb{R}^d} e^{-\frac{\|x-v\|_1}{b}} \mathrm{d}\nu(v)} && \text{(Convolution theorem, PDF of Laplace)} \\
&\le \operatorname*{ess\,sup}_{x\sim P} \frac{\int_{\mathbb{R}^d\times\mathbb{R}^d} e^{-\frac{\|x-u\|_1}{b}} \mathrm{d}\gamma^*(u,v)}{\int_{\mathbb{R}^d\times\mathbb{R}^d} e^{-\frac{\|x-v\|_1}{b}} \mathrm{d}\gamma^*(u,v)} && (\gamma^* \in \Gamma_c(\mu,\nu)) \\
&\le \operatorname*{ess\,sup}_{x\sim P} \frac{\int_{\mathbb{R}^d\times\mathbb{R}^d} e^{-\frac{\|x-v\|_1}{b}+\frac{\|u-v\|_1}{b}} \mathrm{d}\gamma^*(u,v)}{\int_{\mathbb{R}^d\times\mathbb{R}^d} e^{-\frac{\|x-v\|_1}{b}} \mathrm{d}\gamma^*(u,v)} && \text{(triangle's inequality)} \\
&\le \operatorname*{ess\,sup}_{x\sim P} \frac{\int_{\|u-v\|_1\le\Delta} e^{-\frac{\|x-v\|_1}{b}+\frac{\|u-v\|_1}{b}} \mathrm{d}\gamma^*(u,v) + \int_{\|u-v\|_1>\Delta} e^{-\frac{\|x-v\|_1}{b}+\frac{\|u-v\|_1}{b}} \mathrm{d}\gamma^*(u,v)}{\int_{\mathbb{R}^d\times\mathbb{R}^d} e^{-\frac{\|x-v\|_1}{b}} \mathrm{d}\gamma^*(u,v)} \\
&\le \operatorname*{ess\,sup}_{x\sim P} \frac{\int_{\|u-v\|_1\le\Delta} e^{-\frac{\|x-v\|_1}{b}+\frac{\|u-v\|_1}{b}} \mathrm{d}\gamma^*(u,v)}{\int_{\mathbb{R}^d\times\mathbb{R}^d} e^{-\frac{\|x-v\|_1}{b}} \mathrm{d}\gamma^*(u,v)} && (\gamma^*\{\|u-v\|_1 > \Delta\} = 0) \\
&\le \operatorname*{ess\,sup}_{x\sim P} \frac{e^{\frac{\Delta}{b}} \int_{\|u-v\|_1\le\Delta} e^{-\frac{\|x-v\|_1}{b}} \mathrm{d}\gamma^*(u,v)}{\int_{\mathbb{R}^d\times\mathbb{R}^d} e^{-\frac{\|x-v\|_1}{b}} \mathrm{d}\gamma^*(u,v)} \\
&\le \operatorname*{ess\,sup}_{x\sim P} \frac{e^{\frac{\Delta}{b}} \int_{\mathbb{R}^d\times\mathbb{R}^d} e^{-\frac{\|x-v\|_1}{b}} \mathrm{d}\gamma^*(u,v)}{\int_{\mathbb{R}^d\times\mathbb{R}^d} e^{-\frac{\|x-v\|_1}{b}} \mathrm{d}\gamma^*(u,v)} \\
&= e^{\frac{\Delta}{b}}
\end{aligned}
$$

■

F.3 Proof of Lemma 16

Proof of Lemma 16 Since $D_\infty(\mu\|\pi) < \infty$, for any measurable set $S \in \mathcal{F}$, such that $\mu(S) > 0$, we have $\pi(S) > 0$. That is, $\{S \in \mathcal{F} \mid \mu(S) > 0\} \subseteq \{S \in \mathcal{F} \mid \pi(S) > 0\}$. By Definition 11,

$$
\begin{aligned}
D_\infty(\mu,\nu) &= \ln \sup_{S\in\mathcal{F},\ \mu(S)>0} \left[\frac{\mu(S)}{\nu(S)}\right] \\
&= \ln \sup_{S\in\mathcal{F},\ \mu(S)>0,\ \pi(S)>0} \left[\frac{\mu(S)\pi(S)}{\pi(S)\nu(S)}\right] \\
&\le \ln \sup_{S\in\mathcal{F},\ \mu(S)>0} \left[\frac{\mu(S)}{\pi(S)}\right] + \ln \sup_{S\in\mathcal{F},\ \pi(S)>0} \left[\frac{\pi(S)}{\nu(S)}\right] \\
&= D_\infty(\mu\|\pi) + D_\infty(\pi\|\nu).
\end{aligned}
$$

■

F.4 Proof of Theorem 1 and Corollary 4

We first provide proof of the privacy guarantee by reorganizing the proof sketch in Section E, as illustrated in Figure 4.

Let D and D' be neighboring datasets, and let $\mathcal{M}$ be an (ε, δ)-DP mechanism as in Section E. By the equivalence definition of approximate DP (Lemma 14), there exists a hypothetical distribution pair

P, P' such that

$$d_{\mathrm{TV}}(\mathcal{M}(D), P) \le \frac{\delta}{e^\varepsilon + 1}, d_{\mathrm{TV}}(\mathcal{M}(D'), P') \le \frac{\delta}{e^\varepsilon + 1}, D_\infty(P' \| P) \le \varepsilon, \text{ and } D_\infty(P \| P') \le \varepsilon$$

Note that P and P' can both depend on D and D', i.e., $P = P(D, D')$ and $P' = P'(D, D')$, rather than simply $P = P(D)$ and $P' = P'(D')$.

Denote $\mathcal{U}(\cdot) = (1 - \omega) \cdot + \omega \mathrm{Unif}(\Theta)$. After the uniform mixture step (Line 3), the distributions $\mathcal{U}(\mathcal{M}(D)), \mathcal{U}(\mathcal{M}(D')), \mathcal{U}(P), \mathcal{U}(P')$ all satisfy the assumption in Lemma 6 with

$$p_{\min} \ge \frac{\omega}{\mathrm{Vol}(\Theta)} \ge \frac{\omega}{(R/2)^d \mathrm{Vol}\left(\mathbb{B}^d_{\ell_q}(1)\right)}.$$

Therefore, by Lemma 6 (with the case of Θ is an ℓ_q ball) and the fact that $\|\cdot\|_1 \le d^{1-\frac{1}{q}} \|\cdot\|_q$, we have

$$W^{\ell_1}_\infty(\mathcal{U}(\mathcal{M}(D)), \mathcal{U}(P)) \le \Delta, \text{ and } W^{\ell_1}_\infty(\mathcal{U}(\mathcal{M}(D')), \mathcal{U}(P')) \le \Delta,$$

where Δ is defined in Line 2 in Algorithm 1. Therefore, by Lemma 15, we have

$$\begin{aligned}
D_\infty\left(\mathcal{U}(\mathcal{M}(D)) * \mathrm{Laplace}^{\otimes \mathrm{d}}\left(\tfrac{2\Delta}{\varepsilon'}\right) \| \mathcal{U}(P) * \mathrm{Laplace}^{\otimes \mathrm{d}}\left(\tfrac{2\Delta}{\varepsilon'}\right)\right) &\le \varepsilon'/2, \\
D_\infty\left(\mathcal{U}(P) * \mathrm{Laplace}^{\otimes \mathrm{d}}\left(\tfrac{2\Delta}{\varepsilon'}\right) \| \mathcal{U}(\mathcal{M}(D)) * \mathrm{Laplace}^{\otimes \mathrm{d}}\left(\tfrac{2\Delta}{\varepsilon'}\right)\right) &\le \varepsilon'/2, \\
D_\infty\left(\mathcal{U}(\mathcal{M}(D')) * \mathrm{Laplace}^{\otimes \mathrm{d}}\left(\tfrac{2\Delta}{\varepsilon'}\right) \| \mathcal{U}(P') * \mathrm{Laplace}^{\otimes \mathrm{d}}\left(\tfrac{2\Delta}{\varepsilon'}\right)\right) &\le \varepsilon'/2, \\
D_\infty\left(\mathcal{U}(P') * \mathrm{Laplace}^{\otimes \mathrm{d}}\left(\tfrac{2\Delta}{\varepsilon'}\right) \| \mathcal{U}(\mathcal{M}(D')) * \mathrm{Laplace}^{\otimes \mathrm{d}}\left(\tfrac{2\Delta}{\varepsilon'}\right)\right) &\le \varepsilon'/2.
\end{aligned}$$

Finally, with the weak triangle inequality (Lemma 16), we conclude that the output of Algorithm 1 satisfies $(\varepsilon + \varepsilon')$-DP.

Next, we provide the utility bound. Let $u \sim \mathrm{Unif}(0, 1)$ in Line 3 of Algorithm 1.

$$\mathbb{E}\left\|x_{\text{pure}} - x_{\text{apx}}\right\|_q = \mathbb{P}(u > \omega)\mathbb{E}\left[\left\|x_{\text{pure}} - x_{\text{apx}}\right\|_q \mid u > \omega\right] + \mathbb{P}(u \le \omega)\mathbb{E}\left[\left\|x_{\text{pure}} - x_{\text{apx}}\right\|_q \mid u \le \omega\right] \tag{5}$$

$$\begin{aligned}
&\le \mathbb{E}\left[\left(\sum_{i=1}^d |z_i|^q\right)^{\frac{1}{q}}\right] + \omega R && (z_i \sim \mathrm{Laplace}\left(\tfrac{2\Delta}{\varepsilon'}\right)) \\
&\le \left(\mathbb{E}\left[\sum_{i=1}^d |z_i|^q\right]\right)^{\frac{1}{q}} + \omega R && \text{(Jensen's inequality)} \\
&= (d\Gamma(q+1))^{\frac{1}{q}} \frac{2\Delta}{\varepsilon'} + \omega R && (6) \\
&\le \frac{4dqR}{\varepsilon'}\left(\frac{\delta}{2\omega}\right)^{\frac{1}{d}} + \omega R && (7)
\end{aligned}$$

Remark 18 *We can remove the dependence on q in the above bound by controlling $\|x_{\text{pure}} - x_{\text{apx}}\|_1$ via concentration inequalities for sub-exponential random vectors.*

Proof of Corollary 4 Since Θ is an ℓ_q ball, we have $R = 2r = C$. The proof follows directly from Theorem 1 by taking $q = 2$. ■

F.5 Proof of Theorem 2

Proof of Theorem 2 Notice that BinMap is a data-independent deterministic function, thus by post-processing, $z_{\text{bin}} = \mathsf{BinMap}\left(u_{\text{apx}}\right)$ maintains (ε, δ)-DP.

We consider the ℓ_1 norm, i.e., $q = 1$. Let $a = (\frac{1}{2}, \ldots, \frac{1}{2})$. For unit cube $[0, 1]^d$, we have $a + \mathbb{B}^d_{\ell_q}(\frac{1}{2}) \subseteq [0, 1]^d \subseteq a + \mathbb{B}^d_{\ell_q}(\frac{d}{2})$. That is, $[0, 1]^d$ satisfies Assumption 1 with $R = \frac{d}{2}$, $r = \frac{1}{2}$, and $q = 1$. Therefore, by Theorem 1, z_{pure} satisfies $(2\varepsilon, 0)$-DP. After the post-processing, the output of Algorithm 2 maintains $(2\varepsilon, 0)$-DP.

For the utility, by $\delta < \frac{\varepsilon^d}{(2d)^{3d}}$ and Line 2 of Algorithm 1, we have $\Delta \le \frac{\varepsilon}{4d}$.

Let $y_i^{\text{Laplace}} \overset{\text{i.i.d.}}{\sim} \text{Laplace}(2\Delta/\varepsilon')$, be the noise added in Line 4 in Algorithm 1. Then for any i,

$$\mathbb{P}(y_i^{\text{Laplace}} \ge t) = \frac{1}{2}e^{-\frac{t}{2\Delta/\varepsilon}} \le \frac{1}{2}e^{-2dt}.$$

Thus,

$$\mathbb{P}\left(y_i^{\text{Laplace}} > 0.5, \forall i = 1, \dots, d\right) = (1 - \frac{1}{2}e^{-d})^d \ge 1 - \frac{d}{2}e^{-d}, \tag{8}$$

where the last inequality is by Bernoulli's inequality.

Since the rounding function is defined as $\text{Round}_{\{0,1\}^d}(\mathbf{x}) = (\mathbf{1}(x_i \ge 0.5))_{i=1}^d$, we observe that $z_{\text{bin}} = z_{\text{pure}}$ if and only if the following conditions hold simultaneously:

1. $x \sim \text{Unif}(\Theta)$ is sampled in Line 3 of Algorithm 1.
2. For all $i \in [d]$, if $z_{\text{bin}}^{(i)} = 0$, then $y_i^{\text{Lap}} < 0.5$.
3. For all $i \in [d]$, if $z_{\text{bin}}^{(i)} = 1$, then $y_i^{\text{Lap}} > -0.5$.

By the symmetry of Laplace noise, applying Eq. (8), and using the union bound, we obtain

$$\mathbb{P}\left(z_{\text{bin}} = z_{\text{pure}}\right) \ge 1 - \omega - \frac{d}{2}e^{-d} = 1 - 2^{-d} - \frac{d}{2}e^{-d}.$$

■

G Further Discussion of the Optimality of Our Purification Results

In this section, we examine the optimality of the utility guarantees achieved by our purified algorithm, as summarized in Table 1. We compare our results against known information-theoretic lower bounds and the best existing upper bounds, discussing each setting in turn.

For the *DP-ERM* setting (Rows 1 and 2), our bounds (Theorem 3) match the lower bounds reported in [BST14, Table 1, Rows 1 and 3], up to logarithmic factors.

In the *DP-Frank-Wolfe* setting (Row 3), our guarantee matches with the lower bound established in Lemma 30, again up to logarithmic terms.

For the *PTR-type* setting (Row 4), to the best of our knowledge, no pure-DP mechanism has previously been developed in this regime, and thus no direct baseline is available for comparison.

For the *Mode Release* task (Row 5), our result (Theorem 6) matches the lower bound from [CHS14, Proposition 1], up to logarithmic factors.

For *Regression* (Row 6), assuming bounded data and parameter domains, our result (Theorem 7) agrees with the lower bounds in [BST14, Table 1] with respect to n, d, ε, and λ, up to logarithmic factors and constants depending on the data and parameter radii.

Finally, for *Query Release* (Row 7), our utility guarantee (Theorem 8) matches that of the SmallDB and Private Multiplicative Weights algorithms up to logarithmic factors [DR+14, BLR13], which represent the current state of the art for pure-DP query release. Whether this rate can be further improved remains an open question [NS21].

H Deferred Proofs for DP-SGD

The study of DP-ERM is extensive; for other notable contributions, see, e.g., [RBHT12, FTS17, INS+19, SSTT21, MBST22, GTU22, GLL22, RKW23, KLT24].

H.1 Algorithms and Notations

Let $f(\theta; x)$ represent the individual loss function. $\mathcal{L}(\theta) := \frac{1}{n}\sum_{i=1}^n f(\theta; x_i)$ and $\mathcal{L}^* = \min_{\theta \in \mathcal{C}} \mathcal{L}(\theta)$. We also denote $F(\theta) := \sum_{i=1}^n f(\theta; x_i)$ and $F^* = \min_{\theta \in \mathcal{C}} F(\theta)$. L-Lipschitz, β-smooth, or λ-strongly convex are all w.r.t. individual the loss function f. The parameter space Θ in Algorithm 4 is selected as the ℓ_2-ball with diameter C that contains $\mathcal{C}$.

