Private Geometric Median

Mahdi Haghifam* **Thomas Steinke†** **Jonathan Ullman‡**

Abstract

In this paper, we study differentially private (DP) algorithms for computing the geometric median (GM) of a dataset: Given n points, $x_1, \ldots, x_n$ in $\mathbb{R}^d$, the goal is to find a point θ that minimizes the sum of the Euclidean distances to these points, i.e., $\sum_{i=1}^{n} \|\theta - x_i\|_2$. Off-the-shelf methods, such as DP-GD, require strong a priori knowledge locating the data within a ball of radius R, and the excess risk of the algorithm depends linearly on R. In this paper, we ask: can we design an efficient and private algorithm with an excess error guarantee that scales with the (unknown) radius containing the majority of the datapoints? Our main contribution is a pair of polynomial-time DP algorithms for the task of private GM with an excess error guarantee that scales with the effective diameter of the datapoints. Additionally, we propose an inefficient algorithm based on the inverse smooth sensitivity mechanism, which satisfies the more restrictive notion of pure DP. We complement our results with a lower bound and demonstrate the optimality of our polynomial-time algorithms in terms of sample complexity.

1 Introduction

Differentially private (DP) convex optimization is a fundamental task where we approximately minimize a data-dependent convex loss function while limiting what can be learned about individual data points. The predominant algorithm for DP convex optimization is DP (stochastic) gradient descent, or DP-(S)GD, for short [SCS13; BST14]. Given a dataset $\mathbf{X}^{(n)}$ which contains private information, and a loss function $F(\theta; \mathbf{X}^{(n)})$, DP-(S)GD starts with an initial value $\theta_0 \in \mathbb{R}^d$ and iteratively updates it using $\theta_{t+1} = \Pi_\Theta\big(\theta_t - \eta \cdot \big(\nabla_{\theta_t} F(\theta_t; \mathbf{X}^{(n)}) + \xi_t\big)\big)$ where $\eta > 0$ is the step size, ξ_t is noise to ensure DP, $\Theta \subseteq \mathbb{R}^d$ is a closed convex feasible set, and Π_Θ is the Euclidean projection operator. In the most general setting of Lipschitz convex functions, the excess error depends *linearly* on the radius of the set Θ, and this linear dependence is necessary in the worst-case [BST14]. This linear dependence is problematic because we can think of the diameter of the set Θ as capturing a measure of the uncertainty we have about the location of the minimizer, and we want our algorithm to perform well even with a high degree of uncertainty. This linear dependence can be improved under certain unrealistically strong assumptions, such as strong convexity, but it is unclear whether we can improve the dependence on the radius under weaker, more natural conditions. In this paper, as a step towards answering this question, we identify a simple, optimization task—computing the *geometric median*—where we can exponentially improve the dependence on the radius.

We study private algorithms for computing the *geometric median (GM)* of a dataset: We are given a set of n data points $\mathbf{X}^{(n)} \triangleq (x_1, \ldots, x_n) \in (\mathbb{R}^d)^n$, where x_i represents the private information of

*Khoury College of Computer Sciences, Northeastern University. Supported by a Khoury College of Computer Sciences Distinguished Postdoctoral Fellowship. m.haghifam@northeastern.edu

†Google DeepMind.

‡Khoury College of Computer Sciences, Northeastern University. Supported by NSF awards CNS-2232692 and CNS-2247484.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Algorithm	Privacy	Utility $[F(\mathcal{A}_n(\mathbf{X}^{(n)}); \mathbf{X}^{(n)})]$	Run-time	Samples
`LocDPSGD` (Section 3.1)	Approx	$\left(1 + \frac{\sqrt{d}}{n\varepsilon}\right)\mathsf{OPT}$	$n^2 \log(R/r)$ $+n^2 d$	$\frac{\sqrt{d \log(R/r)}}{\varepsilon}$
`LocDPCuttingPlane` (Section 3.2)	Approx	$\left(1 + \frac{\sqrt{d}}{n\varepsilon}\right)\mathsf{OPT}$	$n^2 \log(R/r)$ $+nd^2 + d^{2+\omega}$	$\frac{\sqrt{d \log(R/r)}}{\varepsilon}$
`SInvS` (Section 4)	Pure	$\left(1 + \frac{d \log(R/r)}{n\varepsilon}\right)\mathsf{OPT}$	Exponential	$\frac{d \log(R/r)}{\varepsilon}$
Baseline: DP-(S)GD	Approx	$\mathsf{OPT} + \frac{R\sqrt{d}}{\varepsilon}$	$n^2 d$	N/A

