

569	Contents	
570	1 Introduction	1
571	1.1 Summary of the Results	2
572	1.2 Related Work	3
573	2 Algorithm: Converting Approximate DP to Pure DP	3
574	3 Technical Lemma: from TV distance to ∞-Wasserstein distance	5
575	4 Empirical Risk Minimization with Pure Differential Privacy	6
576	4.1 Purified DP Stochastic Gradient Descent	6
577	4.2 Purified DP Frank-Wolfe Algorithm	6
578	5 Pure DP Data-dependent Mechanisms	7
579	5.1 Propose-Test-Release with Pure Differential Privacy	7
580	5.2 Pure DP Mode Release	7
581	5.3 Pure DP Linear Regression	8
582	6 Pure DP Query Release	8
583	7 Purification as a Tool for Proving Lower Bounds	9
584	8 Conclusion and Limitation	9
585	A Preliminaries	17
586	A.1 Definitions on Distributional Discrepancy	17
587	A.2 Lemmas for DP Analysis of Algorithm 1	18
588	B Supplementary Discussion on Randomized Post-Processing	18
589	C Discussion: Purification on Finite Output Spaces	19
590	D Privacy Analysis: Proof Sketch of Theorem 1	19
591	E Deferred Proofs in Section 2 and Appendix A	20
592	E.1 Proof of Lemma 6	20
593	E.2 Proof of Lemma 14	22
594	E.3 Proof of Lemma 15	22
595	E.4 Proof of Theorem 1 and Corollary 4	23
596	E.5 Proof of Theorem 2	23
597	F Details for DP-SGD	24
598	F.1 Algorithms and Notations	24
599	F.2 Noisy Gradient Descent Using Laplace Mechanism	24
600	F.3 Analysis of DP-SGD	25
601	F.3.1 Privacy Accounting Results	25
602	F.3.2 Convex and Lipschitz case	26
603	F.3.3 Strongly Convex and Lipschitz case	27
604	F.4 Analysis of Purified DP-SGD	28
605	F.4.1 Proof of Theorem 3	28
606	G Details for DP-Frank-Wolfe	29
607	G.1 Approximate DP Frank-Wolfe Algorithm	29
608	G.2 Pure DP Frank-Wolfe Algorithm	30
609	G.3 Proof of Theorem 4	30
610	G.4 Proof of Lemma 28	32
611	H Details for Data-dependent DP mechanism Design	33
612	H.1 Pure DP Propose Test Release	33

613	H.2	Privately Bounding Local Sensitivity	34
614	H.3	Private Mode Release	35
615	H.4	Private Linear Regression Through Adaptive Sufficient Statistics Perturbation . . .	35
616	I	Details for Private Query Release	38
617	I.1	Problem Setting	38
618	I.2	Private Multiplicative Weight Exponential Mechanism	38
619	J	Details for the Lower Bound	40
620	K	Technical Lemmas	42
621	K.1	Supporting Results on Sparse Recovery	42
622	K.2	A Concentration Inequality for Laplace Random Variables	43
623	L	Extended Lower Bounds	44
624	L.1	One-Way Marginal Release	44
625	L.2	Private Selection	46

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 $[n] = \{1, \dots, n\}$. 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 , $\|A\|_q := \sup_{x \in A} \|x\|_q$ represents the ℓ_q radius of set A and $\text{Diam}_q(A) := \sum_{x, y \in A} \|x - y\|_q$ represent the ℓ_1 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 9 ([69], Theorem 6; [27], Definition 3.6) *The Rényi divergence of order ∞ (also known as Max Divergence) between two probability measures μ and ν on a measurable space (Θ, \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 10 *The total variation (TV) distance between two probability measures μ and ν on a measurable space (Θ, \mathcal{F}) is defined as:*

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

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

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

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 $\text{ess sup}_{(x, y) \sim \gamma}$ denotes the essential supremum with respect to measure γ . By [36, Proposition 1], the infimum in this definition is attainable, i.e., there exists $\gamma^* \in \Gamma_c(\mu, \nu)$ such that $W_{\infty}^{\ell_q}(\mu, \nu) = \text{ess sup}_{(x, y) \sim \gamma^*} \|x - y\|_q$.

For Laplace perturbation (Lemma 14), 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 $W_{\infty}^{\ell_1}(\cdot, \cdot) \leq d^{1-\frac{1}{q}} W_{\infty}^{\ell_q}(\cdot, \cdot)$.

Beyond this coupling-based characterization, the following equivalent definition of the ∞ -Wasserstein distance offers a geometric interpretation.

Lemma 12 ([36], 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 $U^{\alpha} := \{x \in \Theta : \|x - U\|_q \leq \alpha\}$.

664 A.2 Lemmas for DP Analysis of Algorithm 1

665 We now introduce the three lemmas for the privacy analysis: the equivalence definition of (ε, δ) -DP
 666 (Lemma 13), the Laplace perturbation (Lemma 14), and the weak triangle inequality for ∞ -Rényi
 667 divergence (Lemma 15).

668 **Lemma 13 (Lemma 3.17 of [24])** *A randomized mechanism \mathcal{M} satisfies (ε, δ) -DP if and only if for*
 669 *all neighboring datasets $D \simeq D'$, there exist probability measures P, P' such that $d_{\text{TV}}(\mathcal{M}(D), P) \leq$*
 670 *$\frac{\delta}{e^\varepsilon + 1}$, $d_{\text{TV}}(\mathcal{M}(D'), P') \leq \frac{\delta}{e^\varepsilon + 1}$, $D_\infty(P \parallel P') \leq \varepsilon$, and $D_\infty(P' \parallel P) \leq \varepsilon$.*

671 **Lemma 14 (Laplace perturbation, adapted from Theorem 3.2 of [61])** *Let μ and ν be probabil-*
 672 *ity distributions on \mathbb{R}^d . Let $\text{Lap}^{\otimes d}(b)$ denote the distribution of $\mathbf{z} \in \mathbb{R}^d$, where $z_i \stackrel{\text{i.i.d.}}{\sim} \text{Lap}(b)$. If*
 673 *$W_\infty^{\ell_1}(\mu, \nu) \leq \Delta$, then*

$$D_\infty\left(\mu * \text{Lap}^{\otimes d}\left(\frac{\Delta}{\varepsilon}\right) \parallel \nu * \text{Lap}^{\otimes d}\left(\frac{\Delta}{\varepsilon}\right)\right) \leq \varepsilon, \text{ and } D_\infty\left(\nu * \text{Lap}^{\otimes d}\left(\frac{\Delta}{\varepsilon}\right) \parallel \mu * \text{Lap}^{\otimes d}\left(\frac{\Delta}{\varepsilon}\right)\right) \leq \varepsilon.$$

674 The proof is provided in Appendix E for completeness. The Laplace mechanism is a special case of
 675 Lemma 14 by setting μ and ν to Dirac distributions at $f(D)$ and $f(D')$, respectively. Lemma 14 can
 676 also be derived from the limit case of [29, Lemma 20] as the Rényi order approaches infinity.

677 **Lemma 15 (Weak Triangle Inequality, adapted from [49], Lemma 4.1)** *Let μ, ν, π be probabil-*
 678 *ity measures on a measurable space (Θ, \mathcal{F}) . If $D_\infty(\mu \parallel \pi) < \infty$ and $D_\infty(\pi \parallel \nu) < \infty$, then*
 679 *$D_\infty(\mu \parallel \nu) \leq D_\infty(\mu \parallel \pi) + D_\infty(\pi \parallel \nu)$.*

680 The proof is deferred to Appendix E. This lemma generalizes [49, Lemma 4.1], which focuses on
 681 discrete distributions. Additionally, Lemma 15 corresponds to the infinite Rényi order limit of [56,
 682 Proposition 11] and [54, lemma 12].

683 B Supplementary Discussion on Randomized Post-Processing

684 In this section, we clarify the distinction between our "purification" process and the randomized
 685 post-processing in [8, Theorem 10]. Define the following function

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

686 By [75], a mechanism \mathcal{M} is (ε, δ) -DP if and only if \mathcal{M} is f_ε -DP.

687 [8, Theorem 10] [20, Theorem 2.10] establish the existence of a (randomized) post-processing
 688 method that transforms a pair of distributions into another pair with a dominating trade-off function.
 689 Specifically, they show that if $T(P, Q) \leq T(P', Q')$, then there exists a randomized algorithm
 690 Proc such that $\text{Proc}(P) = P'$ and $\text{Proc}(Q) = Q'$, where T denotes the trade-off function [20,
 691 Definition 2.1]. Their proof constructs a sequence of transformations and takes the limit. In contrast,
 692 our "purification" process provides a computationally efficient post-processing method for a different
 693 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
 694 neighboring datasets $D \simeq D'$, we seek a randomized post-processing procedure $\mathcal{A}_{\text{pure}}$ such that
 695 $T(\mathcal{A}_{\text{pure}} \circ \mathcal{M}(D), \mathcal{A}_{\text{pure}} \circ \mathcal{M}(D')) \geq f_{(\varepsilon + \varepsilon'), 0}$ while maintaining utility guarantee.

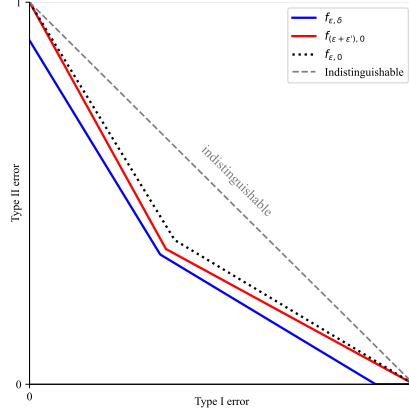


Figure 3: Trade-off functions for (ε, δ) -DP, $(\varepsilon, 0)$ -DP, and $(\varepsilon + \varepsilon', 0)$ -DP. Our method provides a solution to post-process the (ε, δ) -DP distribution pair (in blue) to the $(\varepsilon + \varepsilon', 0)$ -DP pair (in red).

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 $\mathcal{M} : \mathcal{X} \rightarrow \mathcal{Y}$ 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 [41, Lemma 3.2]. However, this strategy does not yield pure DP when the output space is continuous, as the following counterexample shows.

Example 16 Let $f : D \rightarrow [0, 1]$ 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, $\mathcal{M}(D) = \delta \cdot \delta_{f(D)}^{\text{Dirac}} + (1 - \delta)\text{Unif}([0, 1])$, where $\delta_{f(D)}^{\text{Dirac}}$ denotes the Dirac measure at $f(D)$. Then \mathcal{M} satisfies $(0, \delta)$ -DP. Now, consider applying the uniform mixture strategy:

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

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

D 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 13), 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 15), 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?

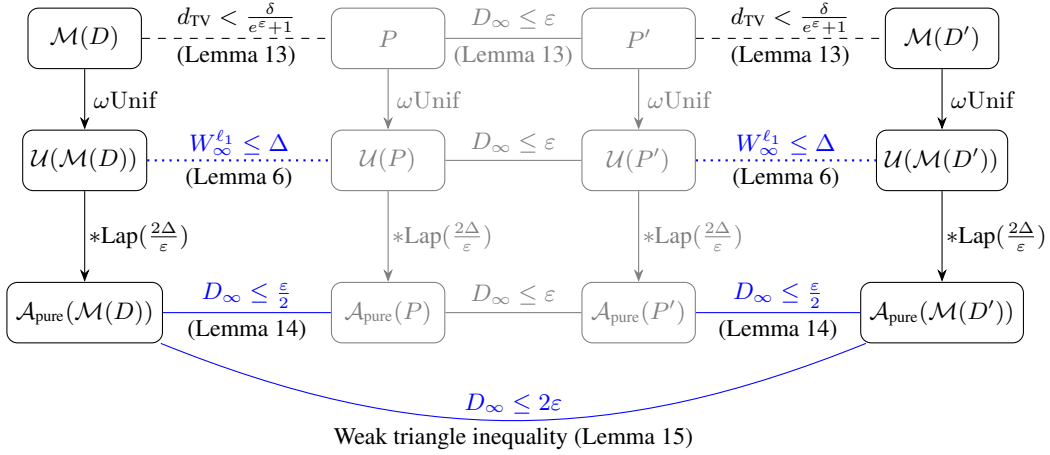


Figure 4: Flowchart illustrating the proof sketch of Theorem 1 and the intuition behind Algorithm 1. The notation $D_\infty \leq \varepsilon$ is an abbreviation for the pair of inequalities $D_\infty(\mu \parallel \nu) \leq \varepsilon$ and $D_\infty(\nu \parallel \mu) \leq \varepsilon$, where μ and 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 $\mathcal{U}(\cdot) = (1 - \omega) \cdot \text{Unif}(\Theta) + \omega \text{Unif}(\cdot)$. The notation $*\text{Lap}$ refers to the convolution with the Laplace distribution, as in Algorithm 1, Line 4.

A natural way to establish the ∞ -Rényi 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 14), 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 $W_\infty^{\ell_1}$ conversion lemma that generalizes [51, 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 E.

E Deferred Proofs in Section 2 and Appendix A

E.1 Proof of Lemma 6

Proof of Lemma 6 Denote $\text{dist}(x, y) = \|x - y\|_q$. 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 $\xi = p_{\min} \cdot \text{Vol}\left(\mathbb{B}_{\ell_q}^d(1)\right) \cdot \left(\frac{r}{4R}\right)^d \cdot \Delta^d$, so that $d_{\text{TV}}(\mu, \nu) < \xi$.

To prove the result that $W_\infty \leq \Delta$, we use the equivalent definition of W_∞ in Lemma 12. By this definition, to prove $W_\infty(\mu, \nu) \leq \Delta$, it suffices to show that

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

where we re-define $A^\Delta := \{x \in \mathbb{R}^d \mid \text{dist}(x, A) \leq \Delta\}$. 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 $\mu(A) \leq \nu(A^\Delta)$. 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

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

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

To prove $\nu(A^\Delta \setminus A) \geq \xi$, 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

753 Property 2 K is contained in Θ , i.e., $K \subseteq \Theta$. This guarantees that $\nu(K) = \nu(K \cap \Theta) \geq p_{\min} \cdot$
 754 $\text{Vol}(K) = \xi$.