Algorithm 3: Differentially Private SGD [ACG+16]

Input: Dataset $D = \{x_1, \ldots, x_n\}$, loss function $f : \mathcal{C} \times D \to \mathbb{R}$, parameters: learning rate η_t, noise scale σ, subsampling rate γ, Lipschitz constant L, parameter space $\mathcal{C} \in \mathbb{R}^d$

Initialize $\theta_0 \in \mathcal{C}$ randomly

for $t \in [T]$ **do**

  Sample a subset S_t by selecting a γ fraction of the dataset without replacement

  Compute gradient: for *each* $i \in S_t$ **do**

    $g_t(x_i) \leftarrow \nabla_\theta f(\theta_t; x_i)$

  Aggregate and add noise: $\hat{g}_t \leftarrow \frac{1}{\gamma}\left(\sum_{i \in S_t} g_t(x_i) + \sigma \mathcal{N}(0, \mathbf{I}_d)\right)$

  Descent: $\theta_{t+1} \leftarrow \mathrm{Proj}_{\mathcal{C}}\left(\theta_t - \eta_t \tilde{g}_t\right)$

Output: $\theta_{\text{out}} = \frac{1}{T}\sum_{t=1}^{T} \theta_t$ if f is convex; $\theta_{\text{out}} = \frac{2}{T(T+1)}\sum_{t=1}^{T} t\theta_t$ if f is strongly convex.

Algorithm 4: Pure DP SGD

Input: Output from DP-SGD Algorithm 3 θ_{apx}, parameter space Θ, privacy parameter ε' and δ, mixture level ω

$\theta_{\text{pure}} \leftarrow \mathcal{A}_{\text{pure}}(\theta_{apx}, \Theta, \varepsilon', \delta, \omega)$ ▷ Algorithm 1

Output: θ_{pure}

H.2 Noisy Gradient Descent Using Laplace Mechanism

Lemma 19 (Laplace Noisy Gradient Descent) *Let the loss function $f : \mathcal{X} \to \mathbb{R}^d$ be convex and L-Lipschitz with respect to $\|\cdot\|_2$ and $\max_{X \simeq X'} \|\nabla f(X) - \nabla f(X')\|_1 \le \Delta_1$. Suppose the parameter space $\mathcal{C} \subset \mathbb{R}^d$ is convex with an ℓ_2 diameter of at most C. Running **full-batch** noisy gradient descent with learning rate $\eta = \frac{C}{\sqrt{T(n^2L^2 + 2d\sigma^2)}}$, number of iterations $T = \frac{\varepsilon n L}{\Delta_1 \sqrt{d}}$, and Laplace noise with parameter $\sigma = \frac{\Delta_1 T}{\varepsilon}$, satisfies ε-DP. Moreover,*

$$\mathbb{E}_{\mathcal{A}}\left(\mathcal{L}(\bar{\theta}) - \mathcal{L}^*\right) \le \mathcal{O}\left(\frac{C\Delta_1^{1/2} L^{1/2} d^{1/4}}{n^{1/2}\varepsilon^{1/2}}\right).$$

and the total number of gradient calculation is $\frac{n^2 \varepsilon L \sqrt{d}}{\Delta_1}$. Without further assumptions on ∇f, we have:

$$\mathbb{E}\left(\mathcal{L}(\bar{\theta}) - \mathcal{L}^*\right) \le \mathcal{O}\left(\frac{CLd^{1/2}}{n^{1/2}\varepsilon^{1/2}}\right)$$

Proof Suppose we run T iterations and the final privacy budget is ε. Then, the privacy budget per iteration is $\varepsilon_0 = \frac{\varepsilon}{T}$, and the parameter of the additive Laplace noise is $\sigma = \Delta_1/\varepsilon_0 = \Delta_1 T/\varepsilon$. By [GG23, Theorem 9.6], we have

$$\begin{aligned}
\mathbb{E}\left(F\left(\frac{1}{T}\sum_{t=1}^{T}\theta_t\right) - F^*\right) &\le \frac{C^2}{T\eta} + \eta(n^2L^2 + 2d\sigma^2) \\
&\le \mathcal{O}\left(\frac{CnL}{\sqrt{T}} + \frac{C\sigma\sqrt{d}}{\sqrt{T}}\right) \\
&= \mathcal{O}\left(\frac{CnL}{\sqrt{T}} + \frac{C\Delta_1\sqrt{dT}}{\varepsilon}\right) \\
&\le \mathcal{O}\left(\frac{C(nL\Delta_1)^{1/2} d^{1/4}}{\varepsilon^{1/2}}\right)
\end{aligned}$$

where the second inequality is obtained by choosing learning rate $\eta = \frac{C}{\sqrt{T(n^2L^2 + 2d\sigma^2)}}$ and the fact $\sqrt{a+b} < \sqrt{a} + \sqrt{b}$ for any positive a and b. The last inequality is by setting $T = \frac{\varepsilon n L}{\Delta_1 \sqrt{d}}$. Divide

both sides by n, we have:

$$\mathbb{E}\left(\mathcal{L}(\bar{\theta}) - \mathcal{L}^*\right) \leq \mathcal{O}\left(\frac{C\Delta_1^{1/2} L^{1/2} d^{1/4}}{n^{1/2}\varepsilon^{1/2}}\right) \tag{9}$$

Without additional information, the best upper bound for Δ_1 is $\sqrt{d}\Delta_2 = \sqrt{d}L$. Plugging this bound to Eq. (9) yields:

$$\mathbb{E}\left(\mathcal{L}(\bar{\theta}) - \mathcal{L}^*\right) \leq \mathcal{O}\left(\frac{CLd^{1/2}}{n^{1/2}\varepsilon^{1/2}}\right)$$

■

H.3 Analysis of DP-SGD

H.3.1 Privacy Accounting Results

Our privacy accounting for DP-SGD is based on Rényi differential privacy [Mir17]. Before stating the privacy accounting result (Corollary 24), we define Rényi Differential privacy and its variant, zero concentrated Differential Privacy [BS16].

Definition 20 (Rényi differential privacy) *A randomized mechanism $\mathcal{M}$ satisfies $(\alpha, \varepsilon(\alpha))$-Rényi Differential Privacy (RDP) if for all neighboring datasets D, D' and for all $\alpha \geq 1$,*

$$D_\alpha(\mathcal{M}(D) \| \mathcal{M}(D')) \leq \varepsilon(\alpha),$$

where $D_\alpha(P\|Q)$ denotes the α-Rényi divergence when $\alpha > 1$; Kullback-Leibler divergence when $\alpha = 1$; max-divergence when $\alpha = \infty$. We refer the readers to [Mir17, Definition 3] for a complete description.

Zero Concentrated Differential Privacy is a special case of Rényi differential privacy when Rényi divergence grows linearly with α, e.g., Gaussian Mechanism.

Definition 21 (Zero Concentrated Differential Privacy (zCDP)) *A randomized mechanism $\mathcal{M}$ satisfies ρ-zCDP if for all neighboring datasets D, D' and for all $\alpha > 1$,*

$$D_\alpha(\mathcal{M}(D) \| \mathcal{M}(D')) \leq \rho\alpha,$$

where $D_\alpha(P\|Q)$ is the Rényi divergence of order α.

Lemma 22 (ρ-zCDP to (ε, δ)-DP) *If mechanism $\mathcal{M}$ satisfies ρ-zCDP, then for any $\delta \in (0, 1)$, $\mathcal{M}$ satisfies $(\rho + 2\sqrt{\rho \log(1/\delta)}, \delta)$-DP.*

Proof Since $\mathcal{M}$ satisfies ρ-zCDP, $\mathcal{M}$ also satisfies $(\alpha, \rho\alpha)$-RDP for any $\alpha > 1$. By [Mir17, Proposition 3], for any $\delta \in (0, 1)$, $\mathcal{M}$ satisfies $(\rho\alpha + \frac{\log(1/\delta)}{\alpha - 1}, \delta)$-DP. The remaining proof is by minimizing $f(\alpha) := \rho\alpha + \frac{\log(1/\delta)}{\alpha-1}$ for $\alpha > 1$. The minimum is $\rho + 2\sqrt{\rho \log(1/\delta)}$, by choosing minimizer $\alpha^* = 1 + \sqrt{\frac{\log(1/\delta)}{\rho}}$. ■

We now introduce RDP accounting results for DP-SGD. We first demonstrate RDP guarantee for one-step DP-SGD (i.e. sub-sampled Gaussian mechanism, Lemma 23) then show the privacy guarantee for multi-step DP-SGD through RDP composition (Corollary 24).

Lemma 23 (RDP guarantee for subsampled Gaussian Mechanism, Theorem 11 in [BDRS18]) *Let $\mathcal{M}$ be a ρ-zCDP Gaussian mechanism, and $\mathcal{S}_\gamma$ be a subsampling procedure on the dataset with subsampling rate γ, then the subsampled Gaussian mechanism $\mathcal{M} \circ \mathcal{S}_\gamma$ satisfies $(\alpha, \varepsilon(\alpha))$-RDP with:*

$$\varepsilon(\alpha) \geq 13\gamma^2\rho\alpha, \quad \textit{for any } \alpha \leq \frac{\log(1/\gamma)}{4\rho} \tag{10}$$

Corollary 24 (RDP guarantee for DP-SGD) *Let $\mathcal{M}$ be a Gaussian mechanism satisfying ρ_0-zCDP and $\mathcal{S}_\gamma$ be a subsampling procedure on the dataset with subsampling rate γ, then T-fold (adaptive) composition of subsampled Gaussian mechanism, $\mathcal{M}_T := \underbrace{(\mathcal{M} \circ \mathcal{S}_\gamma) \circ \ldots \circ (\mathcal{M} \circ \mathcal{S}_\gamma)}_{T \text{ times}}$, satisfies $(\alpha, \varepsilon(\alpha))$-RDP with*

$$\varepsilon(\alpha) \geq 13\gamma^2\rho_0 T\alpha, \quad \textit{for any } \alpha \leq \frac{\log(1/\gamma)}{4\rho_0}. \tag{11}$$

Denote $\rho := 13\gamma^2\rho_0 T$ *for a shorthand, if further* $\rho_0 \le \frac{\log(1/\gamma)}{4(1+\sqrt{\frac{\log(1/\delta)}{\rho}})}$, *the composed mechanism* $\mathcal{M}_T$ *satisfies* $(\rho + 2\sqrt{\rho\log(1/\delta)}, \delta)$*-DP for any* $\delta \in (0,1)$.

Proof By RDP accounting of subsampled gaussian mechanisms Lemma 23 and the composition theorem for RDP ([Mir17, Proposition 1]), we have $\mathcal{M}_T$ satisfies $(\alpha, 13\gamma^2\rho_0 T\alpha)$-RDP for any $\alpha \in (1, \frac{\log(1/\gamma)}{4\rho_0})$. Denote $\rho := 13\gamma^2\rho_0 T$. By Lemma 22, if $1 + \sqrt{\frac{\log(1/\delta)}{\rho}} \le \frac{\log(1/\gamma)}{4\rho_0}$ (i.e., the optimal $\alpha^* \le \frac{\log(1/\gamma)}{4\rho_0}$), we have $\mathcal{M}_T$ satisfies $(\rho + 2\sqrt{\rho\log(1/\delta)}, \delta)$-DP. ■

H.3.2 Convex and Lipschitz case

In this section, we analyze the convergence of DP-SGD in convex and Lipschitz settings.

Lemma 25 (Convergence of DP-SGD in convex and L-Lipschitz setting) *Assume that the individual loss function* f *is convex and* L*-Lipschitz. Running DP-SGD with parameters* $\gamma = \frac{2\sqrt{d\log(1/\delta)}}{n\sqrt{\varepsilon}}$, $\sigma^2 = \frac{416L^2\log(1/\delta)}{\varepsilon}$, $T = \frac{n^2\varepsilon^2}{d\log(1/\delta)}$, $\eta = \sqrt{\frac{C^2}{T\left(n^2L^2 + d\sigma^2/\gamma^2 + nL^2/\gamma\right)}}$ *satisfies* (ε, δ)*-DP for any* $\varepsilon \le (d \wedge 8)\log(1/\delta)$. *Moreover,*

$$\mathbb{E}\left[F\left(\bar{\theta}_T\right)\right] - F^* \le \mathcal{O}\left(\frac{CLd^{1/2}\log^{1/2}(1/\delta)}{\varepsilon}\right),$$

with $\bar{\theta}_T$ *being the averaged estimator. In addition, the number of incremental gradient calls is*

$$\mathcal{G} = \frac{2n^2\varepsilon^{3/2}}{\sqrt{d\log(1/\delta)}}.$$

Proof We first state the privacy guarantee. Since each gaussian mechanism satisfies $L^2/2\sigma^2$-zCDP, by Corollary 24, the composed mechanism satisfies ρ-zCDP with $\rho = \frac{13L^2\gamma^2 T}{2\sigma^2} = \frac{\varepsilon^2}{16\log(1/\delta)}$, and thus (ε, δ)-approximate DP.

Let $\hat{g}_t$ be the output from noisy gradient oracle with variance σ^2 and subsampling rate γ (line 7 of Algorithm 3). The variance of the gradient estimator can be upper bounded by:

$$\mathbb{E}[\|\hat{g}_t - \mathbb{E}(\hat{g}_t)\|_2^2] \le \frac{d\sigma^2}{\gamma^2} + \frac{nL^2}{\gamma}.$$

By [GG23, Theorem 9.6], we have:

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{T}\sum_t (F(\theta_t) - F^*)\right] &\le \frac{C^2}{T\eta} + \eta\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^T \|\nabla F(\theta_t)\|_2^2\right] + \eta\left(\frac{d\sigma^2}{\gamma^2} + \frac{nL^2}{\gamma}\right) \\
&\le \frac{C^2}{T\eta} + \eta\left(n^2L^2 + \frac{d\sigma^2}{\gamma^2} + \frac{nL^2}{\gamma}\right) \\
&\le \sqrt{\frac{C^2}{T}\left(n^2L^2 + \frac{d\sigma^2}{\gamma^2} + \frac{nL^2}{\gamma}\right)} \\
&= \mathcal{O}\left(\sqrt{C^2L^2\left(\frac{n^2}{T} + \frac{d}{\rho} + \frac{n}{T\gamma}\right)}\right),
\end{aligned}$$

where the third inequality is by choosing $\eta = \sqrt{\frac{C^2}{T\left(n^2L^2 + \frac{d\sigma^2}{\gamma^2} + \frac{nL^2}{\gamma}\right)}}$, and the forth line is by $\rho = 13T\gamma^2L^2/2\sigma^2$. By the choice of T and γ, we have $\frac{d}{\rho} \ge \max\{\frac{n^2}{T}, \frac{n}{T\gamma}\}$. This implies:

$$\mathbb{E}\left[\frac{1}{T}\sum_t (F(\theta_t) - F^*)\right] \le \mathcal{O}\left(\sqrt{C^2L^2\left(\frac{n^2}{T} + \frac{d}{\rho} + \frac{n}{T\gamma}\right)}\right) \le \mathcal{O}\left(\frac{CLd^{1/2}}{\rho^{1/2}}\right).$$

Since the target approximate DP privacy budget $\varepsilon = 4\sqrt{\rho \log(1/\delta)}$, we have $\sqrt{\rho} = \frac{\varepsilon}{4\sqrt{\log(1/\delta)}}$. Plugging this into the bound above, we have:

$$\mathbb{E}\left[F(\bar{\theta}_T)\right] - F^* \leq \mathbb{E}\left[\frac{1}{T}\sum_t \left(F(\theta_t) - F^*\right)\right] \leq \mathcal{O}\left(\frac{CLd^{1/2}\log^{1/2}(1/\delta)}{\varepsilon}\right).$$

For the number of incremental gradient calls (denoted as $\mathcal{G}$), we have

$$\mathcal{G} = n\gamma T = \frac{2n^2\varepsilon^{3/2}}{\sqrt{d\log(1/\delta)}}.$$