Table 1: Summary of our results. Here $\mathsf{OPT} = \arg\min_{\theta \in \mathbb{R}^d} F(\theta; \mathbf{X}^{(n)})$ denotes the optimal loss and ω is the matrix-multiplication exponent. The highlighted part is the runtime of the warm-up phase which is the same for `LocDPSGD` and `LocDPCuttingPlane`. We also assume that $\max_{i \in [n]} |\mathbf{X}^{(n)} \cap \mathcal{B}_d(x_i, r)| < 3n/4$. (See Section 3.1, Section 3.2, and Section 4 for the general results without this restriction.) For readability, we omit logarithmic factors that depend on n and d.

one individual, and we are interested in approximately solving the following optimization problem:

$$\theta^\star \triangleq \mathrm{GM}(\mathbf{X}^{(n)}) \in \arg\min_{\theta \in \mathbb{R}^d} F(\theta; \mathbf{X}^{(n)}), \quad \text{where,} \quad F(\theta; \mathbf{X}^{(n)}) \triangleq \sum_{i \in [n]} \|\theta - x_i\|_2. \tag{1}$$

The geometric median generalizes the standard one-dimensional median. The geometric median is a useful tool for robust estimation and aggregation, because it is less sensitive to outliers than the mean of the data, i.e., it is a nontrivial estimator even when $\leq 49\%$ of the input data is arbitrarily corrupted. These properties make GM a popular tool for designing robust versions of distributed optimization methods [CSX17; WLCG20; FGGPS22; AHJSDT22; PKH22; EFGH23], boosting the confidence of weakly concentrated estimators [Min15], clustering [BMM03], etc.

Baseline for Private GM. Since the geometric median is the minimizer of a Lipschitz convex loss function, we can privately approximate it using the standard approach of DP-(S)GD. In particular, if we know a priori that all the data points lie in a known ball of radius R (without loss of generality this ball is centered at the origin, i.e., $\|x_i\|_2 \leq R$ for every $i \in [n]$), then DP-(S)GD guarantees (ε, δ)-DP with the following excess error [BST14]:

$$F(\mathrm{DPGD}_n(\mathbf{X}^{(n)}); \mathbf{X}^{(n)}) - F(\theta^\star; \mathbf{X}^{(n)}) = O\left(\frac{R\sqrt{d \log(1/\delta)}}{\varepsilon}\right). \tag{2}$$

As discussed in the beginning of this section, this guarantee has a significant drawback: the excess error of the algorithm depends *linearly* on the radius R of the a priori bound on the data. This bound could be very loose; it does not scale with the data. Can we do better? What quantity should the excess error guarantee scale with?

It is known that the GM is inside the convex hull of the datapoints. However, this convex hull can have a very large diameter due to a small number of *outliers*, while *most* of the datapoints live in a ball with a small diameter. A key property of GM is robustness to outliers, so we want our accuracy guarantee to also be robust to some outliers. Specifically, if $\geq 51\%$ of the points lie in a ball of diameter $\Delta \ll R$ then the geometric median is $O(\Delta)$ far from that ball (see Lemma C.6 for a more precise statement). Thus, we aim to design a DP algorithm whose error is proportional to the actual scale of the majority of the data, rather than the a priori worst-case bound. However, the algorithm designer does not have a priori knowledge of the location or diameter of a ball that contains most of the data; the algorithm must discover this information from the data. This prompts the following question: *Can we design an efficient and private algorithm with an excess error guarantee that scales with the radius that contains majority of the datapoints?* Our results provide a positive answer.

1.1 Contributions

Our main contribution is a pair of *polynomial-time* DP algorithms for approximating the geometric median with an excess error guarantee that scales with the effective diameter of the datapoints. Also, the sample complexity and the runtime of our algorithms depend logarithmically on the a priori bound R. Both of our algorithms achieve the same excess error bounds up to logarithmic factors, but have incomparable running times. We also give a simple numerical experiment on synthetic data as a proof of concept that our algorithm improves over DP-(S)GD, as presdicted by the theory. In

terms of optimality, we show the our proposed algorithms is optimal in terms of sample complexity. Furthermore, we propose an algorithm based on the *inverse smooth sensitivity* mechanism for the private geometric median problem that satisfies the more restrictive notion of *pure* DP. Below, we give an overview of these algorithms and the techniques involved.