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

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

757 Step 2 Then, construct the ball K with center $c_1 = \omega c + (1 - \omega)y$, radius ωr , where $\omega = \frac{\Delta^2}{4R^2}$,
 758 i.e., $K = \mathcal{B}(c_1, \omega r)$. We will later show that this ball is contained in the convex hull of
 759 $\mathcal{B}(c, r) \cup \{y\}$. This construction is inspired by [53].

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

762 We prove it by contradiction. If instead $\Theta \cap \{x \in \mathbb{R}^d | \text{dist}(x, A) = \Delta/2\} = \emptyset$, then for all $x \in \Theta$,
 763 $\text{dist}(x, A) \neq \Delta/2$. Due to the continuity of dist and the convexity of Θ , we know that only one of
 764 these two statements holds:

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

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

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

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

771 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

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

772 • $(x \notin A)$. If $x \in A$, since A is an open set and $\text{dist}(y, A) = \Delta/2$, we have that $\text{dist}(x, y) >$
 773 $\Delta/2$, which contradicts to (3). Therefore $x \notin A$.

774 • $(x \in A^\Delta)$. We have $\text{dist}(x, A) \leq \text{dist}(x, y) + \text{dist}(y, A) \stackrel{(3)}{\leq} \Delta$, implying that $x \in A^\Delta$.

775 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
 776 $\|\mathbf{v}\|_q \leq 1$. We have

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

777 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 $x \in \Theta$,
 778 which completes the proof.

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

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

782

²We note that the symbol ω in Section E.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.

783 E.2 Proof of Lemma 14

784 **Proof of Lemma 14** denote $P = \mu * \text{Lap}^{\otimes d}(\frac{\Delta}{\varepsilon})$, $Q = \nu * \text{Lap}^{\otimes d}(\frac{\Delta}{\varepsilon})$. Then, P and Q are
 785 absolutely continuous with respect to the Lebesgue measure. This is because for any Lebesgue zero
 786 measure set S , $P(S) = \int_x \mathbb{P}_{\text{Lap}}(S - x) d\mu(x)$, where $S - x = \{y | x + y \in S\}$, and \mathbb{P}_{Lap} denotes
 787 the probability measure of $\text{Lap}^{\otimes d}(\frac{\Delta}{\varepsilon})$. Since $S - x$ is zero measure for all $x \in \mathbb{R}^d$, and \mathbb{P}_{Lap} is
 788 absolutely continuous w.r.t. the Lebesgue measure, we have $\mathbb{P}_{\text{Lap}}(S - x) = 0$ for all $x \in \mathbb{R}^d$. Therefore,
 789 $P(S) = 0$. Thus P is absolutely continuous w.r.t. the Lebesgue measure, and Q similarly follows.

790 Denote p and q the probability density function of P and Q respectively. Since $W_{\infty}^{\ell_1}(\mu, \nu) \leq$
 791 Δ , by [36, Proposition 1], there exists $\gamma^* \in \Gamma_c(\mu, \nu)$, such that $\mathbb{P}_{(u,v) \sim \gamma^*}[\|u - v\|_1 > \Delta] =$
 792 $\gamma^* \{\|u - v\|_1 > \Delta\} = 0$. By the definition of the max divergence (Theorem 6 of [69]), we have

$$\begin{aligned}
 e^{D_{\infty}(P\|Q)} &= \text{ess sup}_{x \sim P} \frac{p(x)}{q(x)} && ([69, \text{Theorem 6}]) \\
 &\leq \text{ess sup}_{x \sim P} \frac{\int_{\mathbb{R}^d} e^{-\frac{\|x-u\|_1}{b}} d\mu(u)}{\int_{\mathbb{R}^d} e^{-\frac{\|x-v\|_1}{b}} d\nu(v)} && (\text{Convolution theorem, PDF of Laplace}) \\
 &\leq \text{ess sup}_{x \sim P} \frac{\int_{\mathbb{R}^d \times \mathbb{R}^d} e^{-\frac{\|x-u\|_1}{b}} d\gamma^*(u, v)}{\int_{\mathbb{R}^d \times \mathbb{R}^d} e^{-\frac{\|x-v\|_1}{b}} d\gamma^*(u, v)} && (\gamma^* \in \Gamma_c(\mu, \nu)) \\
 &\leq \text{ess sup}_{x \sim P} \frac{\int_{\mathbb{R}^d \times \mathbb{R}^d} e^{-\frac{\|x-v\|_1}{b} + \frac{\|u-v\|_1}{b}} d\gamma^*(u, v)}{\int_{\mathbb{R}^d \times \mathbb{R}^d} e^{-\frac{\|x-v\|_1}{b}} d\gamma^*(u, v)} && (\text{triangle's inequality}) \\
 &\leq \text{ess sup}_{x \sim P} \frac{\int_{\|u-v\|_1 \leq \Delta} e^{-\frac{\|x-v\|_1}{b} + \frac{\|u-v\|_1}{b}} d\gamma^*(u, v) + \int_{\|u-v\|_1 > \Delta} e^{-\frac{\|x-v\|_1}{b} + \frac{\|u-v\|_1}{b}} d\gamma^*(u, v)}{\int_{\mathbb{R}^d \times \mathbb{R}^d} e^{-\frac{\|x-v\|_1}{b}} d\gamma^*(u, v)} \\
 &\leq \text{ess sup}_{x \sim P} \frac{\int_{\|u-v\|_1 \leq \Delta} e^{-\frac{\|x-v\|_1}{b} + \frac{\|u-v\|_1}{b}} d\gamma^*(u, v)}{\int_{\mathbb{R}^d \times \mathbb{R}^d} e^{-\frac{\|x-v\|_1}{b}} d\gamma^*(u, v)} && (\gamma^* \{\|u - v\|_1 > \Delta\} = 0) \\
 &\leq \text{ess sup}_{x \sim P} \frac{e^{\frac{\Delta}{b}} \int_{\|u-v\|_1 \leq \Delta} e^{-\frac{\|x-v\|_1}{b}} d\gamma^*(u, v)}{\int_{\mathbb{R}^d \times \mathbb{R}^d} e^{-\frac{\|x-v\|_1}{b}} d\gamma^*(u, v)} \\
 &\leq \text{ess sup}_{x \sim P} \frac{e^{\frac{\Delta}{b}} \int_{\mathbb{R}^d \times \mathbb{R}^d} e^{-\frac{\|x-v\|_1}{b}} d\gamma^*(u, v)}{\int_{\mathbb{R}^d \times \mathbb{R}^d} e^{-\frac{\|x-v\|_1}{b}} d\gamma^*(u, v)} \\
 &= e^{\frac{\Delta}{b}}
 \end{aligned}$$

793

794 E.3 Proof of Lemma 15

795 **Proof of Lemma 15** Since $D_{\infty}(\mu\|\pi) < \infty$, for any measurable set $S \in \mathcal{F}$, such that $\mu(S) > 0$, we
 796 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 9,

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

797

798 E.4 Proof of Theorem 1 and Corollary 4

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

801 Let D and D' be neighboring datasets, and let \mathcal{M} be an (ε, δ) -DP mechanism as in Section D. By
802 the equivalence definition of approximate DP (Lemma 13), there exists a hypothetical distribution
803 pair P, P' such that

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

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

806 Denote $\mathcal{U}(\cdot) = (1 - \omega) \cdot + \omega \text{Unif}(\Theta)$. After the uniform mixture step (Line 3), the distributions
807 $\mathcal{U}(\mathcal{M}(D)), \mathcal{U}(\mathcal{M}(D')), \mathcal{U}(P), \mathcal{U}(P')$ all satisfy the assumption in Lemma 6 with

$$p_{\min} \geq \frac{\omega}{\text{Vol}(\Theta)} \geq \frac{\omega}{(R/2)^d \text{Vol}(\mathbb{B}_{\ell_q}^d(1))}.$$

Therefore, by Lemma 6 and the fact that $\|\cdot\|_1 \leq d^{1-\frac{1}{q}} \|\cdot\|_q$, we have

$$W_\infty^{\ell_1}(\mathcal{U}(\mathcal{M}(D)), \mathcal{U}(P)) \leq \Delta, \text{ and } W_\infty^{\ell_1}(\mathcal{U}(\mathcal{M}(D')), \mathcal{U}(P')) \leq \Delta,$$

where Δ is defined in Line 2 in Algorithm 1. Therefore, by Lemma 14, we have

$$\begin{aligned} D_\infty\left(\mathcal{U}(\mathcal{M}(D)) * \text{Lap}^{\otimes d}\left(\frac{2\Delta}{\varepsilon'}\right) \| \mathcal{U}(P) * \text{Lap}^{\otimes d}\left(\frac{2\Delta}{\varepsilon'}\right)\right) &\leq \varepsilon'/2, \\ D_\infty\left(\mathcal{U}(P) * \text{Lap}^{\otimes d}\left(\frac{2\Delta}{\varepsilon'}\right) \| \mathcal{U}(\mathcal{M}(D)) * \text{Lap}^{\otimes d}\left(\frac{2\Delta}{\varepsilon'}\right)\right) &\leq \varepsilon'/2, \\ D_\infty\left(\mathcal{U}(\mathcal{M}(D')) * \text{Lap}^{\otimes d}\left(\frac{2\Delta}{\varepsilon'}\right) \| \mathcal{U}(P') * \text{Lap}^{\otimes d}\left(\frac{2\Delta}{\varepsilon'}\right)\right) &\leq \varepsilon'/2, \\ D_\infty\left(\mathcal{U}(P') * \text{Lap}^{\otimes d}\left(\frac{2\Delta}{\varepsilon'}\right) \| \mathcal{U}(\mathcal{M}(D')) * \text{Lap}^{\otimes d}\left(\frac{2\Delta}{\varepsilon'}\right)\right) &\leq \varepsilon'/2. \end{aligned}$$

808 Finally, with the weak triangle inequality (Lemma 15), we conclude that the output of Algorithm 1
809 satisfies $(\varepsilon + \varepsilon')$ -DP.

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

$$\begin{aligned} \mathbb{E} \|x_{\text{pure}} - x_{\text{apx}}\|_q &= \mathbb{P}(u > \omega) \mathbb{E} \left[\|x_{\text{pure}} - x_{\text{apx}}\|_q \mid u > \omega \right] + \mathbb{P}(u \leq \omega) \mathbb{E} \left[\|x_{\text{pure}} - x_{\text{apx}}\|_q \mid u \leq \omega \right] \\ &\leq \mathbb{E} \left[\left(\sum_{i=1}^d |z_i|^q \right)^{\frac{1}{q}} \right] + \omega R \\ &\leq \left(\mathbb{E} \left[\sum_{i=1}^d |z_i|^q \right] \right)^{\frac{1}{q}} + \omega R \quad (\text{Jensen's inequality}) \\ &= (d\Gamma(q+1))^{\frac{1}{q}} \frac{2\Delta}{\varepsilon'} + \omega R \\ &\leq \frac{4dqR^2}{r\varepsilon'} \left(\frac{\delta}{2\omega} \right)^{\frac{1}{d}} + \omega R \end{aligned}$$

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

813 E.5 Proof of Theorem 2

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

816 We consider the ℓ_1 norm, i.e., $q = 1$. Let $a = (\frac{1}{2}, \dots, \frac{1}{2})$. For unit cube $[0, 1]^d$, we have $a +$
817 $\mathbb{B}_{\ell_q}^d(\frac{1}{2}) \subseteq [0, 1]^d \subseteq a + \mathbb{B}_{\ell_q}^d(\frac{d}{2})$. That is, $[0, 1]^d$ satisfies Assumption 1 with $R = \frac{d}{2}$, $r = \frac{1}{2}$, and

818 $q = 1$. Therefore, by Theorem 1, z_{pure} satisfies $(2\varepsilon, 0)$ -DP. After the post-processing, the output of
 819 Algorithm 2 maintains $(2\varepsilon, 0)$ -DP.

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

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

$$\mathbb{P}(y_i^{\text{Lap}} \geq t) = \frac{1}{2}e^{-\frac{t}{2\Delta/\varepsilon}} \leq \frac{1}{2}e^{-2dt}.$$

822 Thus,

$$\mathbb{P}\left(y_i^{\text{Lap}} > 0.5, \forall i = 1, \dots, d\right) = \left(1 - \frac{1}{2}e^{-d}\right)^d \geq 1 - \frac{d}{2}e^{-d}, \quad (5)$$

823 where the last inequality is by Bernoulli's inequality.

824 Since the rounding function is defined as $\text{Round}_{\{0,1\}}^d(\mathbf{x}) = (\mathbf{1}(x_i \geq 0.5))_{i=1}^d$, we observe that
 825 $z_{\text{bin}} = z_{\text{pure}}$ if and only if the following conditions hold simultaneously:

826 1. $x \sim \text{Unif}(\Theta)$ is sampled in Line 3 of Algorithm 1.

827 2. For all $i \in [d]$, if $z_{\text{bin}}^{(i)} = 0$, then $y_i^{\text{Lap}} < 0.5$.

828 3. For all $i \in [d]$, if $z_{\text{bin}}^{(i)} = 1$, then $y_i^{\text{Lap}} > -0.5$.

829 By the symmetry of Laplace noise, applying Eq. (5), and using the union bound, we obtain

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

830 ■

831 **F Details for DP-SGD**

832 **F.1 Algorithms and Notations**

833 Let $f(\theta; x)$ represent the individual loss function. $\mathcal{L}(\theta) := \frac{1}{n} \sum_{i=1}^n f(\theta; x_i)$ and $\mathcal{L}^* = \min_{\theta \in \mathcal{C}} \mathcal{L}(\theta)$.

834 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

835 convex are all w.r.t. individual the loss function f . The parameter space Θ in Algorithm 4 is selected
 836 as the ℓ_2 -ball with diameter C that contains \mathcal{C} .