■

H.3.3 Strongly Convex and Lipschitz case

Lemma 26 (Convergence of DP-SGD in strongly convex and L-Lipschitz setting) *Assume individual loss function f is λ-strongly convex and L-Lipschitz. Running DP-SGD with parameters $\gamma = \frac{2\sqrt{d\log(1/\delta)}}{n\sqrt{\varepsilon}}$, $\sigma^2 = \frac{416L^2\log(1/\delta)}{\varepsilon}$, $T = \frac{n^2\varepsilon^2}{d\log(1/\delta)}$, $\eta_t = \frac{2}{n\lambda(t+1)}$ satisfies (ε, δ)-DP for any $\varepsilon \leq (d \wedge 8)\log(1/\delta)$. Moreover,*

$$\mathbb{E}\left[F\left(\frac{2}{T(T+1)}\sum_{t=1}^{T} t\theta_t\right)\right] - F^* \leq \mathcal{O}\left(\frac{dL^2\log(1/\delta)}{n\lambda\varepsilon^2}\right),$$

and the number of incremental gradient calls is

$$\mathcal{G} = \frac{2n^2\varepsilon^{3/2}}{\sqrt{d\log(1/\delta)}}.$$

Proof The derivation of privacy guarantee follows the same procedure as Lemma 25. By Corollary 24, the total zCDP guarantee $\rho = \frac{13L^2\gamma^2T^2}{2\sigma^2}$. Choosing learning rate $\eta_t = \frac{2}{\Lambda(t+1)}$ and apply convergence result from [LJSB12], where $\Lambda = n\lambda$ be strong convex parameter of F, we have:

$$\begin{aligned}
\mathbb{E}\left[F\left(\frac{2}{T(T+1)}\sum_{t=1}^{T} tx_t\right)\right] - F^* &\leq \frac{2\mathbb{E}[\|\hat{g}_t\|_2^2]}{\Lambda(T+1)} \\
&\leq \mathcal{O}\left(\frac{n^2L^2}{\Lambda T} + \frac{d\sigma^2}{\Lambda\gamma^2 T} + \frac{nL^2}{\Lambda\gamma T}\right) \\
&= \mathcal{O}\left(\frac{L^2}{\Lambda}\left(\frac{n^2}{T} + \frac{d}{\rho} + \frac{n}{\gamma T}\right)\right).
\end{aligned}$$

where in the last line we use the fact that $\rho = \frac{13T\gamma^2L^2}{2\sigma^2}$. By the choice of T and γ, we have $\frac{d}{\rho} \geq \max\{\frac{n^2}{T}, \frac{n}{T\gamma}\}$. This implies:

$$\mathbb{E}\left[F\left(\frac{2}{T(T+1)}\sum_{t=1}^{T} tx_t\right)\right] - F^* \leq \mathcal{O}\left(\frac{dL^2}{\Lambda\rho}\right) = \mathcal{O}\left(\frac{dL^2\log(1/\delta)}{n\lambda\varepsilon^2}\right),$$

where the last equality is using the fact that the target privacy budget $\varepsilon = 4\sqrt{\rho\log(1/\delta)}$.

For the number of incremental gradient calls (denote as $\mathcal{G}$), we have the same result as in Lemma 25:

$$\mathcal{G} = \frac{2n^2\varepsilon^{3/2}}{\sqrt{d\log(1/\delta)}}.$$

■

H.4 Analysis of Purified DP-SGD

Lemma 27 (Error from Laplace perturbation) *Suppose $x \in \mathbb{R}^d$ and $\tilde{x} = x + Lap^{\otimes d}(b)$, then*

$$\mathbb{E}[\|x - \tilde{x}\|_2] \leq \sqrt{2d}b.$$

Proof

$$\mathbb{E}[\|x - \tilde{x}\|_2] = \mathbb{E}\left[\sqrt{\|x - \tilde{x}\|_2^2}\right] \leq \sqrt{\mathbb{E}[\|x - \tilde{x}\|_2^2]} = \sqrt{2db^2}$$

■

H.4.1 Proof of Theorem 3

Theorem 10 (Restatement of Theorem 3) *Let the domain $\mathcal{C} \subset \mathbb{R}^d$ be a convex set with ℓ_2 diameter C, and let $f(\cdot; \mathrm{x})$ be L-Lipschitz for all $\mathrm{x} \in \mathcal{X}$. Algorithm 4 satisfies 2ε-pure DP and with $\tilde{\mathcal{O}}(n^2\varepsilon^{3/2}d^{-1})$ incremental gradient calls, the output θ_{out} satisfies:*

1. *If $f(\cdot; \mathrm{x})$ is convex for all $\mathrm{x} \in \mathcal{X}$, then $\mathbb{E}\left[\mathcal{L}(\theta_{\text{out}})\right] - \mathcal{L}^* \leq \tilde{\mathcal{O}}\left(CLd/n\varepsilon\right)$.*
2. *If $f(\cdot; \mathrm{x})$ is λ-strongly convex for all $\mathrm{x} \in \mathcal{X}$, then $\mathbb{E}\left[\mathcal{L}\left(\theta_{\text{out}}\right)\right] - \mathcal{L}(x^*) \leq \tilde{\mathcal{O}}\left(d^2L^2/n^2\lambda\varepsilon^2\right)$.*

Proof Setting $\omega = \frac{1}{n^2}$, $\delta = \frac{2\omega}{2^{4d}d^d n^{2d} C^d} = 2^{1-4d}d^{-d}n^{-2d-2}C^{-d}$, we have

$$\log(1/\delta) = \mathcal{O}\left(d\log(2) + d\log(n) + d\log(d) + d\log(C)\right) = \tilde{\mathcal{O}}(d) \tag{12}$$

Applying Corollary 4 with ω, δ defined above and choose $\varepsilon' = \varepsilon$, the additional error from purification can be upper bounded by $\frac{1}{n^2\varepsilon} + \frac{C}{n^2}$.

(When f is Convex and L-Lipschitz): By Lemma 25 and dividing both sides by n:

$$\begin{aligned}\mathbb{E}\left[\mathcal{L}(\theta_{\text{out}})\right] - \mathcal{L}^* &\leq \frac{L}{n^2\varepsilon} + \frac{CL}{n^2} + \mathcal{O}\left(\frac{CLd^{1/2}\log^{1/2}(1/\delta)}{n\varepsilon}\right) \\ &= \tilde{\mathcal{O}}\left(\frac{L}{n^2\varepsilon} + \frac{CL}{n^2} + \frac{CLd}{n\varepsilon}\right)\end{aligned}$$

(When f is λ-strongly Convex and L-Lipschitz): By Lemma 26 and dividing both sides by n:

$$\begin{aligned}\mathbb{E}\left[\mathcal{L}(\theta_{\text{out}})\right] - \mathcal{L}^* &\leq \frac{L}{n^2\varepsilon} + \frac{CL}{n^2} + \mathcal{O}\left(\frac{dL^2\log(1/\delta)}{n^2\lambda\varepsilon^2}\right) \\ &= \tilde{\mathcal{O}}\left(\frac{L}{n^2\varepsilon} + \frac{CL}{n^2} + \frac{d^2L^2}{n^2\lambda\varepsilon^2}\right)\end{aligned}$$

The number of incremental gradient calls for both cases:

$$\mathcal{G} = \frac{2n^2\varepsilon^{3/2}}{\sqrt{d\log(1/\delta)}} = \tilde{\mathcal{O}}\left(n^2\varepsilon^{3/2}d^{-1}\right)$$

■

I Deferred Proofs for DP-Frank-Wolfe

I.1 Approximate DP Frank-Wolfe Algorithm

Given a dataset $D = x_1, \ldots, x_n$ and a parameter space $\mathcal{C}$, we denote the individual loss function by $f : \mathcal{C} \times \mathcal{X} \to \mathbb{R}$. We define the average empirical loss as follows:

$$\mathcal{L}(\theta) := \frac{1}{n}\sum_{i=1}^{n} f(\theta; x_i) \tag{13}$$

For completeness, we state the Frank-Wolfe algorithm [FW+56] as follows.

The differential private version of Algorithm 5 is modified by using the exponential mechanism to select coordinates in each update. We follow the setting in [TGTZ15] with initialization at point zero.

Algorithm 5: Frank-Wolfe algorithm (Non-Private)

Input: $\mathcal{C} \subseteq \mathbb{R}^d$, loss function $\mathcal{L} : \mathcal{C} \to \mathbb{R}$, number of iterations T, stepsizes η_t.

Choose an arbitrary θ_1 from $\mathcal{C}$

for $t = 1$ *to* $\tilde{T} - 1$ **do**

  Compute $\tilde{\theta}_t = \arg\min_{\theta \in \mathcal{C}} \langle \nabla \mathcal{L}(\theta_t), \theta - \theta_t \rangle$

  Set $\theta_{t+1} = \theta_t + \eta_t(\tilde{\theta}_t - \theta_t)$

Output θ_T

Algorithm 6: Approximate DP Frank-Wolfe Algorithm [TGTZ15]

Input: Dataset $D = \{d_1, \ldots, d_n\}$, loss function f defined in Eq. (13) with ℓ_1-Lipschitz constant L_1, privacy parameters (ε, δ), convex set $\mathcal{C} = \text{conv}(S)$, $\|\mathcal{C}\|_1 = \max_{s \in S} \|s\|_1$.

Initialize $\theta_0 \leftarrow \mathbf{0} \in \mathcal{C}$.

for $t = 0$ *to* $T - 1$ **do**

  $\forall s \in S, \alpha_s \leftarrow \langle s, \nabla \mathcal{L}(\theta_t; \mathcal{D}) \rangle + \text{Lap}\left(\frac{L_1 \|\mathcal{C}\|_1 \sqrt{8T \log(1/\delta)}}{n\varepsilon}\right)$, where $\text{Lap}(\lambda) \sim \frac{1}{2\lambda} e^{-|x|/\lambda}$

  $\tilde{\theta}_t \leftarrow \arg\min_{s \in S} \alpha_s$

  $\theta_{t+1} \leftarrow (1 - \eta_t)\theta_t + \eta_t \tilde{\theta}_t$, where $\eta_t = \frac{2}{t+2}$

Output $\theta_{\text{priv}} = \theta_T$

Lemma 28 (Equation 21 of [TTZ14]) *Running Algorithm 6 for T iterations yields:*

$$\mathbb{E}\left[\mathcal{L}(\theta_T; D) - \min_{\theta \in \mathcal{C}} \mathcal{L}(\theta; D)\right] = O\left(\frac{\Gamma_{\mathcal{L}}}{T} + \frac{L_1 \|\mathcal{C}\|_1 \sqrt{8T \log(1/\delta)} \log(TL_1 \|\mathcal{C}\|_1 \cdot |S|)}{n\varepsilon}\right),$$

where $\Gamma_{\mathcal{L}}$ is the curvature parameter [TGTZ15, Definition 2.1], which can be upper bounded by $\beta \|\mathcal{C}\|_1^2$ if the loss function f is β-smooth [TTZ14].

The sparsity of θ_T is given in the following lemma.

Lemma 29 (Sparsity of DP Frank-Wolfe) *Suppose $S \subset \{x \in \mathbb{R}^d \mid \text{nnz}(x) \leq s\}$. After running Algorithm 6 for T iterates, the output θ_T is $Ts \wedge d$-sparse.*

Proof From Line 10 and 12 of Algorithm 6, we know $\text{nnz}(\theta_{t+1}) \leq \text{nnz}(\theta_t) + s$. Since $\text{nnz}(\theta_0) = 0$, we have $\text{nnz}(\theta_T) \leq Ts$. Since $\theta_T \in \mathbb{R}^d$, we have $\text{nnz}(\theta_T) \leq d$. Therefore, $\text{nnz}(\theta_T) \leq Ts \wedge d$. ■

I.2 Pure DP Frank-Wolfe Algorithm

Algorithm 7: Pure DP Frank-Wolfe Algorithm

Input: Dataset D, loss function $\mathcal{L}$: defined in Eq. (13), DP parameter ε, a convex polytope $\mathcal{C} = \text{conv}(S)$, where S is the vertices set, number of iterations T, a Gaussian random matrix $\Phi \in \mathbb{R}^{k \times d}$ constructed by Lemma 44 satisfying $(1/4, 4T)$-RWC with high probability.

Set parameters $T = \tilde{\mathbf{\Theta}}\left(\sqrt{\frac{\beta \|\mathcal{C}\|_1 n\varepsilon}{L_1}}\right)$, $k = \mathbf{\Theta}\left(T \log\left(\frac{d}{T}\right) + \log n\right)$, $\omega = \frac{1}{n}$, $\delta = \frac{2\omega}{(nk)^k}$.

$\theta_{\text{FW}} \leftarrow$ Algorithm 6 ▷ (ε, δ)-DP FW

$\theta_{\text{apx-}k} \leftarrow \Phi\theta_{\text{FW}} \in \mathbb{R}^k$ ▷ Dimension reduction

$\theta_{\text{pure-}k} \leftarrow \mathcal{A}_{\text{pure}}\left(x_{\text{apx}} = \mathsf{Clip}^{\ell_2}_{2\|\mathcal{C}\|_1}(\theta_{\text{apx-}k}), \Theta = \mathcal{B}^k_{\ell_2}(2\|\mathcal{C}\|_1), \varepsilon' = \varepsilon, \delta, \omega\right)$ ▷ Algorithm 1

$\theta_{\text{pure-}d} \leftarrow \mathcal{M}_{\text{rec}}(\theta_{\text{pure-}k}, \Phi, \xi)$ ▷ Algorithm 8, with ξ defined in Eq. (14)

$\theta_{\text{out}} = \mathsf{Clip}^{\ell_1}_{\|\mathcal{C}\|_1}(\theta_{\text{pure-}d})$

Output: θ_{out}

Algorithm 8: Sparse Vector Recovery $\mathcal{M}_{\text{rec}}(b, \Phi, \xi)$

Input: Noisy measurement b, design matrix Φ, noise tolerant magnitude ξ

Define the feasible set $\mathcal{U} := \{\theta \in \mathbb{R}^d \mid \|\Phi\theta - b\|_1 \leq \xi\}$

Solve $\hat{\theta} = \arg\min_{\theta \in \mathcal{U}} \|\theta\|_1$

Output: $\hat{\theta}$

I.3 Proof of Theorem 4

Proof of Theorem 4 The privacy analysis follows by Theorem 1 and the post-possessing property of DP. The utility analysis follows by bounding the following (a) and (b):

$$\underbrace{\mathbb{E}[\mathcal{L}(\theta_{\text{out}}; D) - \mathcal{L}(\theta_{FW}; D)]}_{(a)} + \underbrace{\mathbb{E}[\mathcal{L}(\theta_{FW}; D) - \mathcal{L}(\theta^*; D)]}_{(b)}$$

For (a): Since $\mathcal{C}$ is an ℓ_1-ball with center $\mathbf{0}$ and ℓ_1 radius $\|\mathcal{C}\|_1$, the vertices set is $S = \{x \mid \|x\|_1 = \|\mathcal{C}\|_1, \text{nnz}(x) = 1\}$. We note that θ_{FW}, the output of Algorithm 6 is T-sparse by Lemma 29.

Denote the "failure" events as follows: $F_1 := \{u \leq \omega \text{ in Line 3 of Algorithm 1}\}$, where the uniform sample is accepted in Algorithm 1; $F_2 := \{\Phi \text{ is not } (e, 4T)\text{-RWC}\}$, where the randomly sampled Φ is not $(e, 4T)$-restricted well-conditioned (RWC); and $F_3 := \{\|\mathbf{z}\|_1 > \xi\}$, where the Laplace noise added in Line 4 of Algorithm 1 exceeds the tolerance threshold. We first bound the failure probability $\mathbb{P}(F_1 \cup F_2 \cup F_3)$, and then analyze the utility under the "success" event $F_1^c \cap F_2^c \cap F_3^c$, followed by an expected overall utility bound.

Since $\omega = \frac{1}{n}$, we have $\mathbb{P}(F_1) = \frac{1}{n}$.

Set the distortion rate as $e = \frac{1}{4}$, and $k = \mathbf{\Theta}\left(T \log\left(\frac{d}{T}\right) + \log n\right)$. Constructing Φ following Lemma 44, and by Lemma 45, $\Phi \in \mathbb{R}^{k \times d}$ is $(4T, e)$-RWC with probability at least $1 - \frac{1}{n}$, i.e., $\mathbb{P}(F_2) \leq \frac{1}{n}$.