Polynomial-Time Algorithms. Both of our algorithms for the private geometric median problem are two-phase algorithms: in the first phase, which we refer to as *warm-up*, the algorithm shrinks the feasible set to a ball whose diameter is proportional to what we call the *quantile radius* in time that depends logarithmically on R. The second phase, which we call *fine-tuning*, uses the output of the warm-up algorithm to further improve the error.

First, we formalize the notion of the quantile radius as the radius of the smallest ball containing sufficiently many points.

Definition 1.1 (Quantile Radius)**.** Fix a dataset $\mathbf{X}^{(n)} = (x_1, \ldots, x_n) \in (\mathbb{R}^d)^n$ and $\theta \in \mathbb{R}^d$. For every $\gamma \in [0, 1]$, define $\Delta_{\gamma n}(\theta) \triangleq \min\{\Delta : |i \in [n] : \|x_i - \theta\| \leq \Delta| \geq \gamma n\}$.

To motivate the idea behind our algorithms, assume the algorithm designer *knew* a ball, with center θ_0 and radius $\hat{\Delta}$ such that $\|\theta_0 - \theta^\star\| \leq O(\hat{\Delta})$ and $\hat{\Delta} = \widetilde{O}(\Delta_{4n/5}(\theta^\star))$. Then, running DP-(S)GD over this ball would give excess error $O(\Delta_{4n/5}(\theta^\star)\sqrt{d}/\varepsilon)$. This guarantee is particularly interesting as the excess error scales with the quantile radius and not the largest possible norm of any point. Also, by definition of the quantile radius and the geometric median loss function, we have that $F(\theta^\star; \mathbf{X}^{(n)}) \geq (1 - \gamma)n\Delta_{\gamma n}(\theta^\star)$. This inequality shows that an algorithm whose excess error depends on $\Delta_{\gamma n}(\theta^\star)$ has a *multiplicative guarantee* rather than the standard additive guarantee for DP-(S)GD. This type of guarantee is particularly desirable for the geometric median since an algorithm with a multiplicative guarantee will be scale free and be adaptive to the niceness of the dataset. However, since we do not know such a pair θ_0 and $\hat{\Delta}$ a priori, the objective of the warm-up algorithm is to privately find these quantities.

The warm-up algorithm is based on the following structural result: given a point θ that satisfies $\|\theta - \theta^\star\| \gtrsim \Delta_{3n/4}(\theta^\star)$, we have $F(\theta; \mathbf{X}^{(n)}) - F(\theta^\star; \mathbf{X}^{(n)}) \gtrsim \|\theta - \theta^\star\|$. (See Lemma 2.6 for a formal statement.) This result implies that, even though $F(\theta; \mathbf{X}^{(n)})$ is not a strongly convex function, we have a *growth condition* such that the excess error increases with the distance to the global minimizer, at least when the excess error is large enough. (In contrast, strong convexity would imply quadratic growth $F(\theta; \mathbf{X}^{(n)}) - F(\theta^\star; \mathbf{X}^{(n)}) \gtrsim \|\theta - \theta^\star\|^2$, rather than linear growth.) Intuitively, this growth condition allows us to take larger step sizes and make progress faster, consuming less of the privacy budget. However, since this growth condition only holds for θ that is more than $\Delta_{3n/4}(\theta^\star)$ away from the minimizer, which is a data-dependent property, we first need to develop a private algorithm to estimate $\Delta_{3n/4}(\theta^\star)$ in order to make use of this property. In Section 2.1, we develop an efficient algorithm, `RadiusFinder`, for this task, which is inspired by [NSV16]. Our procedure assumes that we know some potentially very small lower bound $r \leq \Delta_{3n/4}(\theta^\star)$, which is necessary by the impossibility results in [BNSV15]. Since the sample complexity of this procedure depends only on $\log(1/r)$, we can choose this parameter to be very small. In Section 2.1, we show how to eliminate this assumption at the cost of a small additive error. With high probability, `RadiusFinder` (see Theorem 2.4) outputs $\hat{\Delta}$ such that $\Delta_{3n/4}(\theta^\star) \leq \hat{\Delta} \leq O(\Delta_{4n/5}(\theta^\star))$. Having obtained $\hat{\Delta}$, the second step of the warm-up algorithm is finding a good initialization point. In Section 2.2, we propose `Localization`, based on DP-GD with *geometrically decaying step sizes*, to perform this task. Due to the growth condition we show that DP-GD makes a fast progress towards some point that is within $O(\Delta_{4n/5}(\theta^\star))$ from the optimizer: in $\log(R)$ iterations, with high probability, it outputs θ_0 such that $\theta^\star$ is in the ball of radius $O(\hat{\Delta}) = O(\Delta_{4n/5}(\theta^\star))$ centered at θ_0.