Algorithm 3 Differentially Private SGD [1]

- 1 **Input:** Dataset $D = \{x_1, \dots, x_n\}$, loss function $f : \mathcal{C} \times D \rightarrow \mathbb{R}$, parameters: learning rate η_t , noise scale σ , subsampling rate γ , Lipschitz constant L , parameter space $\mathcal{C} \in \mathbb{R}^d$
 - 2 Initialize $\theta_0 \in \mathcal{C}$ randomly
 - 3 **for** $t \in [T]$ **do**
 - 4 Sample a subset S_t by selecting a γ fraction of the dataset without replacement
 - 5 **Compute gradient:** **for** each $i \in S_t$ **do**
 - 6 $g_t(x_i) \leftarrow \nabla_{\theta} f(\theta_t; x_i)$
 - 7 **end**
 - 7 **Aggregate and add noise:** $\hat{g}_t \leftarrow \frac{1}{\gamma} (\sum_{i \in S_t} g_t(x_i) + \sigma \mathcal{N}(0, \mathbf{I}_d))$
 - 8 **Descent:** $\theta_{t+1} \leftarrow \text{Proj}_{\mathcal{C}}(\theta_t - \eta_t \hat{g}_t)$
 - 9 **end**
 - 9 **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.
-

837 **F.2 Noisy Gradient Descent Using Laplace Mechanism**

838 **Lemma 17 (Laplace Noisy Gradient Descent)** *Let the loss function $f : \mathcal{X} \rightarrow \mathbb{R}^d$ be convex and*
 839 *L -Lipschitz with respect to $\|\cdot\|_2$ and $\max_{X \sim X'} \|\nabla f(X) - \nabla f(X')\|_1 \leq \Delta_1$. Suppose the parameter*
 840 *space $\mathcal{C} \subset \mathbb{R}^d$ is convex with an ℓ_2 diameter of at most C . Running **full-batch** noisy gradient descent*

Algorithm 4 Pure DP SGD

- 1 **Input:** Output from DP-SGD Algorithm 3 θ_{apx} , parameter space Θ , privacy parameter ε' and δ , mixture level ω
2 $\theta_{\text{pure}} \leftarrow \mathcal{A}_{\text{pure}}(\theta_{apx}, \Theta, \varepsilon', \delta, \omega)$ ▷ Algorithm 1
3 **Output:** θ_{pure}
-

841 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
842 parameter $\sigma = \frac{\Delta_1 T}{\varepsilon}$, satisfies ε -DP. Moreover,

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

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

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

845 **Proof** Suppose we run T iterations and the final privacy budget is ε . Then, the privacy budget per
846 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
847 [35, Theorem 9.6], we have

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

848 where the second inequality is obtained by choosing learning rate $\eta = \frac{C}{\sqrt{T(n^2L^2 + 2d\sigma^2)}}$ and the fact
849 $\sqrt{a+b} < \sqrt{a} + \sqrt{b}$ for any positive a and b . The last inequality is by setting $T = \frac{\varepsilon n L}{\Delta_1 \sqrt{d}}$. Divide
850 both sides by n , we have:

$$\mathbb{E} (\mathcal{L}(\bar{\theta}) - \mathcal{L}^*) \leq \mathcal{O} \left(\frac{C \Delta_1^{1/2} L^{1/2} d^{1/4}}{n^{1/2} \varepsilon^{1/2}} \right) \quad (6)$$

851 Without additional information, the best upper bound for Δ_1 is $\sqrt{d}\Delta_2 = \sqrt{d}L$. Plugging this bound
852 to Eq. (6) yields:

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

853 ■

854 F.3 Analysis of DP-SGD

855 F.3.1 Privacy Accounting Results

856 Our privacy accounting for DP-SGD is based on Rényi differential privacy [56]. Before stating the
857 privacy accounting result (Corollary 22), we define Rényi Differential privacy and its variant, zero
858 concentrated Differential Privacy [12].

859 **Definition 18 (Rényi differential privacy)** A randomized mechanism \mathcal{M} satisfies $(\alpha, \varepsilon(\alpha))$ -Rényi
860 Differential Privacy (RDP) if for all neighboring datasets D, D' and for all $\alpha \geq 1$,

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

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 [56, 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 19 (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 20 (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 [56, 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 21) then show the privacy guarantee for multi-step DP-SGD through RDP composition (Corollary 22).

Lemma 21 (RDP guarantee for subsampled Gaussian Mechanism, Theorem 11 in [10]) Let \mathcal{M} be a ρ -zCDP Gaussian mechanism, and \mathcal{S}_γ 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:

$$\varepsilon(\alpha) \geq 13\gamma^2\rho\alpha, \quad \text{for any } \alpha \leq \frac{\log(1/\gamma)}{4\rho} \quad (7)$$

Corollary 22 (RDP guarantee for DP-SGD) Let \mathcal{M} be a Gaussian mechanism satisfying ρ_0 -zCDP and \mathcal{S}_γ be a subsampling procedure on the dataset with subsampling rate γ , then T -fold (adaptive) composition of subsampled Gaussian mechanism, $\mathcal{M}_T := \underbrace{(\mathcal{M} \circ \mathcal{S}_\gamma) \circ \dots \circ (\mathcal{M} \circ \mathcal{S}_\gamma)}_{T \text{ times}}$, satisfies

$(\alpha, \varepsilon(\alpha))$ -RDP with

$$\varepsilon(\alpha) \geq 13\gamma^2\rho_0T\alpha, \quad \text{for any } \alpha \leq \frac{\log(1/\gamma)}{4\rho_0}. \quad (8)$$

Denote $\rho := 13\gamma^2\rho_0T$ for a shorthand, if further $\rho_0 \leq \frac{\log(1/\gamma)}{4(1+\sqrt{\frac{\log(1/\delta)}{\rho}})}$, 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 21 and the composition theorem for RDP ([56, Proposition 1]), we have \mathcal{M}_T satisfies $(\alpha, 13\gamma^2\rho_0T\alpha)$ -RDP for any $\alpha \in (1, \frac{\log(1/\gamma)}{4\rho_0})$. Denote $\rho := 13\gamma^2\rho_0T$. By Lemma 20, if $1 + \sqrt{\frac{\log(1/\delta)}{\rho}} \leq \frac{\log(1/\gamma)}{4\rho_0}$ (i.e., the optimal $\alpha^* \leq \frac{\log(1/\gamma)}{4\rho_0}$), we have \mathcal{M}_T satisfies $(\rho + 2\sqrt{\rho \log(1/\delta)}, \delta)$ -DP. ■

F.3.2 Convex and Lipschitz case

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

Lemma 23 (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 $\gamma = \frac{2\sqrt{d \log(1/\delta)}}{n\sqrt{\varepsilon}}$,

$\sigma^2 = \frac{416L^2 \log(1/\delta)}{\varepsilon}$, $T = \frac{n^2\varepsilon^2}{d \log(1/\delta)}$, $\eta = \sqrt{\frac{C^2}{T(n^2L^2 + d\sigma^2/\gamma^2 + nL^2/\gamma)}}$ satisfies (ε, δ) -DP for any

$\varepsilon \leq (d \wedge 8) \log(1/\delta)$. Moreover,

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

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

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

898 **Proof** We first state the privacy guarantee. Since each gaussian mechanism satisfies $L^2/2\sigma^2$ -zCDP, by
 899 Corollary 22, the composed mechanism satisfies ρ -zCDP with $\rho = \frac{13L^2\gamma^2T}{2\sigma^2} = \frac{\varepsilon^2}{16\log(1/\delta)}$, and thus
 900 (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:

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

901 By [35, Theorem 9.6], we have:

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

902 where the third inequality is by choosing $\eta = \sqrt{\frac{C^2}{T(n^2L^2 + \frac{d\sigma^2}{\gamma^2} + \frac{nL^2}{\gamma})}}$, and the forth line is by $\rho =$
 903 $13T\gamma^2L^2/2\sigma^2$. By the choice of T and γ , we have $\frac{d}{\rho} \geq \max\{\frac{n^2}{T}, \frac{n}{T\gamma}\}$. This implies:

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

904 Since the target approximate DP privacy budget $\varepsilon = 4\sqrt{\rho\log(1/\delta)}$, we have $\sqrt{\rho} = \frac{\varepsilon}{4\sqrt{\log(1/\delta)}}$.

905 Plugging this into the bound above, we have:

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

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

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

907

908 F.3.3 Strongly Convex and Lipschitz case

909 **Lemma 24 (Convergence of DP-SGD in strongly convex and L -Lipschitz setting)** Assume indi-
 910 vidual loss function f is λ -strongly convex and L -Lipschitz. Running DP-SGD with parameters
 911 $\gamma = \frac{2\sqrt{d\log(1/\delta)}}{n\sqrt{\varepsilon}}$, $\sigma^2 = \frac{416L^2\log(1/\delta)}{\varepsilon}$, $T = \frac{n^2\varepsilon^2}{d\log(1/\delta)}$, $\eta_t = \frac{2}{n\lambda(t+1)}$ satisfies (ε, δ) -DP for any
 912 $\varepsilon \leq (d \wedge 8) \log(1/\delta)$. Moreover,

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

913 and the number of incremental gradient calls is

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

914 **Proof** The derivation of privacy guarantee follows the same procedure as Lemma 23. By Corollary 22,
 915 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
 916 result from [48], where $\Lambda = n\lambda$ be strong convex parameter of F , we have:

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

917 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
 918 $\frac{d}{\rho} \geq \max\{\frac{n^2}{T}, \frac{n}{T\gamma}\}$. This implies:

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

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

920 For the number of incremental gradient calls (denote as \mathcal{G}), we have the same result as in Lemma 23:

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

921

922 F.4 Analysis of Purified DP-SGD

Lemma 25 (Error from Laplace perturbation) Suppose $x \in \mathbb{R}^d$ and $\tilde{x} = x + \text{Lap}^{\otimes d}(b)$, then

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

Proof

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

923

924 F.4.1 Proof of Theorem 3

925 **Theorem 10 (Restatement of Theorem 3)** Let the domain $\mathcal{C} \subset \mathbb{R}^d$ be a convex set with ℓ_2 diam-
 926 eter C , and let $f(\cdot; \mathbf{x})$ be L -Lipschitz for all $\mathbf{x} \in \mathcal{X}$. Algorithm 4 satisfies 2ε -pure DP and with
 927 $\tilde{\mathcal{O}}(n^2\varepsilon^{3/2}d^{-1})$ incremental gradient calls, the output θ_{out} satisfies:

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

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

$$\log(1/\delta) = \mathcal{O}(d\log(2) + d\log(n) + d\log(d) + d\log(C)) = \tilde{\mathcal{O}}(d) \quad (9)$$

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

933 (When f is Convex and L -Lipschitz): By Lemma 23 and dividing both sides by n :

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

934 (When f is λ -strongly Convex and L -Lipschitz): By Lemma 24 and dividing both sides by n :

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

935 The number of incremental gradient calls for both cases:

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

936

937 G Details for DP-Frank-Wolfe

938 G.1 Approximate DP Frank-Wolfe Algorithm

939 Given a dataset $D = x_1, \dots, x_n$ and a parameter space \mathcal{C} , we denote the individual loss function by
940 $f : \mathcal{C} \times \mathcal{X} \rightarrow \mathbb{R}$. We define the average empirical loss as follows:

$$\mathcal{L}(\theta) := \frac{1}{n} \sum_{i=1}^n f(\theta; x_i) \tag{10}$$

For completeness, we state the Frank-Wolfe algorithm [31] as follows.

Algorithm 5 Frank-Wolfe algorithm (Non-Private)

4 **Input:** $\mathcal{C} \subseteq \mathbb{R}^d$, loss function $\mathcal{L} : \mathcal{C} \rightarrow \mathbb{R}$, number of iterations T , stepsizes η_t .
5 Choose an arbitrary θ_1 from \mathcal{C}
6 **for** $t = 1$ **to** $\tilde{T} - 1$ **do**
7 Compute $\tilde{\theta}_t = \arg \min_{\theta \in \mathcal{C}} \langle \nabla \mathcal{L}(\theta_t), \theta - \theta_t \rangle$
8 Set $\theta_{t+1} = \theta_t + \eta_t(\tilde{\theta}_t - \theta_t)$
 end
9 Output θ_T

941

942 The differential private version of Algorithm 5 is modified by using the exponential mechanism to
943 select coordinates in each update. We follow the setting in [63] with initialization at point zero.

944 **Lemma 26 (Equation 21 of [64])** *Running Algorithm 6 for T iterations yields:*

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

945 where $\Gamma_{\mathcal{L}}$ is the curvature parameter [63, Definition 2.1], which can be upper bounded by $\beta \|\mathcal{C}\|_1^2$ if
946 the loss function f is β -smooth [64].