Set

$$\xi = \frac{4\Delta}{\varepsilon}(k + \log n), \tag{14}$$

where $\Delta = 8\sqrt{k}\|\mathcal{C}\|_1 \left(\frac{\delta}{2\omega}\right)^{1/k} = \frac{8\|\mathcal{C}\|_1}{n\sqrt{k}}$. Then by Lemma 48 with taking $b = \frac{2\Delta}{\varepsilon}$ and $t = b(k + 2\log n)$, we have $\mathbb{P}(F_3) \leq \frac{1}{n}$.

Under the "success" event $F_1^c \cap F_2^c \cap F_3^c$, consider the variables in Algorithm 7, we have

$$\begin{aligned}
\theta_{\text{pure-}k} &= \mathcal{A}_{\text{pure}}\left(x_{\text{apx}} = \mathsf{Clip}^{\ell_2}_{2\|\mathcal{C}\|_1}(\theta_{\text{apx-}k}), \Theta = \mathcal{B}^k_{\ell_2}(2\|\mathcal{C}\|_1), \varepsilon' = \varepsilon, \delta, \omega\right) \\
&\qquad\qquad \text{(Line 10 of Algorithm 7)} \\
&= \mathcal{A}_{\text{pure}}\left(\theta_{\text{apx-}k}, \Theta = \mathcal{B}^k_{\ell_2}(2\|\mathcal{C}\|_1), \varepsilon' = \varepsilon, \delta, \omega\right) \qquad \text{(Under } F_2^c, \|\theta_{\text{apx-}k}\|_2 \leq (1+e)\|\mathcal{C}\|_1) \\
&= \theta_{\text{apx-}k} + \text{Laplace}^{\otimes \text{d}}\left(\frac{2\Delta}{\varepsilon}\right) \qquad \text{(Under } F_1^c) \\
&= \Phi\theta_{\text{FW}} + \tilde{w}, \text{ where } \|\tilde{w}\|_1 \leq \xi. \qquad \text{(Under } F_3^c)
\end{aligned}$$

By Lemma 46 and $\theta_{\text{pure-}d} = \mathcal{M}_{\text{rec}}(\theta_{\text{pure-}k}, \Phi, \xi)$, since θ_{FW} is T-sparse, we have

$$\|\theta_{\text{pure-}d} - \theta_{\text{FW}}\|_1 \leq \frac{4\sqrt{T}}{\sqrt{1-e}} \cdot \xi.$$

Since $\|\theta_{\text{pure-}k} - \Phi\theta_{\text{FW}}\|_1 = \|\tilde{w}\|_1 \leq \xi$, i.e., θ_{FW} is in the feasible set, we have $\|\theta_{\text{pure-}d}\|_1 \leq \|\theta_{\text{FW}}\|_1 \leq \|\mathcal{C}\|_1$, therefore, by Line 14 in Algorithm 7, $\theta_{\text{out}} = \mathsf{Clip}^{\ell_1}_{\|\mathcal{C}\|_1}(\theta_{\text{pure-}d}) = \theta_{\text{pure-}d}$ (under event F_3^c.)

Since $\mathcal{L}$ is L_1-Lipschitz with respect to ℓ_1 norm, we get an upper bound on (a):

$$\begin{aligned}\mathbb{E}[\mathcal{L}(\theta_{\text{out}}; D) - \mathcal{L}(\theta_{FW}; D) \mid F_1^c \cap F_2^c \cap F_3^c] &\leq L_1 \|\theta_{\text{out}} - \theta_{FW}\|_1 \\ &\leq L_1 \frac{4\sqrt{T}}{\sqrt{1-e}} \cdot \xi \\ &\leq \frac{64}{\sqrt{3}} \frac{L_1 \|\mathcal{C}\|_1 (k + \log n)}{n\varepsilon}\end{aligned} \tag{15}$$

For (b): by Lemma 28, we have:

$$\begin{aligned}\mathbb{E}\left[\mathcal{L}(\theta_{\text{FW}}; D) - \min_{\theta \in \mathcal{C}} \mathcal{L}(\theta; D) \mid F_1^c \cap F_2^c \cap F_3^c\right] &= \mathcal{O}\left(\frac{\Gamma_{\mathcal{L}}}{T} + \frac{L_1 \|\mathcal{C}\|_1 \sqrt{8T \log(1/\delta)} \log(T L_1 \|\mathcal{C}\|_1 \cdot |S|)}{n\varepsilon}\right) \\ &\leq \mathcal{O}\left(\frac{\beta \|\mathcal{C}\|_1^2}{T} + \frac{L_1 \|\mathcal{C}\|_1 \sqrt{8T \log(1/\delta)} \log(T L_1 \|\mathcal{C}\|_1 \cdot |S|)}{n\varepsilon}\right).\end{aligned}$$

Therefore,

$$\begin{aligned}&\mathbb{E}\left[\mathcal{L}(\theta_{\text{out}}; D) - \min_{\theta \in \mathcal{C}} \mathcal{L}(\theta; D) \mid F_1^c \cap F_2^c \cap F_3^c\right] \\ &\leq \mathcal{O}\left(\frac{L_1 \|\mathcal{C}\|_1 (k + \log n)}{n\varepsilon} + \frac{\beta \|\mathcal{C}\|_1^2}{T} + \frac{L_1 \|\mathcal{C}\|_1 \sqrt{8T \log(1/\delta)} \log(T L_1 \|\mathcal{C}\|_1 \cdot |S|)}{n\varepsilon}\right) \\ &= \tilde{\mathcal{O}}\left(\frac{L_1 \|\mathcal{C}\|_1 T}{n\varepsilon} + \frac{\beta \|\mathcal{C}\|_1^2}{T} + \frac{L_1 \|\mathcal{C}\|_1 T}{n\varepsilon}\right) \qquad (\delta = \tfrac{2\omega}{(nk)^k}) \\ &= \tilde{\mathcal{O}}\left(\frac{\beta \|\mathcal{C}\|_1^2}{T} + \frac{L_1 \|\mathcal{C}\|_1 T}{n\varepsilon}\right).\end{aligned}$$

By setting $T = \tilde{\mathbf{\Theta}}(\sqrt{\frac{n\varepsilon\beta\|\mathcal{C}\|_1}{L_1}})$, we have

$$\mathbb{E}\left[\mathcal{L}(\theta_{\text{out}}; D) - \min_{\theta \in \mathcal{C}} \mathcal{L}(\theta; D) \mid F_1^c \cap F_2^c \cap F_3^c\right] \leq \tilde{\mathcal{O}}\left(\frac{(\beta L_1 \|\mathcal{C}\|_1^3)^{1/2}}{(n\varepsilon)^{1/2}}\right). \tag{16}$$

By Line 14 in Algorithm 7, we have $\|\theta_{\text{out}}\|_1 \leq \|\mathcal{C}\|_1$. Therefore,

$$\mathbb{E}\left[\mathcal{L}(\theta_{\text{out}}; D) - \min_{\theta \in \mathcal{C}} \mathcal{L}(\theta; D) \mid F_1 \cup F_2 \cup F_3\right] \leq L_1 \|\theta_{\text{out}} - \theta^*\|_1 \leq 2L_1 \|\mathcal{C}\|_1.$$

$$\begin{aligned}\mathbb{E}\left[\mathcal{L}(\theta_{\text{out}}; D) - \min_{\theta \in \mathcal{C}} \mathcal{L}(\theta; D)\right] &\leq \tilde{\mathcal{O}}\left(\frac{(\beta L_1 \|\mathcal{C}\|_1^3)^{1/2}}{(n\varepsilon)^{1/2}} + \frac{3}{n} 2L_1 \|\mathcal{C}\|_1\right) \\ &\leq \tilde{\mathcal{O}}\left(\frac{(\beta L_1 \|\mathcal{C}\|_1^3)^{1/2}}{(n\varepsilon)^{1/2}}\right)\end{aligned}$$

For the computation cost, in each iteration, the full-batch gradient is calculated; therefore, the cost for calculating θ_{FW} is $Tnd = \tilde{\mathcal{O}}(n^{3/2}d)$. The computation cost for Algorithm 1 is $\mathcal{O}(d)$. Therefore, the computation cost is $\tilde{\mathcal{O}}(n^{3/2}d)$, plus one call of the LASSO solver.

I.4 Proof of Lemma 30

Lemma 30 *Let $\mathcal{A}$ be any ε-DP ERM algorithm. For every parameter n, d, ε. There is a DP-ERM problem with a convex, 1-Lipschitz, 1-smooth loss function, a constrained parameter space $\Theta = \{\theta \in \mathbb{R}^d | \|\theta\|_1 \leq 1\}$ and a dataset $\text{Data} := \{x_1, ..., x_n\} \in \mathcal{X}^n$ that gives rise to the empirical risk $\mathcal{L}(\theta) = \frac{1}{n}\sum_{i=1}^n \ell(\theta; x_i)$, such that with probability at least 0.5, the excess empirical risk*

$$\mathcal{L}(\mathcal{A}(\text{Data})) - \min_{\theta \in \Theta} \mathcal{L}(\theta) \geq \sqrt{\frac{\log(d+1)}{n\varepsilon + \log(4)}} \wedge 1.$$

Definition 31 ((α, β)-accurate ERM algorithm) *Given parameter space Θ, dataspace $\mathcal{X}$, and risk function R, we say an ERM algorithm $\mathcal{M} : \mathcal{X}^n \to \Theta$ is (α, β)-accurate if for any dataset $D \in \mathcal{X}^n$, with probability at least $1 - \beta$ over the randomness of algorithm:*

$$R(\mathcal{M}(D); D) - \min_{\theta \in \Theta} R(\theta; D) \leq \alpha$$

Lemma 32 (Restatement of Lemma 30) *There exists a hard instance with n samples over $\mathcal{B}_1^d$ and a 1-Lipschitz loss function $\mathcal{L}$ such that any ε-pure differential private $(\alpha, 1/2)$-accurate ERM algorithm $\mathcal{M}$ must have:*

$$\alpha \geq \sqrt{\frac{\log(d+1)}{n\varepsilon + \log(4)}} \wedge 1$$

Proof of Lemma 30 We proof by the standard packing argument. Consider $\mathcal{B}_1^d$ and an α-packing over it: $\{\theta_i\}_{i \in [K]}$, with K being packing number $M(\alpha, \mathcal{B}_1^d, \|\cdot\|_2)$. For any $i \in [K]$, let $E_i = \{\theta \in \Theta \mid \|\theta - \theta_i\|_2 \leq \alpha\}$ and $X_i = \underbrace{\{\theta_i, \ldots, \theta_i\}}_{\text{n copies}}$.

We define the error function is $\mathcal{L}(\theta; X_i) = \frac{1}{n} \sum_{j=1}^n \|\theta - X_i(j)\|_2$. Notice that:

$$\begin{aligned} 1 \geq \mathbb{P}(\mathcal{M}(X_i) \notin E_i) &\geq \sum_{j \in [K] \setminus i} \mathbb{P}(\mathcal{M}(X_i) \in E_j) \\ &\geq \sum_{j \in [K] \setminus i} \exp(-n\varepsilon) \times \mathbb{P}(\mathcal{M}(X_j) \in E_j) \\ &\geq \frac{K \exp(-n\varepsilon)}{4} \end{aligned} \tag{17}$$

Taking log of both sides, we have

$$n\varepsilon + \log(4) \geq \log(K) \tag{18}$$

It remains to calculate the packing number K. Notice that $M(\alpha, \mathcal{B}_1^d, \|\cdot\|_2) \asymp N(\alpha, \mathcal{B}_1^d, \|\cdot\|_2) \asymp \exp\left(\frac{1}{\alpha^2} \log(\alpha^2 d)\right)$[4], where we assume $\alpha \gtrsim \frac{1}{\sqrt{d}}$ [Wu16, Equation 15.4]. This implies:

$$\alpha \geq \tilde{\Omega}\left(\frac{1}{\sqrt{n\varepsilon + \log(4)}} \vee \frac{1}{\sqrt{d}}\right) \tag{19}$$

where $\tilde{\Omega}$ hides universal constant and logarithmic terms w.r.t. d. We conclude the proof by observing that $\mathcal{L}(\theta; X_i) \leq 1$, which implies that $\alpha \leq 1$. ■

J Deferred Proofs for Data-dependent DP mechanism Design

J.1 Pure DP Propose Test Release

Definition 33 (Local Sensitivity) *The local sensitivity of a query function q on a dataset X is defined as*

$$\Delta_{\text{Local}}^q(X) := \max_{X' \simeq X} \|q(X) - q(X')\|_2,$$

where $X' \simeq X$ denotes that X' is a neighboring dataset of X.

We first present the original version of PTR in Algorithm 9. The pure DP version, obtained via the purification trick, is given in Algorithm 10. Their privacy guarantees are stated as follows:

Lemma 34 *Algorithm 9 satisfies $(2\varepsilon, \delta)$-DP and its purified version, Algorithm 10 satisfies $(2\varepsilon + \varepsilon', 0)$-DP*

Proof The privacy guarantee for Algorithm 9 is based on [Vad17, Proposition 7.3.2]. For Algorithm 10, the privacy guarantee follows from the post-processing property of differential privacy and the privacy guarantee of Algorithm 1. ■

[4] N denotes the covering number.

Algorithm 9: $\mathcal{M}_{\mathrm{PTR}}(X, q, \varepsilon, \delta, \beta)$: Propose-Test-Release [DL09]

Input: Dataset X; privacy parameters ε, δ; proposed bound β; query function $q : \mathcal{X} \to \Theta$

Compute: $\mathcal{D}_\beta^q(X) = \min_{X'}\{d_{\mathrm{Hamming}}(X, X') \mid \Delta^q_{\mathrm{Local}}(X') > \beta\}$

if $\mathcal{D}_\beta^q(X) + \mathrm{Lap}\left(\frac{1}{\varepsilon}\right) \le \frac{\log(1/\delta)}{\varepsilon}$ **then**

  Output $\perp$

else

  Release $f(X) + \mathrm{Lap}\left(\frac{\beta}{\varepsilon}\right)$

Algorithm 10: Pure DP Propose-Test-Release

Input: Dataset X; privacy parameters $\varepsilon, \varepsilon', \delta$; proposed bound β; query function $q : \mathcal{X} \to \Theta$, level of uniform smoothing ω

$\theta \leftarrow \mathcal{M}_{\mathrm{PTR}}(X, q, \varepsilon, \delta, \beta)$

if $\theta == \perp$ **then**

  $u \sim \mathrm{Unif}(\Theta)$

  $\theta \leftarrow u$

$\theta_{\mathrm{out}} \leftarrow \mathcal{A}_{\mathsf{pure}}(\theta, \Theta, \varepsilon', \varepsilon, \delta, \omega)$ ▷ Algorithm 1

Output: θ_{out}

J.2 Privately Bounding Local Sensitivity

We assume query function $q : \mathcal{X}^* \to \Theta$ with $\Theta \subset \mathbb{R}^d$ being a convex set and $\mathsf{Diam}_2(\Theta) = R$. Assume the global sensitivity of local sensitivity is upper bounded by 1: $\max_{X \simeq X'} \|\Delta^q_{\mathrm{Local}}(X) - \Delta^q_{\mathrm{Local}}(X')\|_2 \le 1$. The purified version of PTR based on privately releasing local sensitivity is stated in Algorithm 11 and its utility guarantee is included in Theorem 11.

Algorithm 11:

Input: Dataset D; privacy parameters $\varepsilon, \varepsilon', \delta$; proposed bound β; query function $q : \mathcal{X}^* \to \Theta \subset \mathbb{R}^d$ with $\mathsf{Diam}_2(\Theta) = R$, level of uniform smoothing ω

$\hat{\beta} = \Delta^q_{\mathrm{Local}}(D) + \mathrm{Lap}\,(1/\varepsilon) + \log(2/\delta)/\varepsilon$

$q_{\mathsf{apx}} \leftarrow \mathrm{Proj}_\Theta(q(D) + \mathrm{Lap}^{\otimes d}(\hat{\beta}/\varepsilon))$

$q_{\mathsf{pure}} \leftarrow \mathcal{A}_{\mathsf{pure}}(\Theta, \varepsilon', \omega, R)$ ▷ Algorithm 1

Output: q_{pure}

Theorem 11 (Restatement of theorem 5) *Algorithm 11 satisfies $(3\varepsilon, 0)$-DP . Moreover,*

$$\mathbb{E}\left[\|q_{\mathrm{out}}(D) - q(D)\|_2\right] \le \tilde{\mathcal{O}}\left(\frac{d^{1/2}\Delta^q_{\mathrm{Local}}(D)}{\varepsilon} + \frac{d^{3/2}}{\varepsilon^2}\right).$$

Proof First notice that $\hat{\beta}$ satisfies ε-DP by the privacy guarantee from Laplace mechanism and the assumption that global sensitivity of $\Delta^q_{\mathrm{Local}}(D)$ is upper bounded by 1. Second, we notice that $\mathbb{P}(\hat{\beta} > \Delta^q_{\mathrm{Local}}(D)) = \mathbb{P}(\mathrm{Lap}(1/\varepsilon) \ge \log(\delta/2)/\varepsilon) = 1 - \delta$. This implies $q(D) + \mathrm{Lap}^{\otimes d}(\hat{\beta}/\varepsilon)$ satisfies (ε, δ)-probabilistic DP ([DRV10, Definition 2.2]), thus satisfies (ε, δ)-DP. By post-processing and simple composition, q_{apx} satisfies $(2\varepsilon, \delta)$-DP. Finally, using $\mathcal{A}_{\mathrm{pure}}$ under appropriate choice of δ, we have q_{pure} satisfies $(3\varepsilon, 0)$-DP by Theorem 1.

We now prove the utility guarantee. For notational convenience, we denote $z_0 \sim \mathrm{Lap}\,(1/\varepsilon)$, $Z_1 \sim \mathrm{Lap}^{\otimes d}(\hat{\beta}/\varepsilon)$ and $Z_2 \sim \mathrm{Lap}^{\otimes d}(\Delta/\varepsilon)$. By definition of q_{pure}, we have:

$$\begin{aligned}
\mathbb{E}\left[\|q_{\mathrm{pure}} - q(D)\|_2\right] &= \mathbb{E}\left[\|q_{apx} - q(D) + Z_2\|_2\right] + \omega C \\
&= \mathbb{E}\left[\|\mathrm{Proj}_\Theta\left(q(D) + Z_1\right) - q(D) + Z_2\|_2\right] + \omega C \\
&\le \mathbb{E}\left[\|Z_1\|_2\right] + \mathbb{E}\left[\|Z_2\|_2\right] + \omega C
\end{aligned} \tag{20}$$

Notice that:

$$\begin{aligned}\mathbb{E}\left[\|Z_1\|_2\right] &\leq \sqrt{\mathbb{E}\left[\|Z_1\|_2^2\right]} \\ &= \sqrt{d\mathbb{E}[Z_{11}^2]} \qquad (Z_{11} \text{ denotes first element of } Z_1) \\ &\leq \frac{\Delta^q_{\text{Local}}(D) + \log(2/\delta)/\varepsilon}{\varepsilon} + \frac{1}{\varepsilon}\mathbb{E}\left[|z_0|\right] \\ &= \mathcal{O}\left(\frac{\sqrt{d}(\Delta^q_{\text{Local}}(D) + {}^{1+\log(2/\delta)}/_{\varepsilon})}{\varepsilon}\right)\end{aligned}$$

Now, set $\omega = \frac{1}{100} \wedge \frac{1}{C\varepsilon^2}$ and $\delta = \frac{2\omega}{(16dC\varepsilon)^d}$. By Corollary 4, we have :

$$\mathbb{E}[\|Z_2\|_2] + \omega C \leq \frac{2}{\varepsilon^2}.$$

Also notice that $\log(2/\delta) = \tilde{\mathcal{O}}(d)$. Thus,

$$\mathbb{E}\left[\|q_{\text{pure}} - q(D)\|_2\right] \leq \tilde{\mathcal{O}}\left(\frac{\sqrt{d}\Delta^q_{\text{Local}}(D)}{\varepsilon} + \frac{d^{3/2}}{\varepsilon^2}\right)$$

■

J.3 Private Mode Release

The mode release algorithm discussed in Section 5.2 is provided in Algorithm 12.

Algorithm 12: Pure DP Mode Release

Input: Dataset D, pure DP parameter ε

Set: $\log(1/\delta) = d\log\left(2d^3/\varepsilon\right)$, where $d = \log_2|\mathcal{X}|$.

Compute the mode $f(D)$ and its frequency occ_1, as well as the frequency occ_2 of the second most frequent item.

Compute the gap: $\mathcal{D}_0^f(D) \leftarrow \left\lceil \frac{\text{occ}_1 - \text{occ}_2}{2} \right\rceil$.

if $\mathcal{D}_0^f(D) - 1 + \text{Lap}\left(\frac{1}{\varepsilon}\right) \leq \frac{\log(1/\delta)}{\varepsilon}$ **then**

  $u_{\text{apx}} \leftarrow \perp$

else

  $u_{\text{apx}} \leftarrow f(D)$

$u_{\text{pure}} \leftarrow$ Algorithm 2 with inputs ε, δ, $u_{\text{apx}}, \mathcal{Y} = \mathcal{X}$, and $\mathsf{Index} = \mathsf{id}$, the identity map.

Output: u_{pure}

Proof of Theorem 6 By [Vad17, Proposition 3.3] and that $dist(D, \{D' : f(D') \neq f(D)\}) = \left\lceil \frac{\text{occ}_1 - \text{occ}_2}{2} \right\rceil$, we know u_{apx} satisfies (ε, δ)-DP. By [Vad17, Proposition 3.4], when $\text{occ}_1 - \text{occ}_2 \geq 4\lceil\ln(1/\delta)/\varepsilon\rceil$, u_{apx} is the exact mode, i.e., $u_{\text{apx}} = f(D)$, with probability at least $1 - \delta$. Choosing $\delta < \frac{\varepsilon^d}{(2d)^{3d}}$, by Theorem 2 and the union bound, we have $\mathbb{P}\left(u_{\text{pure}} = f(D)\right) \geq 1 - \delta - 2^{-d} - \frac{d}{2}e^{-d} \geq 1 - \frac{3}{|\mathcal{X}|}$. ■

J.4 Private Linear Regression Through Adaptive Sufficient Statistics Perturbation

We investigate the problem of differentially private linear regression. Specifically, we consider a fixed design matrix $X \in \mathbb{R}^{n\times d}$ and a response variable $Y \in \mathbb{R}^n$, which represent a collection of data points $(x_i, y_i)_{i=1}^n$, where $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$. Assuming that there exists $\theta^* \in \Theta$ such that $Y = X\theta^*$, our goal is to find a differentially private estimator θ that minimizes the mean squared error:

$$\text{MSE}(\theta) = \frac{1}{2n}\|Y - X\theta\|_2^2$$

We assume prior knowledge of the magnitude of the dataspace: $\|\mathcal{X}\| := \sup_{x\in\mathcal{X}}\|x\|_2$ and $\|\mathcal{Y}\| := \sup_{y\in\mathcal{Y}}\|y\|_2$. Our algorithm operates under the same parameter settings as [Wan18, Algorithm 2]. To enable our purification procedure, we first derive a high-probability upper bound on $\|\tilde{\theta}\|_2$, where

Algorithm 13: $\mathcal{M}_{\text{pure}-\text{AdaSSP}}$

Input: Data X, y; privacy parameters $\epsilon, \varepsilon', \delta$, Bounds: $\|\mathcal{X}\|$, $\|\mathcal{Y}\|$, level of smoothing ω

$\theta_{apx} \leftarrow \text{AdaSSP}(X, y, \varepsilon, \delta, \|\mathcal{X}\|, \|\mathcal{Y}\|)$ ▷ [Wan18, Algorithm 2]

Propose a high probability upper bound $\tilde{R} = \tilde{\mathcal{O}}\left((1+n\varepsilon/d)\|\mathcal{Y}\|/\|\mathcal{X}\|\right)$

Construct trust region $\Theta := \mathcal{B}^d_{\ell_2}(\tilde{R})$

Norm clipping $\theta_{apx} \leftarrow \text{Proj}_{\Theta}(\theta_{apx})$

$\theta_{\text{pure}} \leftarrow \mathcal{A}_{\text{pure}}(\theta_{apx}, \Theta, \varepsilon, \varepsilon', \delta)$

Output: θ_{pure}

$\tilde{\theta}$ is the output of AdaSSP. Subsequently, we clip the output of AdaSSP and apply the purification procedure. The implementation details are provided in Algorithm 13.

Theorem 12 (Restatement of theorem 7) *Assume $X^\top X$ is positive definite and $\|\mathcal{Y}\| \lesssim \|\mathcal{X}\|\|\theta^*\|$. With probability $1-\zeta-1/n^2$, the output θ_{pure} from Algorithm 13 satisfies:*

$$\text{MSE}(\theta_{pure}) - \text{MSE}(\theta^*) \leq \tilde{\mathcal{O}}\left(\frac{\|\mathcal{X}\|^2}{\varepsilon^2 n^4} + \frac{d\|\mathcal{X}\|^2\|\theta^*\|^2}{n\epsilon} \wedge \frac{d^2\|\mathcal{X}\|^4\|\theta^*\|^2}{\epsilon^2 n^2 \lambda_{\min}}\right) \tag{21}$$

where $\lambda_{\min} := \lambda_{\min}(X^\top X/n)$.

Proof First, we introduce a utility Lemma from [Wan18]

Lemma 35 (Theorem 3 from [Wan18]) *Under the setting of [Wan18, Algorithm 2], AdaSSP satisfies (ε, δ)-DP. Assume $\|\mathcal{Y}\| \lesssim \|\mathcal{X}\|\|\theta^*\|$, then with probability $1-\zeta$,*

$$MSE(\tilde{\theta}) - MSE(\theta^*) \leq \mathcal{O}\left(\frac{\sqrt{d\log\left(\frac{d^2}{\zeta}\right)}\|\mathcal{X}\|^2\|\theta^*\|^2}{n\epsilon/\sqrt{\log\left(\frac{6}{\delta}\right)}} \wedge \frac{\|\mathcal{X}\|^4\|\theta^*\|^2\,\text{tr}[(X^TX)^{-1}]}{n\epsilon^2/[\log\left(\frac{6}{\delta}\right)\log\left(\frac{d^2}{\zeta}\right)]}\right) \tag{22}$$

Now, we prove the utility guarantee for our results. In order to operate the purification technique, we need to estimate the range $\|\tilde{\theta}\|_2$ in order to apply the uniform smoothing technique. Notice that

$$\|\tilde{\theta}\|_2 \leq \underbrace{\|(X^\top X + \lambda I_d + E_1)^{-1}\|_2}_{(a)} \underbrace{\|X^\top y + E_2\|_2}_{(b)} \tag{23}$$

with $E_2 \sim \frac{\sqrt{\log(6/\delta)}\|\mathcal{X}\|\|\mathcal{Y}\|}{\varepsilon/3}\mathcal{N}(0, I_d)$ and $E_1 \sim \frac{\sqrt{\log(6/\delta)}\|\mathcal{X}\|^2}{\varepsilon/3}Z$, where $Z \in \mathbb{R}^{d\times d}$ is a symmetric matrix, and each entry in its upper-triangular part (including the diagonal) is independently sampled from $\mathcal{N}(0,1)$. *Under the high probability event in Lemma 35*, we upper bound (a) and (b) separately:

For (a):

$$\|(X^\top X + \lambda I_d + E_1)^{-1}\|_2 = \frac{1}{\lambda_{\min}(X^\top X + \lambda I_d + E_1)}$$

By the choice of λ and concentration of $\|E_1\|_2$, we have $(X^\top X + \lambda I_d + E_1) \succ \frac{1}{2}(X^\top X + \lambda I_d)$, which allows a lower bound on $\lambda_{\min}(X^\top X + \lambda I_d + E_1)$:

$$\begin{aligned}
2\lambda_{\min}(X^\top X + \lambda I_d + E_1) &\geq \lambda_{\min}(X^\top X + \lambda I_d) \\
&\geq \lambda_{\min}(X^\top X) + \lambda \\
&\geq \lambda_{\min}(X^\top X) + \frac{\sqrt{d\log(6/\delta)\log(2d^2/\zeta)}\|\mathcal{X}\|^2}{\varepsilon/3} - \lambda^*_{\min} \\
&\geq \frac{\sqrt{d\log(6/\delta)\log(2d^2/\zeta)}\|\mathcal{X}\|^2}{\varepsilon/3}
\end{aligned}$$

where the third inequality is by setting $\lambda = \max\left\{0, \frac{\sqrt{d\log(6/\delta)\log(2d^2/\zeta)}\|\mathcal{X}\|^2}{\varepsilon/3} - \lambda^*_{\min}\right\}$, where $\lambda^*_{\min}$ is a high probability lower bound on $\lambda_{\min}(X^\top X)$. Thus,

$$\begin{aligned}
\|X^\top X + \lambda I_d + E_1\|_2 &= \frac{1}{\lambda_{\min}(X^\top X + \lambda I_d + E_1)} \\
&\leq \frac{2}{\lambda_{\min}(X^\top X + \lambda I_d)} \\
&\leq \frac{2\varepsilon}{3\sqrt{d\log(6/\delta)\log(2d^2/\zeta)}\|\mathcal{X}\|^2} \\
&= \tilde{\mathcal{O}}\left(\frac{\varepsilon}{\sqrt{d\log(6/\delta)}\|\mathcal{X}\|^2}\right)
\end{aligned}$$

For (b), by triangle inequality

$$(b) = \|X^\top y + E_2\|_2 \leq \|X^\top y\|_2 + \|E_2\|_2 \leq n\|\mathcal{X}\|\|\mathcal{Y}\| + \|E_2\|_2$$

Apply [Wan18, Lemma 6], we have w.p. at least $1-\beta$:

$$\|E_2\|_2 = \mathcal{O}\left(\frac{\sqrt{d}\|\mathcal{X}\|\|\mathcal{Y}\|\sqrt{\log(6/\delta)\log(d/\beta)}}{\varepsilon}\right)$$

Thus, w.p. at least $1-\beta$ over the randomness of E_2,

$$\begin{aligned}
(b) &\leq \mathcal{O}\left(n\|\mathcal{X}\|\|\mathcal{Y}\| + \frac{\sqrt{d}\|\mathcal{X}\|\|\mathcal{Y}\|\sqrt{\log(6/\delta)\log(d/\beta)}}{\varepsilon}\right) \\
&= \tilde{\mathcal{O}}\left(n\|\mathcal{X}\|\|\mathcal{Y}\| + \frac{\sqrt{d}\|\mathcal{X}\|\|\mathcal{Y}\|\sqrt{\log(6/\delta)}}{\varepsilon}\right)
\end{aligned}$$

Putting everything together under the high probability event:

$$\begin{aligned}
\|\tilde{\theta}\|_2 &\leq \tilde{\mathcal{O}}\left(\frac{\|\mathcal{Y}\|}{\|\mathcal{X}\|}\left(1 + \frac{n\varepsilon}{\sqrt{d\log(6/\delta)}}\right)\right) \\
&\leq \tilde{\mathcal{O}}\left(\frac{\|\mathcal{Y}\|}{\|\mathcal{X}\|}\left(1 + \frac{n\varepsilon}{d}\right)\right) := \tilde{r}
\end{aligned}$$