947 The sparsity of θ_T is given in the following lemma.

948 **Lemma 27 (Sparsity of DP Frank-Wolfe)** *Suppose $S \subset \{x \in \mathbb{R}^d \mid \text{nnz}(x) \leq s\}$. After running*
949 *Algorithm 6 for T iterates, the output θ_T is $Ts \wedge d$ -sparse.*

950 **Proof** From Line 14 and 15 of Algorithm 6, we know $\text{nnz}(\theta_{t+1}) \leq \text{nnz}(\theta_t) + s$. Since $\text{nnz}(\theta_0) = 0$,
951 we have $\text{nnz}(\theta_T) \leq Ts$. Since $\theta_T \in \mathbb{R}^d$, we have $\text{nnz}(\theta_T) \leq d$. Therefore, $\text{nnz}(\theta_T) \leq Ts \wedge d$. ■

Algorithm 6 Approximate DP Frank-Wolfe Algorithm [63]

10 **Input:** Dataset $D = \{d_1, \dots, d_n\}$, loss function f defined in Eq. (10) with ℓ_1 -Lipschitz constant L_1 , privacy parameters (ε, δ) , convex set $\mathcal{C} = \text{conv}(S)$, $\|\mathcal{C}\|_1 = \max_{s \in S} \|s\|_1$.
11 Initialize $\theta_0 \leftarrow \mathbf{0} \in \mathcal{C}$.
12 **for** $t = 0$ **to** $T - 1$ **do**
13 $\forall s \in S, \alpha_s \leftarrow \langle s, \nabla \mathcal{L}(\theta_t; \mathcal{D}) \rangle + \text{Lap} \left(\frac{L_1 \|\mathcal{C}\|_1 \sqrt{8T \log(1/\delta)}}{n\varepsilon} \right)$, where $\text{Lap}(\lambda) \sim \frac{1}{2\lambda} e^{-|x|/\lambda}$
14 $\tilde{\theta}_t \leftarrow \arg \min_{s \in S} \alpha_s$
15 $\theta_{t+1} \leftarrow (1 - \eta_t) \theta_t + \eta_t \tilde{\theta}_t$, where $\eta_t = \frac{2}{t+2}$
end
16 **Output** $\theta_{\text{priv}} = \theta_T$

Algorithm 7 Pure DP Frank-Wolfe Algorithm

1 **Input:** Dataset D , loss function \mathcal{L} : defined in Eq. (10), 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 41 satisfying $(1/4, 4T)$ -RWC with high probability.
2 Set parameters $T = \tilde{\Theta} \left(\sqrt{\frac{\beta \|\mathcal{C}\|_1 n \varepsilon}{L_1}} \right)$, $k = \Theta \left(T \log \left(\frac{d}{T} \right) + \log n \right)$, $\omega = \frac{1}{n}$, $\delta = \frac{2\omega}{(nk)^k}$.
3 $\theta_{\text{FW}} \leftarrow \text{Algorithm 6}$ $\triangleright (\varepsilon, \delta)$ -DP FW
4 $\theta_{\text{apx-}k} \leftarrow \Phi \theta_{\text{FW}} \in \mathbb{R}^k$ \triangleright Dimension reduction
5 $\theta_{\text{pure-}k} \leftarrow \mathcal{A}_{\text{pure}} \left(x_{\text{apx}} = \text{Clip}_{2\|\mathcal{C}\|_1}^{\ell_2}(\theta_{\text{apx-}k}), \Theta = \mathcal{B}_{\ell_2}^k(2\|\mathcal{C}\|_1), \varepsilon' = \varepsilon, \delta, \omega \right)$ \triangleright Algorithm 1
6 $\theta_{\text{pure-}d} \leftarrow \mathcal{M}_{\text{rec}}(\theta_{\text{pure-}k}, \Phi, \xi)$ \triangleright Algorithm 8, with ξ defined in Eq. (11)
7 $\theta_{\text{out}} = \text{Clip}_{\|\mathcal{C}\|_1}^{\ell_1}(\theta_{\text{pure-}d})$
8 **Output:** θ_{out}

952 **G.2 Pure DP Frank-Wolfe Algorithm**

953 **G.3 Proof of Theorem 4**

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

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

956 **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 27.

958 Denote the “failure” events as follows: $F_1 := \{u \leq \omega \text{ in Line 3 of Algorithm 1}\}$, where the uniform
959 sample is accepted in Algorithm 1; $F_2 := \{\Phi \text{ is not } (e, 4T)\text{-RWC}\}$, where the randomly sampled Φ
960 is not $(e, 4T)$ -restricted well-conditioned (RWC); and $F_3 := \{\|\mathbf{z}\|_1 > \xi\}$, where the Laplace noise
961 added in Line 4 of Algorithm 1 exceeds the tolerance threshold. We first bound the failure probability
962 $\mathbb{P}(F_1 \cup F_2 \cup F_3)$, and then analyze the utility under the “success” event $F_1^c \cap F_2^c \cap F_3^c$, followed by
963 an expected overall utility bound.

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

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

1 **Input:** Noisy measurement b , design matrix Φ , noise tolerant magnitude ξ
2 Define the feasible set $\mathcal{U} := \{\theta \in \mathbb{R}^d \mid \|\Phi\theta - b\|_1 \leq \xi\}$
3 Solve $\hat{\theta} = \arg \min_{\theta \in \mathcal{U}} \|\theta\|_1$
4 **Output:** $\hat{\theta}$

965 Set the distortion rate as $e = \frac{1}{4}$, and $k = \Theta\left(T \log\left(\frac{d}{T}\right) + \log n\right)$. Constructing Φ following
 966 Lemma 41, and by Lemma 42, $\Phi \in \mathbb{R}^{k \times d}$ is $(4T, e)$ -RWC with probability at least $1 - \frac{1}{n}$, i.e.,
 967 $\mathbb{P}(F_2) \leq \frac{1}{n}$.

968 Set

$$\xi = \frac{4\Delta}{\varepsilon}(k + \log n), \quad (11)$$

969 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 45 with taking $b = \frac{2\Delta}{\varepsilon}$ and $t =$
 970 $b(k + 2\log n)$, we have $\mathbb{P}(F_3) \leq \frac{1}{n}$.

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

$$\begin{aligned} \theta_{\text{pure-}k} &= \mathcal{A}_{\text{pure}}\left(x_{\text{apx}} = \text{Clip}_{2\|\mathcal{C}\|_1}^{\ell_2}(\theta_{\text{apx-}k}), \Theta = \mathcal{B}_{\ell_2}^k(2\|\mathcal{C}\|_1), \varepsilon' = \varepsilon, \delta, \omega\right) \quad (\text{Line 5 of Algorithm 7}) \\ &= \mathcal{A}_{\text{pure}}(\theta_{\text{apx-}k}, \Theta = \mathcal{B}_{\ell_2}^k(2\|\mathcal{C}\|_1), \varepsilon' = \varepsilon, \delta, \omega) \quad (\text{Under } F_2^c, \|\theta_{\text{apx-}k}\|_2 \leq (1+e)\|\mathcal{C}\|_1) \\ &= \theta_{\text{apx-}k} + \text{Lap}^{\otimes d}\left(\frac{2\Delta}{\varepsilon}\right) \quad (\text{Under } F_1^c) \\ &= \Phi\theta_{\text{FW}} + \tilde{w}, \text{ where } \|\tilde{w}\|_1 \leq \xi. \quad (\text{Under } F_3^c) \end{aligned}$$

972 By Lemma 43 and $\theta_{\text{pure-}d} = \mathcal{M}_{\text{rec}}(\theta_{\text{pure-}k}, \Phi, \xi)$, since θ_{FW} is T -sparse, we have

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

973 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 $\|\theta_{\text{pure-}d}\|_1 \leq$
 974 $\|\theta_{\text{FW}}\|_1 \leq \|\mathcal{C}\|_1$, therefore, by Line 7 in Algorithm 7, $\theta_{\text{out}} = \text{Clip}_{\|\mathcal{C}\|_1}^{\ell_1}(\theta_{\text{pure-}d}) = \theta_{\text{pure-}d}$ (under event
 975 F_3^c .)

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

$$\begin{aligned} \mathbb{E}[\mathcal{L}(\theta_{\text{out}}; D) - \mathcal{L}(\theta_{\text{FW}}; D) \mid F_1^c \cap F_2^c \cap F_3^c] &\leq L_1 \|\theta_{\text{out}} - \theta_{\text{FW}}\|_1 \\ &\leq L_1 \frac{4\sqrt{T}}{\sqrt{1-e}} \cdot \xi \\ &\leq \frac{64}{\sqrt{3}} \frac{L_1 \|\mathcal{C}\|_1 (k + \log n)}{n\varepsilon} \end{aligned} \quad (12)$$

977 **For (b):** by Lemma 26, we have:

$$\begin{aligned} \mathbb{E}\left[\mathcal{L}(\theta_{\text{FW}}; D) - \min_{\theta \in \mathcal{C}} \mathcal{L}(\theta; D) \mid F_1^c \cap F_2^c \cap F_3^c\right] &= \mathcal{O}\left(\frac{\Gamma_{\mathcal{L}}}{T} + \frac{L_1 \|\mathcal{C}\|_1 \sqrt{8T \log(1/\delta)} \log(TL_1 \|\mathcal{C}\|_1 \cdot |S|)}{n\varepsilon}\right) \\ &\leq \mathcal{O}\left(\frac{\beta \|\mathcal{C}\|_1^2}{T} + \frac{L_1 \|\mathcal{C}\|_1 \sqrt{8T \log(1/\delta)} \log(TL_1 \|\mathcal{C}\|_1 \cdot |S|)}{n\varepsilon}\right). \end{aligned}$$

978 Therefore,

$$\begin{aligned} &\mathbb{E}\left[\mathcal{L}(\theta_{\text{out}}; D) - \min_{\theta \in \mathcal{C}} \mathcal{L}(\theta; D) \mid F_1^c \cap F_2^c \cap F_3^c\right] \\ &\leq \mathcal{O}\left(\frac{L_1 \|\mathcal{C}\|_1 (k + \log n)}{n\varepsilon} + \frac{\beta \|\mathcal{C}\|_1^2}{T} + \frac{L_1 \|\mathcal{C}\|_1 \sqrt{8T \log(1/\delta)} \log(TL_1 \|\mathcal{C}\|_1 \cdot |S|)}{n\varepsilon}\right) \\ &= \tilde{\mathcal{O}}\left(\frac{L_1 \|\mathcal{C}\|_1 T}{n\varepsilon} + \frac{\beta \|\mathcal{C}\|_1^2}{T} + \frac{L_1 \|\mathcal{C}\|_1 T}{n\varepsilon}\right) \quad (\delta = \frac{2\omega}{(nk)^k}) \\ &= \tilde{\mathcal{O}}\left(\frac{\beta \|\mathcal{C}\|_1^2}{T} + \frac{L_1 \|\mathcal{C}\|_1 T}{n\varepsilon}\right). \end{aligned}$$

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

$$\mathbb{E} \left[\mathcal{L}(\theta_{\text{out}}; D) - \min_{\theta \in \mathcal{C}} \mathcal{L}(\theta; D) \mid F_1^c \cap F_2^c \cap F_3^c \right] \leq \tilde{\mathcal{O}} \left(\frac{(\beta L_1 \|\mathcal{C}\|_1^3)^{1/2}}{(n\varepsilon)^{1/2}} \right). \quad (13)$$

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

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

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

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

983 G.4 Proof of Lemma 28

Lemma 28 *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 $\Theta = \{\theta \in \mathbb{R}^d \mid \|\theta\|_1 \leq 1\}$ and a dataset $\text{Data} := \{x_1, \dots, x_n\} \in \mathcal{X}^n$ 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*

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

984 **Definition 29** ((α, β) -accurate ERM algorithm) *Given parameter space Θ , dataspace \mathcal{X} , and risk*
 985 *function R , we say an ERM algorithm $\mathcal{M} : \mathcal{X}^n \rightarrow \Theta$ is (α, β) -accurate if for any dataset $D \in \mathcal{X}^n$,*
 986 *with probability at least $1 - \beta$ over the randomness of algorithm:*

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

987 **Lemma 30 (Restatement of Lemma 28)** *There exists a hard instance with n samples over \mathcal{B}_1^d and a*
 988 *1-Lipschitz loss function \mathcal{L} such that any ε -pure differential private $(\alpha, 1/2)$ -accurate ERM algorithm*
 989 *\mathcal{M} must have:*

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

990 **Proof of Lemma 28** We proof by the standard packing argument. Consider \mathcal{B}_1^d and an α -packing
 991 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 $E_i = \{\theta \in$
 992 $\Theta \mid \|\theta - \theta_i\|_2 \leq \alpha\}$ and $X_i = \underbrace{\{\theta_i, \dots, \theta_i\}}_{n \text{ copies}}.$

993 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:

$$\begin{aligned} 1 &\geq \mathbb{P}(\mathcal{M}(X_i) \notin E_i) \geq \sum_{j \in [K] \setminus i} \mathbb{P}(\mathcal{M}(X_i) \in E_j) \\ &\geq \sum_{j \in [K] \setminus i} \exp(-n\varepsilon) \times \mathbb{P}(\mathcal{M}(X_j) \in E_j) \\ &\geq \frac{K \exp(-n\varepsilon)}{4} \end{aligned} \quad (14)$$

994 Taking log of both sides, we have

$$n\varepsilon + \log(4) \geq \log(K) \quad (15)$$

995 It remains to calculate the packing number K . Notice that $M(\alpha, \mathcal{B}_1^d, \|\cdot\|_2) \asymp N(\alpha, \mathcal{B}_1^d, \|\cdot\|_2) \asymp$
 996 $\exp\left(\frac{1}{\alpha^2} \log(\alpha^2 d)\right)^3$, where we assume $\alpha \gtrsim \frac{1}{\sqrt{d}}$ [76, Equation 15.4]. This implies:

$$\alpha \geq \tilde{\Omega}\left(\frac{1}{\sqrt{n\varepsilon + \log(4)}} \vee \frac{1}{\sqrt{d}}\right) \quad (16)$$

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

999 H Details for Data-dependent DP mechanism Design

1000 H.1 Pure DP Propose Test Release

1001 **Definition 31 (Local Sensitivity)** *The local sensitivity of a query function q on a dataset X is defined*
 1002 *as*

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

1003 where $X' \simeq X$ denotes that X' is a neighboring dataset of X .