Now, we apply purification Algorithm 1 with $\Theta = \mathcal{B}^d_{\ell_2}(\tilde{r})$, $\omega = \frac{1}{n^2}$, $\delta = \frac{2\omega}{(16d^{3/2}\tilde{r}n^2)^d}$. This parameter configuration implies $\log(1/\delta) = \tilde{\mathcal{O}}(d)$ and Wasserstein-∞ distance $\Delta = \frac{1}{4n^2 d}$ (Line 2 of Algorithm 1)

Finally, it remains to bound the estimation error for θ_{pure}. Let's denote the additive Laplace noise from purification by $Z_2 \sim \text{Lap}^{\otimes d}(\Delta/\varepsilon)$, and the purified estimator $\theta_{pure} := \tilde{\theta} + Z_2$. Under the event that the purification algorithm doesn't output uniform noise, which happens w.p. at least $1-\omega$:

$$\begin{aligned}
\text{MSE}(\theta_{pure}) - \text{MSE}(\tilde{\theta}) &= \frac{1}{2n}\|y - X\theta_{pure}\|_2^2 - \frac{1}{2n}\|y - X\tilde{\theta}\|_2^2 \\
&= \frac{1}{n}Z_2^\top X^\top X Z_2 \\
&\leq \frac{1}{n}\lambda_{\max}(X^\top X)\,\|Z_2\|_2^2 \\
&\leq \|\mathcal{X}\|^2\|Z_2\|_1^2 \\
&\leq \frac{\|\mathcal{X}\|^2}{\varepsilon^2 n^4}
\end{aligned}$$

where the last inequality holds w.p. $1-1/n^2$ by Lemma 48, as we derived below:

$$\begin{aligned}\|Z_2\|_1 &\le \frac{2d\Delta}{\varepsilon}+\frac{2\Delta}{\varepsilon}\log(n^2)\\ &\le \tilde{\mathcal{O}}\left(\frac{1}{\varepsilon n^2}\right)\end{aligned}$$

Under the high probability event of Lemma 35 and Algorithm 1, which happens with probability at least $1-\zeta-1/n^2$:

$$\begin{aligned}\mathrm{MSE}(\theta_{pure})-\mathrm{MSE}(\theta^*) &= \mathrm{MSE}(\theta_{pure})-\mathrm{MSE}(\tilde{\theta})+\mathrm{MSE}(\tilde{\theta})-\mathrm{MSE}(\theta^*)\\ &\le \tilde{\mathcal{O}}\left(\frac{\|\mathcal{X}\|^2}{\varepsilon^2 n^4}+\frac{d\|\mathcal{X}\|^2\|\theta^*\|^2}{n\epsilon}\wedge\frac{d\|\mathcal{X}\|^4\|\theta^*\|^2\,\mathrm{tr}[(X^TX)^{-1}]}{n\epsilon^2}\right) \qquad (24)\\ &\le \tilde{\mathcal{O}}\left(\frac{\|\mathcal{X}\|^2}{\varepsilon^2 n^4}+\frac{d\|\mathcal{X}\|^2\|\theta^*\|^2}{n\epsilon}\wedge\frac{d^2\|\mathcal{X}\|^4\|\theta^*\|^2}{n\epsilon^2\lambda_{\min}(X^\top X)}\right)\end{aligned}$$

By denoting $\lambda_{\min}=\lambda_{\min}\left(\frac{X^\top X}{n}\right)$, we have:

$$\mathrm{MSE}(\theta_{pure})-\mathrm{MSE}(\theta^*)\le\tilde{\mathcal{O}}\left(\frac{\|\mathcal{X}\|^2}{\varepsilon^2 n^4}+\frac{d\|\mathcal{X}\|^2\|\theta^*\|^2}{n\epsilon}\wedge\frac{d^2\|\mathcal{X}\|^4\|\theta^*\|^2}{\epsilon^2 n^2\lambda_{\min}}\right) \qquad (25)$$

■

K Deferred Proofs for Private Query Release

K.1 Problem Setting

Let data universe $\mathcal{X}=\{0,1\}^d$ and denote $N:=|\mathcal{X}|$. The dataset, $D\in\mathcal{X}^n$ is represented as a histogram $D\in\mathbb{N}^{|\mathcal{X}|}$ with $\|D\|_1=n$. We consider bounded linear query function $q:\mathcal{X}\to[0,1]$ and workload Q with size K. For a shorthand, we denote:

$$Q(D):=(q_1(D),\dots,q_k(D))^\top:=\left(\frac{1}{n}\sum_{i\in[n]}q_1(d_i),\dots,\frac{1}{n}\sum_{i\in[n]}q_k(d_i)\right)^\top$$

K.2 Private Multiplicative Weight Exponential Mechanism

We first introduce private multiplicative weight exponential algorithm (MWEM):

Algorithm 14: Multipliative Weight Exponential Mechanism $\texttt{MWEM}(D,Q,T,\rho)$ [HLM12]

Input: Data set $D\in\mathbb{N}^{|\mathcal{X}|}$, set Q of linear queries; Number of iterations $T\in\mathbb{N}$; zCDP Privacy parameter $\rho>0$.

Set: number of data points $n\leftarrow\|D\|_1$, initial distribution $p_0\leftarrow\frac{\mathbf{1}_{|\mathcal{X}|}}{|\mathcal{X}|}$, privacy budget for each mechanism $\varepsilon_0\leftarrow\sqrt{\rho/T}$

Define: $\mathrm{Score}(\,\cdot\,;\hat{p},p)=|\langle\cdot,\hat{p}\rangle-\langle\cdot,p\rangle|$

for $t=1$ *to* T **do**

  $q_t\leftarrow\texttt{ExpoMech}\left(Q,\varepsilon_0,\mathrm{Score}(\,\cdot\,;p_{t-1},p)\right)$

  $m_t\leftarrow\langle q_t,X\rangle+\mathrm{Laplace}(1/n\varepsilon_0)$

  Multiplicative weights update: let p_{t+1} be the distribution over $\mathcal{X}$ with entries satisfy:

$$q_t(x)\propto q_{t-1}(x)\cdot\exp(q_t(x)\cdot(m_t-q_t(p_{t-1}))/2),\ \forall x\in\mathcal{X}$$

Output: $D_{\text{out}}\leftarrow\frac{n}{T}\sum_{t=0}^{T-1}p_t$.

Algorithm 15: `ProportionalSampling`$(\mathcal{X}, p, m)$

Input: Dataspace $\mathcal{X}$, Probability vector $p \in \mathbb{R}^{|\mathcal{X}|}$, sample size m

for $i = 1$ *to* m **do**

  $s_i \leftarrow$ `UnSortedProportionalSampling`$(p, \mathcal{X})$ ▷ e.g., Alias method [Wal74]

Output: $\{s_1, ...s_m\}$

Algorithm 16: Pure DP Multiplicative Weight Exponential Mechanism

Input: Dataset $D \in \mathbb{N}^{|\mathcal{X}|}$ with size $\|D\|_1 = n$, Query set Q, privacy parameters ϵ, δ, accuracy parameter α

Set: Number of iterations $T = \tilde{\mathcal{O}}(\varepsilon^{2/3} n^{2/3} d^{1/3})$, size of subsampled dataset $m = n^{2/3}\varepsilon^{2/3} d^{-2/3}$, zCDP budget $\rho = \varepsilon^2/16\log(1/\delta)$, $\text{Score}(q; \hat{p}, p) = |\langle q, \hat{p}\rangle - \langle q, p\rangle|, \ \forall q \in Q$

Initilize: $p_1 \leftarrow \frac{\mathbf{1}_{|\mathcal{X}|}}{|\mathcal{X}|}, \ p \leftarrow \frac{D}{\|D\|_1}$

$\hat{D} \leftarrow$ `MWEM`(D, Q, T, ρ) ▷ Algorithm 14

$Y \leftarrow$ `ProportionalSampling`$(\mathcal{X}, \hat{D}/n, m)$ ▷ Algorithm 16

$\hat{Y} \leftarrow \mathcal{A}_{\text{pure-discrete}}(\varepsilon, \delta, Y, \mathcal{X}^m)$

Output: $\hat{Y}$

Lemma 36 (Privacy and Utility of MWEM [HLM12]) *Algorithm 14 instantiated as* `MWEM`$(D, Q, T, \varepsilon^2/16\log(1/\delta))$ *satisfies* (ε, δ)*-DP. With probability* $1 - 2\beta T$*, PMW and the output* $\hat{D}$ *such that:*

$$\|Q(\hat{D}) - Q(D)\|_\infty \leq \mathcal{O}\left(\sqrt{\frac{d}{T}} + \frac{\sqrt{T\log(1/\delta)}\log(K/\beta)}{n\varepsilon}\right) \tag{26}$$

Proof We first state the privacy guarantee. Since each iteration satisfies zCDP guarantee ρ/T-zCDP, the total zCDP guarantee for T iterations is ρ. By Lemma 22, the whole algorithm satisfies $(4\sqrt{\rho\log(1/\delta)}, \delta)$-DP. Plugging in the choice of $\rho = \varepsilon^2/16\log(1/\delta)$, we have `MWEM`$(D, Q, T, \varepsilon^2/16\log(1/\delta))$ satisfies (ε, δ)-DP. The utility guarantee follows [HLM12, Theorem 2.2]. Specifically, we choose $\text{adderr} = 2\sqrt{T/\rho}\log(K/\beta)$, this yields the utility guarantee stated in Theorem with probability $1 - 2\beta T$. ■

Lemma 37 (Sampling bound, Lemma 4.3 in [DR+14]) *Let data* $X = (a_1, ..., a_N)$ *with* $\sum_{i=1}^N a_i = 1$ *and* $a_i \geq 0$. $Y \sim \text{Multinomial}(m, X)$. *Then we have:*

$$\mathbb{P}[\|Q(Y) - Q(X)\|_\infty \geq \alpha] \leq 2|Q|\exp(-2m\alpha^2)$$

Proof The proof follows the proof of [DR+14, Lemma 4.3]. Since we have $Y = (Y_1, ..., Y_m)$ with $Y_i \overset{\text{iid}}{\sim} \text{Multinomial}(1, X)$, for any $q \in Q$, we have $q(Y) = \frac{1}{m}\sum_{i=1}^m q(Y_i)$ and $\mathbb{E}[q(Y)] = q(X)$. By the Chernoff bound and a union bound over all queries in Q, we have $\mathbb{P}[\|Q(Y) - Q(X)\|_\infty \geq \alpha] \leq 2|Q|\exp(-2m\alpha^2)$. ■

Theorem 13 (Restatement of Theorem 8) *Algorithm 16 satisfies* 2ε*-DP. Moreover, the output* $\hat{Y}$ *satisfies*

$$\|Q(D) - Q(\hat{Y})\|_\infty \leq \tilde{\mathcal{O}}\left(\frac{d^{1/3}}{n^{1/3}\varepsilon^{1/3}}\right),$$

and the runtime is $\tilde{\mathcal{O}}(nK + \varepsilon^{2/3} n^{2/3} d^{1/3} NK + N)$.

Proof We first state the privacy guarantee. By Lemma 36, the output from multiplicative weight exponential mechanism, $\hat{D}$, satisfies (ε, δ)-DP. By post-processing, Y is also (ε, δ)-DP. Thus, apply Theorem 2, the purified $\hat{Y}$ satisfies 2ε-DP.

The utility guarantee is via bounding the following terms:

$$\|Q(D) - Q(\hat{Y})\|_\infty \leq \underbrace{\|Q(D) - Q(\hat{D})\|_\infty}_{(a)} + \underbrace{\|Q(\hat{D}) - Q(Y)\|_\infty}_{(b)} + \underbrace{\|Q(Y) - Q(\hat{Y})\|_\infty}_{(c)}$$

For (c): Since the output space $\mathcal{Y} = \mathcal{X}^m$, if use binary encoding, the length of code is $\log_2(|\mathcal{Y}|) = md := \tilde{d}$. Thus, by Theorem 2, we choose $\delta = \frac{\varepsilon^{\tilde{d}}}{(2\tilde{d})^{3\tilde{d}}}$, this implies $\log(1/\delta) = \mathcal{O}(md\log(2md/\varepsilon))$ and failure probability $\beta_0 = \frac{1}{2^{md}} + \frac{md}{\exp(md)}$.

For (a): In order to minimize upper bound in Equation 26, we choose $T = \frac{d^{1/2}n\varepsilon}{\log^{1/2}(1/\delta)\log(K/\beta)}$, this implies $\|Q(D) - Q(\hat{D})\|_\infty \leq \frac{d^{1/4}\log^{1/4}(1/\delta)\log^{1/2}(K/\beta)}{(n\varepsilon)^{1/2}} = \mathcal{O}\left(\frac{d^{1/2}m^{1/4}\log^{1/4}(2md/\varepsilon)\log^{1/2}(K/\beta)}{(n\varepsilon)^{1/2}}\right)$

For (b): using Sampling bound (Lemma 37) and setting failure probability $\beta_1 = 2K\exp(-2m\alpha^2)$, we have $\|Q(\hat{D}) - Q(Y)\|_\infty \leq \mathcal{O}\left(\frac{\log^{1/2}(2K/\beta_1)}{m^{1/2}}\right)$

Finally, we choose $m = (n\varepsilon/d)^{2/3}$ to balance the error between (a) and (b). This implies:

$$\|Q(\hat{D}) - Q(Y)\|_\infty \leq \mathcal{O}\left(\frac{d^{1/3}}{(n\varepsilon)^{1/3}}\left(\log^{1/2}(2K/\beta_1) + \log^{1/2}(K/\beta)\log^{1/4}(2d^{1/3}n^{2/3}\varepsilon^{-1/3})\right)\right)$$

Set $\beta = \frac{1}{2Tn}$ and $\beta_1 = \frac{1}{n}$, and by $\beta_0 \leq \frac{2(n\varepsilon)^{2/3}d^{1/3}}{2^{(n\varepsilon)^{2/3}d^{1/3}}} = o(\frac{1}{n})$, we have

$$\begin{aligned}\mathbb{E}[\|Q(\hat{D}) - Q(Y)\|_\infty] &\leq \mathcal{O}\left(\frac{d^{1/3}}{(n\varepsilon)^{1/3}}\left(\log^{1/2}(2nK) + \log^{1/2}(Kd^{1/3}n^{5/3}\varepsilon^{1/3})\log^{1/4}(2d^{1/3}n^{2/3}\varepsilon^{-1/3})\right)\right) \\ &\quad + (T\beta + \beta_1 + \beta_0) \\ &= \tilde{\mathcal{O}}\left(\frac{d^{1/3}}{(n\varepsilon)^{1/3}} + \frac{1}{n}\right)\end{aligned}$$

The computational guarantee follows: (1) Since $T = \frac{d^{1/2}n\varepsilon}{\log^{1/2}(1/\delta)\log(K/\beta)} = \tilde{\mathcal{O}}\left(\varepsilon^{2/3}n^{2/3}d^{1/3}\right)$. The runtime for MWEM is $\tilde{\mathcal{O}}(nK + \varepsilon^{2/3}n^{2/3}d^{1/3}NK)$ [HLM12]; (2) For subsampling, by runtim analysis of Alias method [Wal74], the preprocessing time is $\mathcal{O}(N)$ and the query time is $\mathcal{O}(m)$ for generating m samples. Thus, total runtime is $\mathcal{O}(N + (n\varepsilon)^{2/3}d^{-2/3})$; (3) Finally, note that the query time for Algorithm 2 is $\mathcal{O}(\tilde{d}) = \mathcal{O}(d^{1/3}(n\varepsilon)^{2/3})$. We conclude that the runtime for Algorithm 16 is $\tilde{\mathcal{O}}(nK + \varepsilon^{2/3}n^{2/3}d^{1/3}NK + N)$. ■

L Deferred Proofs for Lower Bounds

L.1 Proof of Theorem 9

Lemma 38 (Lemma 5.1 in [BST14]) *Let $n, d \in \mathbb{N}$ and $\epsilon > 0$. There is a number $M = \Omega(\min(n, d/\epsilon))$ such that for every ϵ-differentially private algorithm $\mathcal{A}$, there is a dataset $D = \{x_1, \ldots, x_n\} \subset \{-1/\sqrt{d}, 1/\sqrt{d}\}^d$ with $\|\sum_{i=1}^n x_i\|_2 \in [M-1, M+1]$ such that, with probability at least $1/2$ (taken over the algorithm random coins), we have*

$$\left\|\mathcal{A}(D) - \bar{D}\right\|_2 = \Omega\left(\min\left(1, \frac{d}{\epsilon n}\right)\right)$$

where $\bar{D} = \frac{1}{n}\sum_{i=1}^n x_i$.

Theorem 14 (Restatement of Theorem 9) *Denote $\mathcal{D} := \{-1/\sqrt{d}, 1/\sqrt{d}\}^d$. Let $\varepsilon \leq \mathcal{O}(1)$, and $\delta \in \left(\frac{1}{\exp(4d\log(d)\log^2(nd))}, \frac{1}{4n^d\log^{2d}(8d)}\right)$. For any (ε, δ)-DP mechanism $\mathcal{M}$, there exist a dataset $D \in \mathcal{D}^n$ such that w.p. at least $1/4$ over the randomness of $\mathcal{M}$:*

$$\|\mathcal{M}(D) - \bar{D}\|_2 \geq \tilde{\Omega}\left(\frac{\sqrt{d\log(1/\delta)}}{\varepsilon n}\right)$$

Here, $\tilde{\Omega}(\cdot)$ hides all polylogarithmic factors, except those with respect to δ.

Proof Suppose there exists an (ε, δ)-differentially private mechanism $\mathcal{M}$ such that with probability at least $3/4$ over the randomness of $\mathcal{M}$, for any $D \in \mathcal{D}$,

$$\|\mathcal{M}(D) - \bar{D}\|_2 \leq \frac{\sqrt{d \log(1/\delta)}}{n\varepsilon a}$$

where a is a term involving n and d, to be specified later. Let $\frac{\sqrt{d \log(1/\delta)}}{n\varepsilon a} \leq \frac{d}{n\varepsilon \log^{1/2} d}$ implies:

$$\delta > \exp(-a^2 d / \log(d)) \tag{27}$$

We execute Algorithm 1 to purify $\mathcal{M}$ directly over the output space $[-1/\sqrt{d}, 1/\sqrt{d}]^d$. Let Y denote the output of Algorithm 1 and $U \sim \mathrm{Unif}([-1/\sqrt{d}, 1/\sqrt{d}]^d)$. The remainder of the proof involves bounding the additional errors introduced during the purification process. By triangle inequality we have

$$\|\bar{D} - Y\|_2 \leq \underbrace{\|\bar{D} - \mathcal{M}(D)\|_2}_{\text{(a)}} + \underbrace{\|\mathcal{M}(D) - Y\|_2}_{\text{(b)}}.$$

Notice that under the event that Line 3 of Algorithm 1 doesn't return the uniform random variable, which happens with probability $1 - \omega$, we have $Y = \mathcal{M}(X) + \mathrm{Laplace}^{\otimes d}(2\Delta/\varepsilon)$, so term (b) equals the 2-norm of the Laplace perturbation.

For the remaining proofs, we choose the mixing level $\omega = 1/8$ in Algorithm 1. We now justify the choice of δ:

Observe that since $Y = \mathcal{M}(X) + \mathrm{Laplace}^{\otimes d}(2\Delta/\varepsilon)$, term (b), which accounts for the error introduced by Laplace noise. With probability at least $7/8$ by the concentration of the L_2 norm of Laplace vector:

$$\text{(b)} \leq \frac{2\sqrt{d}\Delta \log(8d)}{\varepsilon}$$

Thus, without loss of generality, we require $\frac{2\sqrt{d}\Delta \log(8d)}{\varepsilon} \leq \frac{16d}{n\varepsilon \log(8d)}$, this implies

$$\Delta \leq \frac{8\sqrt{d}}{n \log^2(8d)}$$

Notice that:

$$\Delta = d^{1-\frac{1}{q}} \cdot \frac{2R^2}{r} \left(\frac{\delta}{2\omega}\right)^{1/d}$$

Choosing $q = \infty$ (corresponding to the use of ℓ_∞ norm in W-∞ distance), and noticing $R = 2/\sqrt{d}$ and $r = 1/\sqrt{d}$, we obtain the condition:

$$\Delta = 8\sqrt{d} \left(\frac{\delta}{2\omega}\right)^{1/d} \leq \frac{8\sqrt{d}}{n \log^2(8d)}$$

which further implies:

$$\delta \leq \frac{1}{4n^d \log^{2d}(8d)} \tag{28}$$

By Eq. (32) and Eq. (33), we have:

$$\delta \in \left(\exp(-a^2 d / \log(d)), \quad \frac{1}{4n^d \log^{2d}(8d)}\right) \tag{29}$$

The constrained above yields a lower bound on a^2, after some relaxation for simplicity (and assume $d \geq \log(8d)$ and $d \log(n) > \log(4)$):

$$a^2 \geq 2 \log(d) \log(nd) \tag{30}$$

Thus, we set $a = 2\log(d)\log(nd)$ to satisfy the constraint stated in Eq. (35), and now

$$\delta \in \left(\frac{1}{\exp(4d\log(d)\log^2(nd))}, \quad \frac{1}{4n^d\log^{2d}(8d)}\right) \tag{31}$$

When δ is within the range above, we have:

$$\log(1/\delta) < 4d\log(d)\log^2(nd)$$

This implies that, w.p. at least $1/2$ over the randomness of $\mathcal{M}$ and purification algorithm:

$$\|\bar{D} - Y\|_2 \leq \frac{d}{n\varepsilon\log^{1/2}(d)} + \frac{16d}{n\varepsilon\log(8d)},$$

which violates the lower bound stated in Lemma 38. Thus, for any (ε, δ)-DP mechanism $\mathcal{M}$ with δ being in the range of Eq. (36), there exists a dataset $D \in \mathcal{D}$, such that with probability greater than $1/4$ over the randomness of $\mathcal{M}$:

$$\|\mathcal{M}(D) - \bar{D}\|_2 \geq \left(\frac{\sqrt{d\log(1/\delta)}}{2n\varepsilon\log(d)\log(nd)}\right) = \tilde{\Omega}\left(\frac{\sqrt{d\log(1/\delta)}}{n\varepsilon}\right)$$

Here, $\tilde{\Omega}(\cdot)$ hides all polylogarithmic factors, except those with respect to δ. ■

L.2 More Examples of Lower Bounds via the Purification Trick

In this section, we present an extended result for Theorem 9. Additionally, we establish a lower bound for the discrete setting, as stated in Theorem 16, thereby demonstrating that the purifying recipe for proving lower bound remains applicable in the discrete case.

L.2.1 One-Way Marginal Release

We establish a stronger version of Theorem 9, as stated in Theorem 15, which strengthens Theorem 9 by establishing that the lower bound holds for any $c \in (0, 1)$, rather than being restricted to $c = 1/2$.

Theorem 15 (Restatement of Theorem 9) *Denote $\mathcal{D} := \{-1/\sqrt{d}, 1/\sqrt{d}\}^d$. Let $\varepsilon \leq \mathcal{O}(1)$, for any $c \in (0, 1)$, and $\delta \in \left(\frac{1}{n^{2d}d^{2d}}, \frac{1}{4n^d\log^{2d}(8d)}\right)$. For any (ε, δ)-DP mechanism $\mathcal{M}$, there exist a dataset $D \in \mathcal{D}^n$ such that with probability at least $1/4$ over the randomness of $\mathcal{M}$:*

$$\|\mathcal{M}(D) - \bar{D}\|_2 \geq \tilde{\Omega}\left(\max_{c\in(0,1)} \frac{d^c\log^{1-c}(1/\delta)}{\varepsilon n}\right).$$

Here, $\tilde{\Omega}(\cdot)$ hides all polylogarithmic factors, except those with respect to δ.

Proof Suppose for some $c \in (0, 1)$, there exists an (ε, δ)-differentially private mechanism $\mathcal{M}$ such that with probability at least $3/4$ over the randomness of $\mathcal{M}$, for any $D \in \mathcal{D}$,

$$\|\mathcal{M}(D) - \bar{D}\|_2 \leq \frac{d^c\log^{1-c}(1/\delta)}{n\varepsilon a}$$

where a is a term involving n and d, to be specified later. For the purpose of causing contradiction, we let:

$$\frac{d^c\log^{1-c}(1/\delta)}{n\varepsilon a} \leq \frac{d}{n\varepsilon\log^{1-c}(d)}$$

This implies:

$$\delta > \exp(-a^{\frac{1}{1-c}}d/\log(d)) \tag{32}$$

We execute Algorithm 1 to purify $\mathcal{M}$ directly over the output space $[-1/\sqrt{d}, 1/\sqrt{d}]^d$. Let Y denote the output of Algorithm 1 and $U \sim \mathrm{Unif}([-1/\sqrt{d}, 1/\sqrt{d}]^d)$. The remainder of the proof involves bounding the additional errors introduced during the purification process. By triangle inequality we have

$$\|\bar{D} - Y\|_2 \leq \underbrace{\|\bar{D} - \mathcal{M}(D)\|_2}_{\text{(a)}} + \underbrace{\|\mathcal{M}(D) - Y\|_2}_{\text{(b)}}.$$

Notice that under the event that Line 3 of Algorithm 1 doesn't return the uniform random variable, which happens with probability $1-\omega$, we have $Y = \mathcal{M}(X) + \text{Laplace}^{\otimes d}(2\Delta/\varepsilon)$, so term (b) equals the 2-norm of the Laplace perturbation.

For the remaining proofs, we choose the mixing level $\omega = 1/8$ in Algorithm 1. We now justify the choice of δ:

Observe that since $Y = \mathcal{M}(X) + \text{Laplace}^{\otimes d}(2\Delta/\varepsilon)$, term (b), which accounts for the error introduced by Laplace noise. With probability at least $7/8$ by the concentration of the L_2 norm of Laplace vector:

$$\text{(b)} \leq \frac{2\sqrt{d}\Delta \log(8d)}{\varepsilon}$$

Thus, without loss of generality, we will require

$$\frac{2\sqrt{d}\Delta \log(8d)}{\varepsilon} \leq \frac{16d}{n\varepsilon \log(8d)}$$

This implies:

$$\Delta \leq \frac{8\sqrt{d}}{n \log^2(8d)}$$

Notice that:

$$\Delta = d^{1-\frac{1}{q}} \cdot \frac{2R^2}{r} \left(\frac{\delta}{2\omega}\right)^{1/d}$$

Choosing $q = \infty$ (corresponding to the use of ℓ_∞ norm in the Wassertain-∞ distance), and noticing $R = 2/\sqrt{d}$ and $r = 1/\sqrt{d}$, we obtain the condition:

$$\Delta = 8\sqrt{d}\left(\frac{\delta}{2\omega}\right)^{1/d} \leq \frac{8\sqrt{d}}{n \log^2(8d)}$$

which further implies:

$$\delta \leq \frac{1}{4n^d \log^{2d}(8d)} \tag{33}$$

By Eq. (32) and Eq. (33), we have:

$$\delta \in \left(\exp\left(-a^{\frac{1}{1-c}} d/\log(d)\right), \quad \frac{1}{4n^d \log^{2d}(8d)}\right) \tag{34}$$

The constrained above yields a lower bound on a, after some relaxation for simplicity (and assume $d \geq \log(8d)$ and $d\log(n) > \log(4)$):

$$a^{\frac{1}{1-c}} \geq 2\log(d)\log(nd) \tag{35}$$

Thus, we set $a = (2\log(d)\log(nd))^{1-c}$ to satisfy the constraint stated in Eq. (35), and now

$$\delta \in \left(\frac{1}{\exp(2d\log(nd))}, \quad \frac{1}{4n^d \log^{2d}(8d)}\right) = \left(\frac{1}{n^{2d}d^{2d}}, \quad \frac{1}{4n^d \log^{2d}(8d)}\right) \tag{36}$$

When δ is within the range above, we have:

$$\log(1/\delta) < 2d\log(nd)$$

This implies that, w.p. at least $1/2$ over the randomness of $\mathcal{M}$ and purification algorithm:

$$\|\bar{D} - Y\|_2 \leq \frac{d}{n\varepsilon \log^{1-c}(d)} + \frac{16d}{n\varepsilon \log(8d)},$$

which violates the lower bound stated in Lemma 38. Thus, for any (ε, δ)-DP mechanism $\mathcal{M}$ with δ being in the range of Eq. (36), there exists a dataset $D \in \mathcal{D}$, such that with probability greater than $1/4$ over the randomness of $\mathcal{M}$:

$$\|\mathcal{M}(D) - \bar{D}\|_2 \geq \left(\frac{d^c \log^{1-c}(1/\delta)}{n\varepsilon \left(2\log(d)\log(nd)\right)^{1-c}} \right) = \tilde{\Omega}\left(\frac{d^c \log^{1-c}(1/\delta)}{n\varepsilon} \right)$$

Here, $\tilde{\Omega}(\cdot)$ hides all polylogarithmic factors, except those with respect to δ. Since the above derivation holds for arbitraty $c \in (0,1)$, this implies with probability at least $1/4$ over the randomness of the algorithm $\mathcal{M}$, we have:

$$\|\mathcal{M}(D) - \bar{D}\|_2 \geq \tilde{\Omega}\left(\max_{c\in(0,1)} \frac{d^c \log^{1-c}(1/\delta)}{n\varepsilon} \right)$$

■

L.2.2 Private Selection

We begin by stating a lower bound for pure differential privacy in the selection setting, as established in [CHS14].

Lemma 39 (Proposition 1 in [CHS14]) *Let $\varepsilon \in (0,1)$, $n \geq 2$ and denote item set to be $\mathcal{U}$. For any ε-DP mechanism $\mathcal{A}$, there exist a domain $\mathcal{X}$ and a function $f(i,\cdot)$ which is $(1/n)$-Lipschitz for all item $i \in \mathcal{U}$ such that the following holds with probability at least $1/2$ over the randomness of the algorithm:*

$$\max_{i\in\mathcal{U}} f(i;D) - f(\mathcal{A}(D);D) \geq \Omega\left(\frac{\log(K)}{\varepsilon n}\right).$$

Theorem 16 (Lower bound for private selection) *Let $\varepsilon \in (0,1)$, $\delta \in \left(\frac{\varepsilon^{3d}}{(2d)^{3d}}, \frac{\varepsilon^d}{(2d)^{3d}}\right)$, $n \geq 2$, and $K := |\mathcal{U}| \geq 7$ where $\mathcal{U}$ is the item set. For any (ε,δ)-DP mechanism $\mathcal{A}$, there exist a domain $\mathcal{X}$ and a function $f(i,\cdot)$ which is $(1/n)$-Lipschitz for all item $i \in \mathcal{U}$ such that the following holds with probability at least $1/2$ over the randomness of the algorithm:*

$$\max_{i\in\mathcal{U}} f(i;D) - f(\mathcal{A}(D);D) \geq \Omega\left(\max_{c\in(0,1)} \frac{\log^c K \log^{1-c}(1/\delta)}{\varepsilon n} \right).$$

Proof Without loss of generality, we set $d = \lceil \log_2 K \rceil$, we have that $\log K = \Theta(d)$. For any $c \in (0,1)$, assume there exists an (ε,δ)-DP algorithm such that with probability at least $\frac{1}{2} + 2^{-d} + \frac{d}{2\exp(d)}$, for any $D \in \mathcal{X}^n$, we have $\max_{i\in\mathcal{U}} f(i,D) - f(\mathcal{A}(D),D) = \Omega\left(\frac{d^c \log^{1-c}(1/\delta)}{\varepsilon n a}\right)$, with a being some term involved with n, d which will be specified later.