1004 We first present the original version of PTR in Algorithm 9. The pure DP version, obtained via the
 1005 purification trick, is given in Algorithm 10. Their privacy guarantees are stated as follows:

1006 **Lemma 32** *Algorithm 9 satisfies $(2\varepsilon, \delta)$ -DP and its purified version, Algorithm 10 satisfies $(2\varepsilon +$
 1007 $\varepsilon', 0)$ -DP*

1008 **Proof** The privacy guarantee for Algorithm 9 is based on [67, Proposition 7.3.2]. For Algorithm 10,
 1009 the privacy guarantee follows from the post-processing property of differential privacy and the privacy
 1010 guarantee of Algorithm 1. ■

Algorithm 9 $\mathcal{M}_{\text{PTR}}(X, q, \varepsilon, \delta, \beta)$: Propose-Test-Release [22]

```

1 Input: Dataset  $X$ ; privacy parameters  $\varepsilon, \delta$ ; proposed bound  $\beta$ ; query function  $q : \mathcal{X} \rightarrow \Theta$ 
2 Compute:  $\mathcal{D}_\beta^q(X) = \min_{X'} \{d_{\text{Hamming}}(X, X') \mid \Delta_{\text{Local}}^q(X') > \beta\}$ 
3 if  $\mathcal{D}_\beta^q(X) + \text{Lap}\left(\frac{1}{\varepsilon}\right) \leq \frac{\log(1/\delta)}{\varepsilon}$  then
4   | Output  $\perp$ 
5 else
6   | Release  $f(X) + \text{Lap}\left(\frac{\beta}{\varepsilon}\right)$ 
end

```

Algorithm 10 Pure DP Propose-Test-Release

```

1 Input: Dataset  $X$ ; privacy parameters  $\varepsilon, \varepsilon', \delta$ ; proposed bound  $\beta$ ; query function  $q : \mathcal{X} \rightarrow \Theta$ , level  

  of uniform smoothing  $\omega$ 
2  $\theta \leftarrow \mathcal{M}_{\text{PTR}}(X, q, \varepsilon, \delta, \beta)$ 
3 if  $\theta == \perp$  then
4   |  $u \sim \text{Unif}(\Theta)$ 
5   |  $\theta \leftarrow u$ 
end
6  $\theta_{\text{out}} \leftarrow \mathcal{A}_{\text{pure}}(\theta, \Theta, \varepsilon', \varepsilon, \delta, \omega)$ 
7 Output:  $\theta_{\text{out}}$ 

```

▷ Algorithm 1

³ N denote covering number

1011 H.2 Privately Bounding Local Sensitivity

1012 We assume query function $q : \mathcal{X}^* \rightarrow \Theta$ with $\Theta \subset \mathbb{R}^d$ being a convex set and $\text{Diam}_2(\Theta) = R$.
 1013 Assume the global sensitivity of local sensitivity is upper bounded by 1: $\max_{X \simeq X'} \|\Delta_{\text{Local}}^q(X) - \Delta_{\text{Local}}^q(X')\|_2 \leq 1$. 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

- 1 **Input:** Dataset D ; privacy parameters $\varepsilon, \varepsilon', \delta$; proposed bound β ; query function $q : \mathcal{X}^* \rightarrow \Theta \subset \mathbb{R}^d$ with $\text{Diam}_2(\Theta) = R$, level of uniform smoothing ω
 - 2 $\hat{\beta} = \Delta_{\text{Local}}^q(D) + \text{Lap}(1/\varepsilon) + \log(2/\delta)/\varepsilon$
 - 3 $q_{\text{apx}} \leftarrow \text{Proj}_{\Theta}(q(D) + \text{Lap}^{\otimes d}(\hat{\beta}/\varepsilon))$
 - 4 $q_{\text{pure}} \leftarrow \mathcal{A}_{\text{pure}}(\Theta, \varepsilon', \omega, R)$ ▷ Algorithm 1
 - 5 **Output:** q_{pure}
-

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

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

1016 **Proof** First notice that $\hat{\beta}$ satisfies ε -DP by the privacy guarantee from Laplace mechanism and the
 1017 assumption that global sensitivity of $\Delta_{\text{Local}}^q(D)$ is upper bounded by 1. Second, we notice that
 1018 $\mathbb{P}(\hat{\beta} > \Delta_{\text{Local}}^q(D)) = \mathbb{P}(\text{Lap}(1/\varepsilon) \geq \log(2/\delta)/\varepsilon) = 1 - \delta$. This implies $q(D) + \text{Lap}^{\otimes d}(\hat{\beta}/\varepsilon)$
 1019 satisfies (ε, δ) -probabilistic DP ([25, Definition 2.2]), thus satisfies (ε, δ) -DP. By post-processing and
 1020 simple composition, q_{apx} satisfies $(2\varepsilon, \delta)$ -DP. Finally, using $\mathcal{A}_{\text{pure}}$ under appropriate choice of δ , we
 1021 have q_{pure} satisfies $(3\varepsilon, 0)$ -DP by Theorem 1.

1022 We now prove the utility guarantee. For notational convenience, we denote $z_0 \sim \text{Lap}(1/\varepsilon)$, $Z_1 \sim$
 1023 $\text{Lap}^{\otimes d}(\hat{\beta}/\varepsilon)$ and $Z_2 \sim \text{Lap}^{\otimes d}(\Delta/\varepsilon)$. By definition of q_{pure} , we have:

$$\begin{aligned} \mathbb{E} [\|q_{\text{pure}} - q(D)\|_2] &= \mathbb{E} [\|q_{\text{apx}} - q(D) + Z_2\|_2] + \omega C \\ &= \mathbb{E} [\|\text{Proj}_{\Theta}(q(D) + Z_1) - q(D) + Z_2\|_2] + \omega C \\ &\leq \mathbb{E} [\|Z_1\|_2] + \mathbb{E} [\|Z_2\|_2] + \omega C \end{aligned} \tag{17}$$

1024 Notice that:

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

1025 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 :

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

1026 Also notice that $\log(2/\delta) = \tilde{\mathcal{O}}(d)$. Thus,

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

1027 ■

Algorithm 12 Pure DP Mode Release

1 **Input:** Dataset D , pure DP parameter ε
 2 **Set:** $\log(1/\delta) = d \log(2d^3/\varepsilon)$, where $d = \log_2 |\mathcal{X}|$.
 3 Compute the mode $f(D)$ and its frequency occ_1 , as well as the frequency occ_2 of the second most frequent item.
 4 Compute the gap: $\mathcal{D}_0^f(D) \leftarrow \lceil \frac{\text{occ}_1 - \text{occ}_2}{2} \rceil$.
 5 **if** $\mathcal{D}_0^f(D) - 1 + \text{Lap}(\frac{1}{\varepsilon}) \leq \frac{\log(1/\delta)}{\varepsilon}$ **then**
 6 $u_{\text{apx}} \leftarrow \perp$
 7 **else**
 8 $u_{\text{apx}} \leftarrow f(D)$
 9 **end**
 9 $u_{\text{pure}} \leftarrow$ Algorithm 2 with inputs $\varepsilon, \delta, u_{\text{apx}}, \mathcal{Y} = \mathcal{X}$, and Index = id, the identity map.
 10 **Output:** u_{pure}

1028 H.3 Private Mode Release

1029 The mode release algorithm discussed in Section 5.2 is provided in Algorithm 12.

1030 **Proof of Theorem 6** By [67, Proposition 3.3] and that $\text{dist}(D, \{D' : f(D') \neq f(D)\}) =$
 1031 $\lceil \frac{\text{occ}_1 - \text{occ}_2}{2} \rceil$, we know u_{apx} satisfies (ε, δ) -DP. By [67, Proposition 3.4], when $\text{occ}_1 - \text{occ}_2 \geq$
 1032 $4 \lceil \ln(1/\delta)/\varepsilon \rceil$, u_{apx} is the exact mode, i.e., $u_{\text{apx}} = f(D)$, with probability at least $1 - \delta$. Choosing
 1033 $\delta < \frac{\varepsilon^d}{(2d)^{3d}}$, by Theorem 2 and the union bound, we have $\mathbb{P}(u_{\text{pure}} = f(D)) \geq 1 - \delta - 2^{-d} - \frac{d}{2}e^{-d} \geq$
 1034 $1 - \frac{3}{|\mathcal{X}|}$. ■

1035 H.4 Private Linear Regression Through Adaptive Sufficient Statistics Perturbation

1036 We investigate the problem of differentially private linear regression. Specifically, we consider a
 1037 fixed design matrix $X \in \mathbb{R}^{n \times d}$ and a response variable $Y \in \mathbb{R}^n$, which represent a collection of
 1038 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
 1039 $Y = X\theta^*$, our goal is to find a differentially private estimator θ that minimizes the mean squared
 1040 error:

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

1041 We assume prior knowledge of the magnitude of the dataspace: $\|\mathcal{X}\| := \sup_{x \in \mathcal{X}} \|x\|_2$ and $\|\mathcal{Y}\| :=$
 1042 $\sup_{y \in \mathcal{Y}} \|y\|_2$. Our algorithm operates under the same parameter settings as [73, Algorithm 2]. To
 1043 enable our purification procedure, we first derive a high-probability upper bound on $\|\tilde{\theta}\|_2$, where
 1044 $\tilde{\theta}$ is the output of AdaSSP. Subsequently, we clip the output of AdaSSP and apply the purification
 1045 procedure. The implementation details are provided in Algorithm 13.

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

1 **Input:** Data X, y ; privacy parameters $\epsilon, \epsilon', \delta$, Bounds: $\|\mathcal{X}\|, \|\mathcal{Y}\|$, level of smoothing ω
 2 $\theta_{\text{apx}} \leftarrow \text{AdaSSP}(X, y, \epsilon, \delta, \|\mathcal{X}\|, \|\mathcal{Y}\|)$ ▷ [73, Algorithm 2]
 3 Propose a high probability upper bound $\tilde{R} = \tilde{\mathcal{O}}((1+n\epsilon/d)\|\mathcal{Y}\|/\|\mathcal{X}\|)$
 4 Construct trust region $\Theta := \mathcal{B}_{\ell_2}^d(\tilde{R})$
 5 Norm clipping $\theta_{\text{apx}} \leftarrow \text{Proj}_{\Theta}(\theta_{\text{apx}})$
 6 $\theta_{\text{pure}} \leftarrow \mathcal{A}_{\text{pure}}(\theta_{\text{apx}}, \Theta, \epsilon, \epsilon', \delta)$
 7 **Output:** θ_{pure}

1046 **Theorem 12 (Restatement of theorem 7)** Assume $X^\top X$ is positive definite and $\|\mathcal{Y}\| \lesssim \|\mathcal{X}\| \|\theta^*\|$.
 1047 With probability $1 - \zeta - 1/n^2$, the output θ_{pure} from Algorithm 13 satisfies:

$$\text{MSE}(\theta_{\text{pure}}) - \text{MSE}(\theta^*) \leq \tilde{\mathcal{O}} \left(\frac{\|\mathcal{X}\|^2}{\epsilon^2 n^4} + \frac{d \|\mathcal{X}\|^2 \|\theta^*\|^2}{n \epsilon} \wedge \frac{d^2 \|\mathcal{X}\|^4 \|\theta^*\|^2}{\epsilon^2 n^2 \lambda_{\min}} \right) \quad (18)$$

1048 where $\lambda_{\min} := \lambda_{\min}(X^\top X/n)$.

1049 **Proof** First, we introduce a utility Lemma from [73]

1050 **Lemma 33 (Theorem 3 from [73])** Under the setting of [73, Algorithm 2], AdaSSP satisfies (ε, δ) -
 1051 DP. Assume $\|\mathcal{Y}\| \lesssim \|\mathcal{X}\| \|\theta^*\|$, then with probability $1 - \zeta$,

$$MSE(\tilde{\theta}) - MSE(\theta^*) \leq \mathcal{O} \left(\frac{\sqrt{d \log \left(\frac{d^2}{\zeta} \right)} \|\mathcal{X}\|^2 \|\theta^*\|^2}{n\varepsilon / \sqrt{\log \left(\frac{6}{\delta} \right)}} \wedge \frac{\|\mathcal{X}\|^4 \|\theta^*\|^2 \text{tr}[(X^T X)^{-1}]}{n\varepsilon^2 / [\log \left(\frac{6}{\delta} \right) \log \left(\frac{d^2}{\zeta} \right)]} \right) \quad (19)$$

1052 Now, we prove the utility guarantee for our results. In order to operate the purification technique, we
 1053 need to estimate the range $\|\tilde{\theta}\|_2$ in order to apply the uniform smoothing technique. Notice that

$$\|\tilde{\theta}\|_2 \leq \underbrace{\|(X^T X + \lambda I_d + E_1)^{-1}\|_2}_{(a)} \underbrace{\|X^T y + E_2\|_2}_{(b)} \quad (20)$$

1054 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
 1055 matrix, and each entry in its upper-triangular part (including the diagonal) is independently sampled
 1056 from $\mathcal{N}(0, 1)$. Under the high probability event in Lemma 33, we upper bound (a) and (b) separately:

1057 **For (a):**

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

1058 By the choice of λ and concentration of $\|E_1\|_2$, we have $(X^T X + \lambda I_d + E_1) \succ \frac{1}{2}(X^T X + \lambda I_d)$,
 1059 which allows a lower bound on $\lambda_{\min}(X^T X + \lambda I_d + E_1)$:

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

1060 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
 1061 λ_{\min}^* is a high probability lower bound on $\lambda_{\min}(X^T X)$. Thus,

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

1062 For (b), by triangle inequality

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

1063 Apply [73, Lemma 6], we have w.p. at least $1 - \beta$:

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

1064 Thus, w.p. at least $1 - \beta$ over the randomness of E_2 ,

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

1065 Putting everything together under the high probability event:

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

1066 Now, we apply purification Algorithm 1 with $\Theta = \mathcal{B}_{\ell_2}^d(\tilde{r})$, $\omega = \frac{1}{n^2}$, $\delta = \frac{2\omega}{(16d^{3/2}\tilde{r}n^2)^d}$. This
 1067 parameter configuration implies $\log(1/\delta) = \tilde{\mathcal{O}}(d)$ and Wasserstein- ∞ distance $\Delta = \frac{1}{4n^2d}$ (Line 2 of
 1068 Algorithm 1)

1069 Finally, it remains to bound the estimation error for θ_{pure} . Let's denote the additive Laplace noise
 1070 from purification by $Z_2 \sim \text{Lap}^{\otimes d}(\Delta/\varepsilon)$, and the purified estimator $\theta_{pure} := \tilde{\theta} + Z_2$. Under the
 1071 event that the purification algorithm doesn't output uniform noise, which happens w.p. at least $1 - \omega$:

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

1072 where the last inequality holds w.p. $1 - 1/n^2$ by Lemma 45, as we derived below:

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

1073 Under the high probability event of Lemma 33 and Algorithm 1, which happens with probability at
 1074 least $1 - \zeta - 1/n^2$:

$$\begin{aligned} \text{MSE}(\theta_{pure}) - \text{MSE}(\theta^*) &= \text{MSE}(\theta_{pure}) - \text{MSE}(\tilde{\theta}) + \text{MSE}(\tilde{\theta}) - \text{MSE}(\theta^*) \\ &\leq \tilde{\mathcal{O}} \left(\frac{\|\mathcal{X}\|^2}{\varepsilon^2 n^4} + \frac{d\|\mathcal{X}\|^2\|\theta^*\|^2}{n\varepsilon} \wedge \frac{d\|\mathcal{X}\|^4\|\theta^*\|^2 \text{tr}[(X^\top X)^{-1}]}{n\varepsilon^2} \right) \quad (21) \\ &\leq \tilde{\mathcal{O}} \left(\frac{\|\mathcal{X}\|^2}{\varepsilon^2 n^4} + \frac{d\|\mathcal{X}\|^2\|\theta^*\|^2}{n\varepsilon} \wedge \frac{d^2\|\mathcal{X}\|^4\|\theta^*\|^2}{n\varepsilon^2 \lambda_{\min}(X^\top X)} \right) \end{aligned}$$

1075 By denoting $\lambda_{\min} = \lambda_{\min} \left(\frac{X^\top X}{n} \right)$, we have:

$$\text{MSE}(\theta_{pure}) - \text{MSE}(\theta^*) \leq \tilde{\mathcal{O}} \left(\frac{\|\mathcal{X}\|^2}{\varepsilon^2 n^4} + \frac{d\|\mathcal{X}\|^2\|\theta^*\|^2}{n\varepsilon} \wedge \frac{d^2\|\mathcal{X}\|^4\|\theta^*\|^2}{\varepsilon^2 n^2 \lambda_{\min}} \right) \quad (22)$$

1076

■

1077 I Details for Private Query Release

1078 I.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 $q : \mathcal{X} \rightarrow [0, 1]$ and workload Q with size K . For a shorthand, we denote:

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

1079 I.2 Private Multiplicative Weight Exponential Mechanism

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

Algorithm 14 Multipliative Weight Exponential Mechanism $\text{MWEM}(D, Q, T, \rho)$ [38]

- 1 **Input:** Data set $D \in \mathbb{N}^{|\mathcal{X}|}$, set Q of linear queries; Number of iterations $T \in \mathbb{N}$; zCDP Privacy parameter $\rho > 0$.
 - 2 **Set:** number of data points $n \leftarrow \|D\|_1$, initial distribution $p_0 \leftarrow \frac{1_{|\mathcal{X}|}}{|\mathcal{X}|}$, privacy budget for each mechanism $\varepsilon_0 \leftarrow \sqrt{\rho/T}$
 - 3 **Define:** $\text{Score}(\cdot; \hat{p}, p) = |\langle \cdot, \hat{p} \rangle - \langle \cdot, p \rangle|$
 - 4 **for** $t = 1$ **to** T **do**
 - 5 $q_t \leftarrow \text{ExpoMech}(Q, \varepsilon_0, \text{Score}(\cdot; p_{t-1}, p))$
 - 6 $m_t \leftarrow \langle q_t, X \rangle + \text{Laplace}(1/n\varepsilon_0)$
 - 7 *Multiplicative weights update:* let p_{t+1} be the distribution over \mathcal{X} with entries satisfy:
$$q_t(x) \propto q_{t-1}(x) \cdot \exp(q_t(x) \cdot (m_t - q_t(p_{t-1}))/2), \forall x \in \mathcal{X}$$
 - end**
 - 8 **Output:** $D_{\text{out}} \leftarrow \frac{n}{T} \sum_{t=0}^{T-1} p_t$.
-

1080

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

- 1 **Input:** Dataspace \mathcal{X} , Probability vector $p \in \mathbb{R}^{|\mathcal{X}|}$, sample size m
 - 2 **for** $i = 1$ **to** m **do**
 - 3 $s_i \leftarrow \text{UnSortedProportionalSampling}(p, \mathcal{X})$ ▷ e.g., Alias method [71]
 - end**
 - 4 **Output:** $\{s_1, \dots, s_m\}$
-

Algorithm 16 Pure DP Multiplicative Weight Exponential Mechanism

- 1 **Input:** Dataset $D \in \mathbb{N}^{|\mathcal{X}|}$ with size $\|D\|_1 = n$, Query set Q , privacy parameters ϵ, δ , accuracy parameter α
 - 2 **Set:** Number of iterations $T = \tilde{O}(\epsilon^{2/3} n^{2/3} d^{1/3})$, size of subsampled dataset $m = n^{2/3} \epsilon^{2/3} d^{-2/3}$, zCDP budget $\rho = \epsilon^2 / 16 \log(1/\delta)$, $\text{Score}(q; \hat{p}, p) = |\langle q, \hat{p} \rangle - \langle q, p \rangle|, \forall q \in Q$
 - 3 **Initilize:** $p_1 \leftarrow \frac{1_{|\mathcal{X}|}}{|\mathcal{X}|}, p \leftarrow \frac{D}{\|D\|_1}$
 - 4 $\hat{D} \leftarrow \text{MWEM}(D, Q, T, \rho)$ ▷ Algorithm 14
 - 5 $Y \leftarrow \text{ProportionalSampling}(\mathcal{X}, \hat{D}/n, m)$ ▷ Algorithm 16
 - 6 $\hat{Y} \leftarrow \mathcal{A}_{\text{pure-discrete}}(\epsilon, \delta, Y, \mathcal{X}^m)$
 - 7 **Output:** \hat{Y}
-

1081 **Lemma 34 (Privacy and Utility of MWEM [38])** *Algorithm 14 instantiated as*

1082 *$\text{MWEM}(D, Q, T, \varepsilon^2/16 \log(1/\delta))$ satisfies (ε, δ) -DP. With probability $1 - 2\beta T$, PMW and the output \hat{D}*
 1083 *such that:*

$$\|Q(\hat{D}) - Q(D)\|_\infty \leq \mathcal{O} \left(\sqrt{\frac{d}{T}} + \frac{\sqrt{T \log(1/\delta) \log(K/\beta)}}{n\varepsilon} \right) \quad (23)$$

1084 **Proof** We first state the privacy guarantee. Since each iteration satisfies zCDP guarantee
 1085 ρ/T -zCDP, the total zCDP guarantee for T iterations is ρ . By Lemma 20, the whole algo-
 1086 rithm satisfies $(4\sqrt{\rho \log(1/\delta)}, \delta)$ -DP. Plugging in the choice of $\rho = \varepsilon^2/16 \log(1/\delta)$, we have
 1087 $\text{MWEM}(D, Q, T, \varepsilon^2/16 \log(1/\delta))$ satisfies (ε, δ) -DP. The utility guarantee follows [38, Theorem 2.2].
 1088 Specifically, we choose $\text{adderr} = 2\sqrt{T/\rho} \log(K/\beta)$, this yields the utility guarantee stated in
 1089 Theorem with probability $1 - 2\beta T$. ■

1090 **Lemma 35 (Sampling bound, Lemma 4.3 in [24])** *Let data $X = (a_1, \dots, a_N)$ with $\sum_{i=1}^N a_i = 1$*
 1091 *and $a_i \geq 0$. $Y \sim \text{Multinomial}(m, X)$. Then we have:*

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

1092 **Proof** The proof follows the proof of [24, Lemma 4.3]. Since we have $Y = (Y_1, \dots, Y_m)$ with
 1093 $Y_i \stackrel{\text{iid}}{\sim} \text{Multinomial}(1, X)$, for any $q \in Q$, we have $q(Y) = \frac{1}{m} \sum_{i=1}^m q(Y_i)$ and $\mathbb{E}[q(Y)] = q(X)$.
 1094 By the Chernoff bound and a union bound over all queries in Q , we have $\mathbb{P}[\|Q(Y) - Q(X)\|_\infty \geq$
 1095 $\alpha] \leq 2|Q| \exp(-2m\alpha^2)$. ■

1096 **Theorem 13 (Restatement of Theorem 8)** *Algorithm 16 satisfies 2ε -DP. Moreover, the output \hat{Y}*
 1097 *satisfies*

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

1098 *and the runtime is $\tilde{\mathcal{O}}(nK + \varepsilon^{2/3} n^{2/3} d^{1/3} NK + N)$.*

1099 **Proof** We first state the privacy guarantee. By Lemma 34, the output from multiplicative weight
 1100 exponential mechanism, \hat{D} , satisfies (ε, δ) -DP. By post-processing, Y is also (ε, δ) -DP. Thus, apply
 1101 Theorem 2, the purified \hat{Y} satisfies 2ε -DP.

1102 The utility guarantee is via bounding the following terms:

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

1103 For (c): Since the output space $\mathcal{Y} = \mathcal{X}^m$, if use binary encoding, the length of code is $\log_2(|\mathcal{Y}|) =$
 1104 $md := \tilde{d}$. Thus, by Theorem 2, we choose $\delta = \frac{\varepsilon^d}{(2d)^{3d}}$, this implies $\log(1/\delta) = \mathcal{O}(md \log(2md/\varepsilon))$
 1105 and failure probability $\beta_0 = \frac{1}{2^{md}} + \frac{md}{\exp(md)}$.

1106 For (a): In order to minimize upper bound in Equation 23, we choose $T = \frac{d^{1/2} n \varepsilon}{\log^{1/2}(1/\delta) \log(K/\beta)}$, this
 1107 implies $\|Q(D) - Q(\hat{D})\|_\infty \leq \frac{d^{1/4} \log^{1/4}(1/\delta) \log^{1/2}(K/\beta)}{(n\varepsilon)^{1/2}} = \mathcal{O} \left(\frac{d^{1/2} m^{1/4} \log^{1/4}(2md/\varepsilon) \log^{1/2}(K/\beta)}{(n\varepsilon)^{1/2}} \right)$

1108 For (b): using Sampling bound (Lemma 35) and setting failure probability $\beta_1 = 2K \exp(-2m\alpha^2)$,
 1109 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:

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

1110 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

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

1111 The computational guarantee follows: (1) Since $T = \frac{d^{1/2}n\varepsilon}{\log^{1/2}(1/\delta)\log(K/\beta)} = \tilde{\mathcal{O}}(\varepsilon^{2/3}n^{2/3}d^{1/3})$. The
 1112 runtime for MWEM is $\tilde{\mathcal{O}}(nK + \varepsilon^{2/3}n^{2/3}d^{1/3}NK)$ [38]; (2) For subsampling, by runtime analysis
 1113 of Alias method [71], the preprocessing time is $\mathcal{O}(N)$ and the query time is $\mathcal{O}(m)$ for generating
 1114 m samples. Thus, total runtime is $\mathcal{O}(N + (n\varepsilon)^{2/3}d^{-2/3})$; (3) Finally, note that the query time
 1115 for Algorithm 2 is $\mathcal{O}(\bar{d}) = \mathcal{O}(d^{1/3}(n\varepsilon)^{2/3})$. We conclude that the runtime for Algorithm 16 is
 1116 $\tilde{\mathcal{O}}(nK + \varepsilon^{2/3}n^{2/3}d^{1/3}NK + N)$. ■

1117 J Details for the Lower Bound

1118 **Lemma 36 (Lemma 5.1 in [6])** *Let $n, d \in \mathbb{N}$ and $\epsilon > 0$. There is a number $M = \Omega(\min(n, d/\epsilon))$
 1119 such that for every ϵ -differentially private algorithm \mathcal{A} , there is a dataset $D = \{x_1, \dots, x_n\} \subset$
 1120 $\{-1/\sqrt{d}, 1/\sqrt{d}\}^d$ with $\|\sum_{i=1}^n x_i\|_2 \in [M - 1, M + 1]$ such that, with probability at least $1/2$ (taken
 1121 over the algorithm random coins), we have*

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

1122 where $\bar{D} = \frac{1}{n} \sum_{i=1}^n x_i$.

1123 **Theorem 14 (Restatement of Theorem 9)** *Denote $\mathcal{D} := \{-1/\sqrt{d}, 1/\sqrt{d}\}^d$. Let $\varepsilon \leq \mathcal{O}(1)$, and
 1124 $\delta \in \left(\frac{1}{\exp(4d \log(d) \log^2(nd))}, \frac{1}{4n^d \log^{2d}(8d)}\right)$. For any (ε, δ) -DP mechanism \mathcal{M} , there exist a dataset
 1125 $D \in \mathcal{D}^n$ such that w.p. at least $1/4$ over the randomness of \mathcal{M} :*

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

1126 Here, $\tilde{\Omega}(\cdot)$ hides all polylogarithmic factors, except those with respect to δ .

1127 **Proof** Suppose there exists an (ε, δ) -differentially private mechanism \mathcal{M} such that with probability
 1128 at least $3/4$ over the randomness of \mathcal{M} , for any $D \in \mathcal{D}$,

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

1129 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:

$$\delta > \exp(-a^2 d / \log(d)) \quad (24)$$

1130 We execute Algorithm 1 to purify \mathcal{M} directly over the output space $[-1/\sqrt{d}, 1/\sqrt{d}]^d$. Let Y denote the
 1131 output of Algorithm 1 and $U \sim \text{Unif}([-1/\sqrt{d}, 1/\sqrt{d}]^d)$. The remainder of the proof involves bounding
 1132 the additional errors introduced during the purification process. By triangle inequality we have

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

1133 Notice that under the event that Line 3 of Algorithm 1 doesn't return the uniform random variable,
 1134 which happens with probability $1 - \omega$, we have $Y = \mathcal{M}(X) + \text{Laplace}^{\otimes d}(2\Delta/\varepsilon)$, so term (b) equals
 1135 the 2-norm of the Laplace perturbation.