First, to ensure the quality of purification, we need to set $\delta \leq \varepsilon^d (2d)^{-3d}$, this ensures with probability at least $1 - 2^{-d} - \frac{d}{2}\exp(-d)$ over the randomness of purification algorithm, we have $\mathcal{A}^{\text{purified}}(D) = \mathcal{A}(D)$.

Further, in order to fulfill contrast argument, without loss of generality, we require

$$\frac{d^c \log^{1-c}(1/\delta)}{\varepsilon n a} \leq \frac{d}{\varepsilon n \log^{1-c}(d)}$$

which implies:

$$\delta > \exp(-a^{\frac{1}{1-c}} d/\log(d))$$

Thus, to ensure the lower bound of δ doesn't exceed the upper bound of δ, we require:

$$a^{\frac{1}{1-c}} \geq \log(d)\log\left(\frac{8d^3}{\varepsilon}\right)$$

So, we set $a^{\frac{1}{1-c}} = 3\log(d)\log\left(\frac{2d}{\varepsilon}\right)$, this implies

$$\delta \in \left(\frac{1}{\exp(3d\log(2d/\varepsilon))}, \quad \frac{\varepsilon^d}{(2d)^{3d}} \right)$$

This implies with probability at least $1/2$,

$$\max_{i\in\mathcal{U}} f(i;D) - f(\mathcal{A}^{\text{purified}}(D);D) \le \mathcal{O}\left(\frac{d}{\varepsilon n \log^{1-c}(d)}\right)$$

Observe that, under the assumption of $K := |\mathcal{U}| \ge 7$ which implies $d \ge \exp(1)$, the inequality above contradicts Lemma 39. Since $c \in (0,1)$ was chosen arbitrarily, this completes the proof of the stated theorem. ■

L.3 An alternative proof for Theorem 9

A $\tilde{\Omega}(d)$ lower bound for mean estimation under (ε,δ)-DP can also be proved using a packing argument when δ is exponentially small, as detailed in the theorem below.

Theorem 17 (Packing lower bound for (ε,δ)-DP mean estimation) *Fix constants $\varepsilon > 0$ and $\alpha \in (0,1]$. Let the data domain be $\mathcal{X} = [-1,1]^d$, and let $\mu(D) = \frac{1}{n}\sum_{i=1}^n x_i$ for $D \in \mathcal{X}^n$. Suppose a mechanism $\mathcal{M} : \mathcal{X}^n \to \mathbb{R}^d$ is (ε,δ)-DP under the replace-one neighboring relation and, for every dataset D, satisfies*

$$\mathbb{P}\big[\|\mathcal{M}(D) - \mu(D)\|_2 \le \alpha\big] \;\ge\; 2/3.$$

If $\delta \le \frac{1}{6}\,e^{-n\varepsilon}$, then

$$n \;\ge\; \frac{\log 2}{\varepsilon}\,(d-1).$$

Proof For each $v \in \{\pm 1\}^d$, set $\mu_v := v \in [-1,1]^d$ and define $D^{(v)} = (\mu_v,\dots,\mu_v) \in \mathcal{D}^n$. Then $\mu(D^{(v)}) = \mu_v$. For $u \ne v$, $\|\mu_u - \mu_v\|_2 = 2\sqrt{|\{j : u_j \ne v_j\}|} \ge 2$, so the α-balls

$$A_v := \{y \in \mathbb{R}^d :\; \|y - \mu_v\|_2 \le \alpha\}$$

are pairwise disjoint for any $\alpha \le 1$. By the accuracy assumption, for all v, we have

$$\mathbb{P}\big[\mathcal{M}(D^{(v)}) \in A_v\big] \;\ge\; 2/3. \tag{37}$$

If two datasets differ in at most h positions, then for any measurable S, by the group privacy, we have

$$\mathbb{P}[\mathcal{M}(D) \in S] \;\ge\; e^{-h\varepsilon}\Big(\mathbb{P}[\mathcal{M}(D') \in S] - \delta_h\Big), \qquad \delta_h \;\le\; \delta\sum_{i=0}^{h-1} e^{i\varepsilon} \;\le\; \delta\,\frac{e^{h\varepsilon}-1}{e^{\varepsilon}-1} \;\le\; \delta\, e^{h\varepsilon}.$$

Notice that in our setting, $\operatorname{dist}(D^{(u)}, D^{(v)}) = n$ (when every coordinate is substituted), hence for $S = A_v$,

$$\mathbb{P}\big[\mathcal{M}(D^{(u)}) \in A_v\big] \;\ge\; e^{-n\varepsilon}\Big(\mathbb{P}\big[\mathcal{M}(D^{(v)}) \in A_v\big] - \delta_n\Big) \;\ge\; e^{-n\varepsilon}\cdot\frac{2}{3} \;-\; \delta, \tag{38}$$

where we used $\delta_n \le \delta e^{n\varepsilon}$ so $e^{-n\varepsilon}\delta_n \le \delta$.

For fixed u, by the disjointness of the A_v's, we have:

$$1 \;\ge\; \sum_v \mathbb{P}\big[\mathcal{M}(D^{(u)}) \in A_v\big] \;=\; \underbrace{\mathbb{P}\big[\mathcal{M}(D^{(u)}) \in A_u\big]}_{\ge 2/3 \text{ by (37)}} \;+\; \sum_{v\ne u} \underbrace{\mathbb{P}\big[\mathcal{M}(D^{(u)}) \in A_v\big]}_{\ge e^{-n\varepsilon}\frac{2}{3} - \delta \text{ by (38)}}.$$

which implies:

$$1 \;\ge\; \frac{2}{3} + (2^d - 1)\Big(\tfrac{2}{3}e^{-n\varepsilon} - \delta\Big) \quad\Rightarrow\quad 2^d \;\le\; 1 + \frac{1/3}{\frac{2}{3}e^{-n\varepsilon} - \delta}.$$

If $\delta \le \frac{1}{6}e^{-n\varepsilon}$ then the denominator is at least $\frac{1}{2}e^{-n\varepsilon}$, and so

$$2^d \;\le\; 1 + \frac{1/3}{\frac{1}{2}e^{-n\varepsilon}} \;\le\; 2e^{n\varepsilon}.$$

Taking logrithm on both sides yields $d\log 2 \le n\varepsilon + \log 2$, i.e., $n \ge \frac{\log 2}{\varepsilon}(d-1)$. ■

M Technical Lemmas

M.1 Supporting Results on Sparse Recovery

For completeness, we introduce the results from sparse recovery [Tia24] that is used in Section 4.2 and Appendix I.

Definition 40 (Numerical sparsity) *A vector x is s-numerically sparse if $\frac{\|x\|_1^2}{\|x\|_2^2} \leq s$.*

Numerical sparsity extends the traditional notion of sparsity. By definition, an s-sparse vector is also s-numerically sparse. A notable property of numerical sparsity is that the difference between a sparse vector and a numerically sparse vector remains numerically sparse, as stated in the following lemma.

Lemma 41 (Difference of numerically sparse vectors) *Let $x \in \mathbb{R}^d$ be an s-sparse vector. For any vector $x' \in \mathbb{R}^d$ satisfying $\|x'\|_1 \leq \|x\|_1$, the difference $x' - x$ is $4s$-numerically sparse.*

Proof Let $S := \{i \in [d] \mid x[i] \neq 0\}$ and denote $v := x' - x$. We have

$$\|x'\|_1 = \|x+v_S+v_{S^c}\|_1 = \|x+v_S\|_1+\|v_{S^c}\|_1 \geq \|x\|_1-\|v_S\|_1+\|v_{S^c}\|_1 \geq \|x'\|_1-\|v_S\|_1+\|v_{S^c}\|_1,$$

which implies $\|v_S\|_1 \geq \|v_{S^c}\|_1$. Therefore,

$$\begin{aligned}\|v\|_1 &= \|v_S\|_1 + \|v_{S^c}\|_1 \\ &\leq 2\|v_S\|_1 \\ &\leq 2\sqrt{s}\|v_S\|_2 \\ &\leq 2\sqrt{s}\|v\|_2\end{aligned}$$

which implies $\|v\|_1^2 \leq 4s\|v\|_2^2$. Thus, by Definition 40, v satisfies $4s$-numerically sparse. ■

If the vector x is s-sparse, we can reduce its dimension while preserving the ℓ_2 norm using matrices that satisfy the Restricted Isometry Property.

Definition 42 ((e, s)-Restricted isometry property (RIP)) *A matrix $\Phi \in \mathbb{R}^{k\times d}$ satisfies the (e, s)-Restricted Isometry Property (RIP) if, for any s-sparse vector $x \in \mathbb{R}^d$ and some $e \in (0, 1)$, the following holds:*

$$(1 - e)\|x\|_2^2 \leq \|\Phi x\|_2^2 \leq (1 + e)\|x\|_2^2.$$

For numerically sparse vectors, we can reduce the dimension while preserving utility by matrices satisfying a related condition – the Restricted well-conditioned (RWC).

Definition 43 ((e, s)-Restricted well-conditioned (RWC) ([Tia24], Definition 4)) *A matrix $\Phi \in \mathbb{R}^{k\times d}$ is (e, s)-Restricted well-conditioned (RWC) if, for any s-numerically sparse vector $x \in \mathbb{R}^d$ and some $e \in (0, 1)$, we have*

$$(1 - e)\|x\|_2^2 \leq \|\Phi x\|_2^2 \leq (1 + e)\|x\|_2^2.$$

Lemma 44 ([Tia24], Lemma 2; [CT05], Theorem 1.4) *Let $\Phi \in \mathbb{R}^{k\times d}$ whose entries are independent and identically distributed Gaussian with mean zero and variance $\mathcal{N}(0, \frac{1}{k})$. For $e, \zeta \in (0, 1)$, if*

$$k \geq C \cdot \frac{s \log\left(\frac{d}{s}\right) + \log\left(\frac{1}{\zeta}\right)}{e^2},$$

for an appropriate constant C, then Φ satisfies (e, s)-RIP with probability $\geq 1 - \zeta$.

There is a connection between RIP and RWC matrices:

Lemma 45 ([Tia24], Lemma 5) *For $\Phi \in \mathbb{R}^{m\times n}$ and $e \in (0, 1)$, if Φ is $(\frac{e}{5}, \frac{25s}{e^2})$-RIP, then Φ is also (e, s)-RWC.*

Finally, we provide the following guarantee for Algorithm 8.

Lemma 46 (ℓ_1 error guarantee from sparse recovery) *Let $\Phi \in \mathbb{R}^{m\times n}$, and $\theta_* \in \mathbb{R}^n$. Given noisy observation $b = \Phi\theta_* + \tilde{w}$ with bounded ℓ_1 norm of noise, i.e. $\|\tilde{w}\|_1 \leq \xi$, consider the following noisy sparse recovery problem:*

$$\begin{aligned}&\hat{\theta} = \arg\min_{\theta} \|\theta\|_1 \\ &\text{s.t. } \|\Phi\theta - b\|_1 \leq \xi\end{aligned}$$

where $\xi > 0$ is the constraint of the noise magnitude. Suppose that θ is an s-sparse vector in $\mathbb{R}^n$ and that Φ is a $(4s, e)$-RWC matrix. Then, the following ℓ_1 estimation error bound holds:

$$\|\hat{\theta} - \theta_*\|_1 \leq \frac{4\sqrt{s}}{\sqrt{1-e}} \cdot \xi$$

Moreover, the problem can be solved in $\mathcal{O}((3m + 4n + 1)^{1.5}(2n + m)\texttt{prec})$ arithmetic operations in the worst case, with each operation being performed to a precision of $\mathcal{O}(\texttt{prec})$ bits.

Proof For utility guarantee: By $\|\tilde{w}\|_1 \leq \xi$, θ_* is a feasible solution. Thus, we have $\|\hat{\theta}\|_1 \leq \|\theta_*\|_1$, which implies $h := \hat{\theta} - \theta_*$ is $4s$-numerically sparse by Lemma 41. Since Φ is $(e, 4s)$-RWC, we have:

$$(1-e)\|h\|_2^2 \leq \|\Phi h\|_2^2 \leq (1+e)\|h\|_2^2 \tag{39}$$

Now it remains to bound $\|h\|_1$:

$$\begin{aligned}
\|h\|_1 &\leq \sqrt{4s}\|h\|_2 \\
&\leq \sqrt{4s} \cdot \frac{\|\Phi h\|_2}{\sqrt{1-e}} \\
&\leq \frac{2\sqrt{s}}{\sqrt{1-e}}(\|\Phi\hat{\theta} - b\|_1 + \|\Phi\theta_* - b\|_1) \\
&\leq \frac{4\sqrt{s}}{\sqrt{1-e}} \cdot \xi
\end{aligned} \tag{40}$$

where the last inequality is by feasibility of $\hat{\theta}$ and the structure of b.

Now we prove the runtime guarantee. We first reformulate this problem to Linear Programming:

$$(P) \qquad \min_{\theta, u^+, u^-, v} \sum_{i=1}^{n} (u_i^+ + u_i^-)$$

subject to:

$$\begin{aligned}
&\theta_i = u_i^+ - u_i^-, \quad u_i^+, u_i^- \geq 0, \quad \forall i = 1, \ldots, n, \\
&\Phi_j\theta - b_j \leq v_j, \quad \forall j = 1, \ldots, m, \\
&-(\Phi_j\theta - b_j) \leq v_j, \quad \forall j = 1, \ldots, m, \\
&\sum_{j=1}^{m} v_j \leq \xi, \\
&v_j \geq 0, \quad \forall j = 1, \ldots, m.
\end{aligned}$$

The problem (P) has $2n + m$ variables and $2m + 2n + 1$ constraints. By [Vai89], this can be solved in $\mathcal{O}((3m + 4n + 1)^{1.5}(2n + m)B))$ arithmetic operations in the worst case, with each operation being performed to a precision of $\mathcal{O}(B)$ bits. ■

M.2 A Concentration Inequality for Laplace Random Variables

Definition 47 (Laplace Distribution) *$X \sim Lap(b)$ if its probability density function satisfies $f_X(t) = \frac{1}{2b}\exp\left(-\frac{|x|}{b}\right)$.*

Lemma 48 (Concentration of the ℓ_1 norm of Laplace vector) *Let $X = (x_1, \ldots, x_k)$ with each x_i independently identically distributed as $Lap(b)$. Then, with probability at least $1 - \zeta$,*

$$\|X\|_1 \leq 2kb + 2b\log(1/\zeta).$$

Proof $\|X\|_1 = \sum_{i=1}^{k} |x_i|$ follows the Gamma distribution $\Gamma(k, b)$, with probability density function $f(x) = \frac{1}{\Gamma(k)b^k}x^{k-1}e^{-\frac{x}{b}}$. Applying the Chernoff's tail bound of Gamma distribution $\Gamma(k, b)$, we have

$$\mathbb{P}(\|X\|_1 \geq t) \leq \left(\frac{t}{kb}\right)^k e^{k-\frac{t}{b}}, \text{ for } t > kb.$$

Taking $t = 2kb + 2b\log(1/\zeta)$, we have

$$\begin{aligned}
\mathbb{P}(\|X\|_1 \geq 2kb + 2b\log(1/\zeta)) &\leq \left(\frac{2kb + 2b\log(1/\zeta)}{kb}\right)^k e^{k - \frac{2kb+2b\log(1/\zeta)}{b}} \\
&\leq 2^k \left(1 + \frac{\log(1/\zeta)}{k}\right)^k e^{-k}\zeta^2 \\
&\leq \frac{1}{\zeta}\zeta^2 \left(\frac{2}{e}\right)^k \\
&\leq \zeta.
\end{aligned}$$

■

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: All claims are well supported.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: We provided limitations.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: We provide complete assumptions and proofs.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [NA]

Justification: NA.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [NA]

Justification: NA.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).

- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification: NA.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: NA.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.

- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: NA.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: NA.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: NA.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: NA.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: NA.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: NA.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: NA.

    Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: Only for editing purposes.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.