1136 For the remaining proofs, we choose the mixing level $\omega = 1/8$ in Algorithm 1. We now justify the
1137 choice of δ :

1138 Observe that since $Y = \mathcal{M}(X) + \text{Laplace}^{\otimes d}(2\Delta/\varepsilon)$, term (b), which accounts for the error intro-
1139 duced by Laplace noise. With probability at least $7/8$ by the concentration of the L_2 norm of Laplace
1140 vector:

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

1141 Thus, without loss of generality, we require $\frac{2\sqrt{d}\Delta \log(8d)}{\varepsilon} \leq \frac{16d}{n\varepsilon \log(8d)}$, this implies

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

1142 Notice that:

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

1143 Choosing $q = \infty$ (corresponding to the use of ℓ_∞ norm in W - ∞ distance), and noticing $R = 2/\sqrt{d}$
1144 and $r = 1/\sqrt{d}$, we obtain the condition:

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

1145 which further implies:

$$\delta \leq \frac{1}{4n^d \log^{2d}(8d)} \quad (25)$$

1146 By Eq. (31) and Eq. (32), we have:

$$\delta \in \left(\exp(-a^2 d / \log(d)), \frac{1}{4n^d \log^{2d}(8d)} \right) \quad (26)$$

1147 The constrained above yields a lower bound on a^2 , after some relaxation for simplicity (and assume
1148 $d \geq \log(8d)$ and $d \log(n) > \log(4)$):

$$a^2 \geq 2 \log(d) \log(nd) \quad (27)$$

1149 Thus, we set $a = 2 \log(d) \log(nd)$ to satisfy the constraint stated in Eq. (34), and now

$$\delta \in \left(\frac{1}{\exp(4d \log(d) \log^2(nd))}, \frac{1}{4n^d \log^{2d}(8d)} \right) \quad (28)$$

1150 When δ is within the range above, we have:

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

1151 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 36. Thus, for any (ε, δ) -DP mechanism \mathcal{M} with δ
being in the range of Eq. (35), there exists a dataset $D \in \mathcal{D}$, such that with probability greater than
 $1/4$ over the randomness of \mathcal{M} :

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

1152 Here, $\tilde{\Omega}(\cdot)$ hides all polylogarithmic factors, except those with respect to δ . ■

1153 K Technical Lemmas

1154 K.1 Supporting Results on Sparse Recovery

1155 For completeness, we introduce the results from sparse recovery [66] that is used in Section 4.2 and
1156 Appendix G.

1157 **Definition 37 (Numerical sparsity)** A vector x is s -numerically sparse if $\frac{\|x\|_1^2}{\|x\|_2^2} \leq s$.

1158 Numerical sparsity extends the traditional notion of sparsity. By definition, an s -sparse vector is also
1159 s -numerically sparse. A notable property of numerical sparsity is that the difference between a sparse
1160 vector and a numerically sparse vector remains numerically sparse, as stated in the following lemma.

1161 **Lemma 38 (Difference of numerically sparse vectors)** Let $x \in \mathbb{R}^d$ be an s -sparse vector. For any
1162 vector $x' \in \mathbb{R}^d$ satisfying $\|x'\|_1 \leq \|x\|_1$, the difference $x' - x$ is $4s$ -numerically sparse.

1163 **Proof** Let $S := \{i \in [d] \mid x[i] \neq 0\}$ and denote $v := x' - x$. We have

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

1164 which implies $\|v_S\|_1 \geq \|v_{S^c}\|_1$. Therefore,

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

1165 which implies $\|v\|_1^2 \leq 4s\|v\|_2^2$. Thus, by Definition 37, v satisfies $4s$ -numerically sparse. \blacksquare

1166 If the vector x is s -sparse, we can reduce its dimension while preserving the ℓ_2 norm using matrices
1167 that satisfy the Restricted Isometry Property.

1168 **Definition 39 ((e, s) -Restricted isometry property (RIP))** A matrix $\Phi \in \mathbb{R}^{k \times d}$ satisfies the (e, s) -
1169 Restricted Isometry Property (RIP) if, for any s -sparse vector $x \in \mathbb{R}^d$ and some $e \in (0, 1)$, the
1170 following holds:

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

1171 For numerically sparse vectors, we can reduce the dimension while preserving utility by matrices
1172 satisfying a related condition – the Restricted well-conditioned (RWC).

1173 **Definition 40 ((e, s) -Restricted well-conditioned (RWC) ([66], Definition 4))** A matrix $\Phi \in$
1174 $\mathbb{R}^{k \times d}$ is (e, s) -Restricted well-conditioned (RWC) if, for any s -numerically sparse vector $x \in \mathbb{R}^d$ and
1175 some $e \in (0, 1)$, we have

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

1176 **Lemma 41 ([66], Lemma 2; [15], Theorem 1.4)** Let $\Phi \in \mathbb{R}^{k \times d}$ whose entries are independent and
1177 identically distributed Gaussian with mean zero and variance $\mathcal{N}(0, \frac{1}{k})$. For $e, \zeta \in (0, 1)$, if

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

1178 for an appropriate constant C , then Φ satisfies (e, s) -RIP with probability $\geq 1 - \zeta$.

1179 There is a connection between RIP and RWC matrices:

1180 **Lemma 42 ([66], 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
1181 (e, s) -RWC.

1182 Finally, we provide the following guarantee for Algorithm 8.

1183 **Lemma 43 (ℓ_1 error guarantee from sparse recovery)** Let $\Phi \in \mathbb{R}^{m \times n}$, and $\theta_* \in \mathbb{R}^n$. Given
1184 noisy observation $b = \Phi\theta_* + \tilde{w}$ with bounded ℓ_1 norm of noise, i.e. $\|\tilde{w}\|_1 \leq \xi$, consider the following
1185 noisy sparse recovery problem:

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

1186 where $\xi > 0$ is the constraint of the noise magnitude. Suppose that θ is an s -sparse vector in \mathbb{R}^n and
 1187 that Φ is a $(4s, e)$ -RWC matrix. Then, the following ℓ_1 estimation error bound holds:

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

1188 Moreover, the problem can be solved in $\mathcal{O}((3m + 4n + 1)^{1.5}(2n + m)\text{prec})$ arithmetic operations
 1189 in the worst case, with each operation being performed to a precision of $\mathcal{O}(\text{prec})$ bits.

1190 **Proof** For utility guarantee: By $\|\tilde{w}\|_1 \leq \xi$, θ_* is a feasible solution. Thus, we have $\|\hat{\theta}\|_1 \leq \|\theta_*\|_1$,
 1191 which implies $h := \hat{\theta} - \theta_*$ is $4s$ -numerically sparse by Lemma 38. Since Φ is $(e, 4s)$ -RWC, we have:

$$(1 - e)\|h\|_2^2 \leq \|\Phi h\|_2^2 \leq (1 + e)\|h\|_2^2 \quad (29)$$

1192 Now it remains to bound $\|h\|_1$:

$$\begin{aligned} \|h\|_1 &\leq \sqrt{4s}\|h\|_2 \\ &\leq \sqrt{4s} \cdot \frac{\|\Phi h\|_2}{\sqrt{1-e}} \\ &\leq \frac{2\sqrt{s}}{\sqrt{1-e}} (\|\Phi \hat{\theta} - b\|_1 + \|\Phi \theta_* - b\|_1) \\ &\leq \frac{4\sqrt{s}}{\sqrt{1-e}} \cdot \xi \end{aligned} \quad (30)$$

1193 where the last inequality is by feasibility of $\hat{\theta}$ and the structure of b .

1194 Now we prove the runtime guarantee. We first reformulate this problem to Linear Programming:

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

1195 subject to:

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

1196 The problem (P) has $2n + m$ variables and $2m + 2n + 1$ constraints. By [68], this can be solved
 1197 in $\mathcal{O}((3m + 4n + 1)^{1.5}(2n + m)B)$ arithmetic operations in the worst case, with each operation
 1198 being performed to a precision of $\mathcal{O}(B)$ bits. ■

1199 K.2 A Concentration Inequality for Laplace Random Variables

1200 **Definition 44 (Laplace Distribution)** $X \sim \text{Lap}(b)$ if its probability density function satisfies

1201 $f_X(t) = \frac{1}{2b} \exp\left(-\frac{|x|}{b}\right).$

1202 **Lemma 45 (Concentration of the ℓ_1 norm of Laplace vector)** Let $X = (x_1, \dots, x_k)$ with each
 1203 x_i independently identically distributed as $\text{Lap}(b)$. Then, with probability at least $1 - \zeta$,

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

1204 **Proof** $\|X\|_1 = \sum_{i=1}^k |x_i|$ follows the Gamma distribution $\Gamma(k, b)$, with probability density function
 1205 $f(x) = \frac{1}{\Gamma(k)b^k} x^{k-1} e^{-\frac{x}{b}}$. Applying the Chernoff's tail bound of Gamma distribution $\Gamma(k, b)$, we
 1206 have

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

1207 Taking $t = 2kb + 2b \log(1/\zeta)$, we have

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

1208 ■

1209 L Extended Lower Bounds

1210 In this section, we present an extended result for Theorem 9. Additionally, we establish a lower bound
1211 for the discrete setting, as stated in Theorem 16, thereby demonstrating that the purifying recipe for
1212 proving lower bound remains applicable in the discrete case.

1213 L.1 One-Way Marginal Release

1214 We establish a stronger version of Theorem 9, as stated in Theorem 15, which strengthens Theorem 9
1215 by establishing that the lower bound holds for any $c \in (0, 1)$, rather than being restricted to $c = 1/2$.

1216 **Theorem 15 (Restatement of Theorem 9)** Denote $\mathcal{D} := \{-1/\sqrt{d}, 1/\sqrt{d}\}^d$. Let $\varepsilon \leq \mathcal{O}(1)$, for any
1217 $c \in (0, 1)$, 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
1218 $D \in \mathcal{D}^n$ such that with probability at least $1/4$ over the randomness of \mathcal{M} :

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

1219 Here, $\tilde{\Omega}(\cdot)$ hides all polylogarithmic factors, except those with respect to δ .

1220 **Proof** Suppose for some $c \in (0, 1)$, there exists an (ε, δ) -differentially private mechanism \mathcal{M} such
1221 that with probability at least $3/4$ over the randomness of \mathcal{M} , for any $D \in \mathcal{D}$,

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

1222 where a is a term involving n and d , to be specified later. For the purpose of causing contradiction,
1223 we let:

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

1224 This implies:

$$\delta > \exp(-a^{\frac{1}{1-c}} d / \log(d)) \tag{31}$$

1225 We execute Algorithm 1 to purify \mathcal{M} directly over the output space $[-1/\sqrt{d}, 1/\sqrt{d}]^d$. Let Y denote the
1226 output of Algorithm 1 and $U \sim \text{Unif}([-1/\sqrt{d}, 1/\sqrt{d}]^d)$. The remainder of the proof involves bounding
1227 the additional errors introduced during the purification process. By triangle inequality we have

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

1228 Notice that under the event that Line 3 of Algorithm 1 doesn't return the uniform random variable,
1229 which happens with probability $1 - \omega$, we have $Y = \mathcal{M}(X) + \text{Laplace}^{\otimes d}(2\Delta/\varepsilon)$, so term (b) equals
1230 the 2-norm of the Laplace perturbation.

1231 For the remaining proofs, we choose the mixing level $\omega = 1/8$ in Algorithm 1. We now justify the
1232 choice of δ :

1233 Observe that since $Y = \mathcal{M}(X) + \text{Laplace}^{\otimes d}(2\Delta/\varepsilon)$, term (b), which accounts for the error intro-
 1234 duced by Laplace noise. With probability at least $7/8$ by the concentration of the L_2 norm of Laplace
 1235 vector:

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

1236 Thus, without loss of generality, we will require

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

1237 This implies:

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

1238 Notice that:

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

1239 Choosing $q = \infty$ (corresponding to the use of ℓ_∞ norm in the Wassertain- ∞ distance), and noticing
 1240 $R = 2/\sqrt{d}$ and $r = 1/\sqrt{d}$, we obtain the condition:

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

1241 which further implies:

$$\delta \leq \frac{1}{4n^d \log^{2d}(8d)} \quad (32)$$

1242 By Eq. (31) and Eq. (32), we have:

$$\delta \in \left(\exp \left(-a^{\frac{1}{1-c}} d / \log(d) \right), \frac{1}{4n^d \log^{2d}(8d)} \right) \quad (33)$$

1243 The constrained above yields a lower bound on a , after some relaxation for simplicity (and assume
 1244 $d \geq \log(8d)$ and $d \log(n) > \log(4)$):

$$a^{\frac{1}{1-c}} \geq 2 \log(d) \log(nd) \quad (34)$$

1245 Thus, we set $a = (2 \log(d) \log(nd))^{1-c}$ to satisfy the constraint stated in Eq. (34), and now

$$\delta \in \left(\frac{1}{\exp(2d \log(nd))}, \frac{1}{4n^d \log^{2d}(8d)} \right) = \left(\frac{1}{n^{2d} d^{2d}}, \frac{1}{4n^d \log^{2d}(8d)} \right) \quad (35)$$

1246 When δ is within the range above, we have:

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

1247 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 36. Thus, for any (ε, δ) -DP mechanism \mathcal{M} with δ
 being in the range of Eq. (35), there exists a dataset $D \in \mathcal{D}$, such that with probability greater than
 $1/4$ over the randomness of \mathcal{M} :

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

1248 Here, $\tilde{\Omega}(\cdot)$ hides all polylogarithmic factors, except those with respect to δ . Since the above derivation
 1249 holds for arbitrary $c \in (0, 1)$, this implies with probability at least $1/4$ over the randomness of the
 1250 algorithm \mathcal{M} , we have:

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

1251

■

1252 L.2 Private Selection

1253 We begin by stating a lower bound for pure differential privacy in the selection setting, as established
1254 in [16].

1255 **Lemma 46 (Proposition 1 in [16])** *Let $\varepsilon \in (0, 1)$, $n \geq 2$ and denote item set to be \mathcal{U} . For any*
1256 *ε -DP mechanism \mathcal{A} , there exist a domain \mathcal{X} and a function $f(i, \cdot)$ which is $(1/n)$ -Lipschitz for all*
1257 *item $i \in \mathcal{U}$ such that the following holds with probability at least $1/2$ over the randomness of the*
1258 *algorithm:*

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

1259 **Theorem 16 (Lower bound for private selection)** *Let $\varepsilon \in (0, 1)$, $\delta \in \left(\frac{\varepsilon^{3d}}{(2d)^{3d}}, \frac{\varepsilon^d}{(2d)^{3d}}\right)$, $n \geq 2$,*
1260 *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}*
1261 *and a function $f(i, \cdot)$ which is $(1/n)$ -Lipschitz for all item $i \in \mathcal{U}$ such that the following holds with*
1262 *probability at least $1/2$ over the randomness of the algorithm:*

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

1263 **Proof** Without loss of generality, we set $d = \lceil \log_2 K \rceil$, we have that $\log K = \Theta(d)$. For any $c \in$
1264 $(0, 1)$, assume there exists an (ε, δ) -DP algorithm such that with probability at least $\frac{1}{2} + 2^{-d} + \frac{d}{2 \exp(d)}$,
1265 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
1266 term involved with n, d which will be specified later.

1267 First, to ensure the quality of purification, we need to set $\delta \leq \varepsilon^d (2d)^{-3d}$, this ensures with probability
1268 at least $1 - 2^{-d} - \frac{d}{2} \exp(-d)$ over the randomness of purification algorithm, we have $\mathcal{A}^{\text{purified}}(D) =$
1269 $\mathcal{A}(D)$.

1270 Further, in order to fulfill contrast argument, without loss of generality, we require

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

1271 which implies:

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

1272 Thus, to ensure the lower bound of δ doesn't exceed the upper bound of δ , we require:

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

1273 So, we set $a^{\frac{1}{1-c}} = 3 \log(d) \log\left(\frac{2d}{\varepsilon}\right)$, this implies

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

1274 This implies with probability at least $1/2$,

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

1275 Observe that, under the assumption of $K := |\mathcal{U}| \geq 7$ which implies $d \geq \exp(1)$, the inequality above
1276 contradicts Lemma 46. Since $c \in (0, 1)$ was chosen arbitrarily, this completes the proof of the stated
1277 theorem. ■

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]

1325 Justification: We provided limitations.

1326 Guidelines:

- 1327 • The answer NA means that the paper has no limitation while the answer No means that
1328 the paper has limitations, but those are not discussed in the paper.
- 1329 • The authors are encouraged to create a separate "Limitations" section in their paper.
- 1330 • The paper should point out any strong assumptions and how robust the results are to
1331 violations of these assumptions (e.g., independence assumptions, noiseless settings,
1332 model well-specification, asymptotic approximations only holding locally). The authors
1333 should reflect on how these assumptions might be violated in practice and what the
1334 implications would be.
- 1335 • The authors should reflect on the scope of the claims made, e.g., if the approach was
1336 only tested on a few datasets or with a few runs. In general, empirical results often
1337 depend on implicit assumptions, which should be articulated.
- 1338 • The authors should reflect on the factors that influence the performance of the approach.
1339 For example, a facial recognition algorithm may perform poorly when image resolution
1340 is low or images are taken in low lighting. Or a speech-to-text system might not be
1341 used reliably to provide closed captions for online lectures because it fails to handle
1342 technical jargon.
- 1343 • The authors should discuss the computational efficiency of the proposed algorithms
1344 and how they scale with dataset size.
- 1345 • If applicable, the authors should discuss possible limitations of their approach to
1346 address problems of privacy and fairness.
- 1347 • While the authors might fear that complete honesty about limitations might be used by
1348 reviewers as grounds for rejection, a worse outcome might be that reviewers discover
1349 limitations that aren't acknowledged in the paper. The authors should use their best
1350 judgment and recognize that individual actions in favor of transparency play an impor-
1351 tant role in developing norms that preserve the integrity of the community. Reviewers
1352 will be specifically instructed to not penalize honesty concerning limitations.

1353 3. Theory assumptions and proofs

1354 Question: For each theoretical result, does the paper provide the full set of assumptions and
1355 a complete (and correct) proof?

1356 Answer: [Yes]

1357 Justification: We provide complete assumptions and proofs.

1358 Guidelines:

- 1359 • The answer NA means that the paper does not include theoretical results.
- 1360 • All the theorems, formulas, and proofs in the paper should be numbered and cross-
1361 referenced.
- 1362 • All assumptions should be clearly stated or referenced in the statement of any theorems.
- 1363 • The proofs can either appear in the main paper or the supplemental material, but if
1364 they appear in the supplemental material, the authors are encouraged to provide a short
1365 proof sketch to provide intuition.
- 1366 • Inversely, any informal proof provided in the core of the paper should be complemented
1367 by formal proofs provided in appendix or supplemental material.
- 1368 • Theorems and Lemmas that the proof relies upon should be properly referenced.

1369 4. Experimental result reproducibility

1370 Question: Does the paper fully disclose all the information needed to reproduce the main ex-
1371 perimental results of the paper to the extent that it affects the main claims and/or conclusions
1372 of the paper (regardless of whether the code and data are provided or not)?

1373 Answer: [NA]
 1374 Justification: NA.
 1375 Guidelines:

- 1376 • The answer NA means that the paper does not include experiments.
- 1377 • If the paper includes experiments, a No answer to this question will not be perceived
 1378 well by the reviewers: Making the paper reproducible is important, regardless of
 1379 whether the code and data are provided or not.
- 1380 • If the contribution is a dataset and/or model, the authors should describe the steps taken
 1381 to make their results reproducible or verifiable.
- 1382 • Depending on the contribution, reproducibility can be accomplished in various ways.
 1383 For example, if the contribution is a novel architecture, describing the architecture fully
 1384 might suffice, or if the contribution is a specific model and empirical evaluation, it may
 1385 be necessary to either make it possible for others to replicate the model with the same
 1386 dataset, or provide access to the model. In general, releasing code and data is often
 1387 one good way to accomplish this, but reproducibility can also be provided via detailed
 1388 instructions for how to replicate the results, access to a hosted model (e.g., in the case
 1389 of a large language model), releasing of a model checkpoint, or other means that are
 1390 appropriate to the research performed.
- 1391 • While NeurIPS does not require releasing code, the conference does require all submis-
 1392 sions to provide some reasonable avenue for reproducibility, which may depend on the
 1393 nature of the contribution. For example
 - 1394 (a) If the contribution is primarily a new algorithm, the paper should make it clear how
 1395 to reproduce that algorithm.
 - 1396 (b) If the contribution is primarily a new model architecture, the paper should describe
 1397 the architecture clearly and fully.
 - 1398 (c) If the contribution is a new model (e.g., a large language model), then there should
 1399 either be a way to access this model for reproducing the results or a way to reproduce
 1400 the model (e.g., with an open-source dataset or instructions for how to construct
 1401 the dataset).
 - 1402 (d) We recognize that reproducibility may be tricky in some cases, in which case
 1403 authors are welcome to describe the particular way they provide for reproducibility.
 1404 In the case of closed-source models, it may be that access to the model is limited in
 1405 some way (e.g., to registered users), but it should be possible for other researchers
 1406 to have some path to reproducing or verifying the results.

1407 5. Open access to data and code

1408 Question: Does the paper provide open access to the data and code, with sufficient instruc-
 1409 tions to faithfully reproduce the main experimental results, as described in supplemental
 1410 material?

1411 Answer: [NA]

1412 Justification: NA.

1413 Guidelines:

- 1414 • The answer NA means that paper does not include experiments requiring code.
- 1415 • Please see the NeurIPS code and data submission guidelines ([https://nips.cc/](https://nips.cc/public/guides/CodeSubmissionPolicy)
 1416 [public/guides/CodeSubmissionPolicy](https://nips.cc/public/guides/CodeSubmissionPolicy)) for more details.
- 1417 • While we encourage the release of code and data, we understand that this might not be
 1418 possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not
 1419 including code, unless this is central to the contribution (e.g., for a new open-source
 1420 benchmark).

- 1421 • The instructions should contain the exact command and environment needed to run to
1422 reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
1423
- 1424 • The authors should provide instructions on data access and preparation, including how
1425 to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- 1426 • The authors should provide scripts to reproduce all experimental results for the new
1427 proposed method and baselines. If only a subset of experiments are reproducible, they
1428 should state which ones are omitted from the script and why.
- 1429 • At submission time, to preserve anonymity, the authors should release anonymized
1430 versions (if applicable).
- 1431 • Providing as much information as possible in supplemental material (appended to the
1432 paper) is recommended, but including URLs to data and code is permitted.

1433 6. Experimental setting/details

1434 Question: Does the paper specify all the training and test details (e.g., data splits, hyper-
1435 parameters, how they were chosen, type of optimizer, etc.) necessary to understand the
1436 results?

1437 Answer: [NA]

1438 Justification: NA.

1439 Guidelines:

- 1440 • The answer NA means that the paper does not include experiments.
- 1441 • The experimental setting should be presented in the core of the paper to a level of detail
1442 that is necessary to appreciate the results and make sense of them.
- 1443 • The full details can be provided either with the code, in appendix, or as supplemental
1444 material.

1445 7. Experiment statistical significance

1446 Question: Does the paper report error bars suitably and correctly defined or other appropriate
1447 information about the statistical significance of the experiments?

1448 Answer: [NA]

1449 Justification: NA.

1450 Guidelines:

- 1451 • The answer NA means that the paper does not include experiments.
- 1452 • The authors should answer "Yes" if the results are accompanied by error bars, confi-
1453 dence intervals, or statistical significance tests, at least for the experiments that support
1454 the main claims of the paper.
- 1455 • The factors of variability that the error bars are capturing should be clearly stated (for
1456 example, train/test split, initialization, random drawing of some parameter, or overall
1457 run with given experimental conditions).
- 1458 • The method for calculating the error bars should be explained (closed form formula,
1459 call to a library function, bootstrap, etc.)
- 1460 • The assumptions made should be given (e.g., Normally distributed errors).
- 1461 • It should be clear whether the error bar is the standard deviation or the standard error
1462 of the mean.
- 1463 • It is OK to report 1-sigma error bars, but one should state it. The authors should
1464 preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis
1465 of Normality of errors is not verified.

1466 • For asymmetric distributions, the authors should be careful not to show in tables or
1467 figures symmetric error bars that would yield results that are out of range (e.g. negative
1468 error rates).

1469 • If error bars are reported in tables or plots, The authors should explain in the text how
1470 they were calculated and reference the corresponding figures or tables in the text.

1471 8. Experiments compute resources

1472 Question: For each experiment, does the paper provide sufficient information on the com-
1473 puter resources (type of compute workers, memory, time of execution) needed to reproduce
1474 the experiments?

1475 Answer: [NA]

1476 Justification: NA.

1477 Guidelines:

- 1478 • The answer NA means that the paper does not include experiments.
- 1479 • The paper should indicate the type of compute workers CPU or GPU, internal cluster,
1480 or cloud provider, including relevant memory and storage.
- 1481 • The paper should provide the amount of compute required for each of the individual
1482 experimental runs as well as estimate the total compute.
- 1483 • The paper should disclose whether the full research project required more compute
1484 than the experiments reported in the paper (e.g., preliminary or failed experiments that
1485 didn't make it into the paper).

1486 9. Code of ethics

1487 Question: Does the research conducted in the paper conform, in every respect, with the
1488 NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

1489 Answer: [Yes]

1490 Justification: The research conducted in the paper conform, in every respect, with the
1491 NeurIPS Code of Ethics.

1492 Guidelines:

- 1493 • The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- 1494 • If the authors answer No, they should explain the special circumstances that require a
1495 deviation from the Code of Ethics.
- 1496 • The authors should make sure to preserve anonymity (e.g., if there is a special consid-
1497 eration due to laws or regulations in their jurisdiction).

1498 10. Broader impacts

1499 Question: Does the paper discuss both potential positive societal impacts and negative
1500 societal impacts of the work performed?

1501 Answer: [NA]

1502 Justification: NA.

1503 Guidelines:

- 1504 • The answer NA means that there is no societal impact of the work performed.
- 1505 • If the authors answer NA or No, they should explain why their work has no societal
1506 impact or why the paper does not address societal impact.
- 1507 • Examples of negative societal impacts include potential malicious or unintended uses
1508 (e.g., disinformation, generating fake profiles, surveillance), fairness considerations
1509 (e.g., deployment of technologies that could make decisions that unfairly impact specific
1510 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:

- 1603 • The answer NA means that the paper does not involve crowdsourcing nor research with
1604 human subjects.
- 1605 • Depending on the country in which research is conducted, IRB approval (or equivalent)
1606 may be required for any human subjects research. If you obtained IRB approval, you
1607 should clearly state this in the paper.
- 1608 • We recognize that the procedures for this may vary significantly between institutions
1609 and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the
1610 guidelines for their institution.
- 1611 • For initial submissions, do not include any information that would break anonymity (if
1612 applicable), such as the institution conducting the review.

1613 16. Declaration of LLM usage

1614 Question: Does the paper describe the usage of LLMs if it is an important, original, or
1615 non-standard component of the core methods in this research? Note that if the LLM is used
1616 only for writing, editing, or formatting purposes and does not impact the core methodology,
1617 scientific rigorousness, or originality of the research, declaration is not required.

1618 Answer: [NA]

1619 Justification: Only for editing purposes.

1620 Guidelines:

- 1621 • The answer NA means that the core method development in this research does not
1622 involve LLMs as any important, original, or non-standard components.
- 1623 • Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>)
1624 for what should or should not be described.