DP Cutting Plane Method for Private GM. The main drawback of using DP-SGD for the fine-tuning stage is that its run-time can be large when $n \gg d$. To address this, we design the second fine-tuning algorithm, `LocDPCuttingPlane`, based on private variant of the cutting plane method that has faster running time when n is large. There are two challenges in the analysis: by using the noisy gradients, we cannot argue that the optimal point always lives in the intersection of the cutting planes, which is a crucial part of the standard analysis. The second challenge is that the cutting plane method is not a *descent* method in the sense that the loss function is not decreasing with the iteration, and we need to privately select an iterate with small loss. The challenge for developing the private variant here is that the loss $F(\theta; \mathbf{X}^{(n)})$ has sensitivity proportional to R, so running the exponential

mechanism in the natural way incurs loss proportional to R. We address both of these challenges and develop an algorithm whose excess error is proportional to $\Delta_{4n/5}(\theta^\star)$.

Pure DP algorithm for Private Geometric Median Problem. In Section 4, we propose a pure $(\varepsilon, 0)$-DP algorithm for the geometric median problem, albeit a computationally inefficient one. Our algorithm is based on the *inverse smooth sensitivity* mechanism of [AD20]. At a high level, the algorithm outputs $\theta \in \mathbb{R}^d$ with a probability proportional to $\exp(-\varepsilon \cdot \mathrm{len}(\mathbf{X}^{(n)}, \theta)/2)$ where $\mathrm{len}(\mathbf{X}^{(n)}, \theta)$ is the minimum number of data points from $\mathbf{X}^{(n)}$ that needs to be modified to obtain a dataset $\tilde{\mathbf{X}}^{(n)}$ such that the geometric median of $\tilde{\mathbf{X}}^{(n)}$ be equal θ. Our analysis shows that the proposed mechanism outputs $\hat{\theta} = \mathrm{GM}(\tilde{\mathbf{X}}^{(n)})$ such that $\tilde{\mathbf{X}}^{(n)}$ and $\mathbf{X}^{(n)}$ differ in at most $k^\star = O(d \log(R)/\varepsilon)$ with a high probability. Then, by a careful sensitivity analysis, we show $\|\hat{\theta} - \theta^\star\|$ can be upper bounded by the $F(\theta^\star; \mathbf{X}^{(n)})$. Using this result we provide an algorithm with a multiplicative guarantee. Moreover, we show $\|\hat{\theta} - \theta^\star\|$ is upper bounded $O(\Delta_{\gamma n}(\theta^\star))$ for some $\gamma \in (1/2, 1]$.

Lower bound on the Sample Complexity. We show every (ε, δ)-DP algorithm requires $\tilde{\Omega}(\sqrt{d}/\varepsilon)$ samples so that it satisfies $\mathbb{E}_{\hat{\theta} \sim \mathcal{A}_n(\mathbf{X}^{(n)})}[F(\hat{\theta}; \mathbf{X}^{(n)})] \leq (1+\alpha) \min_{\theta \in \mathbb{R}^d} F(\theta, \mathbf{X}^{(n)})$ for a constant α. This result shows that the sample complexity of our polynomial-time algorithms is nearly optimal.

A summary of the results is provided in Table 1, comparing the proposed algorithms in terms of privacy, utility, runtime, and sample complexity. As discussed earlier, algorithms with error adaptive to the quantile radius can achieve a nearly multiplicative guarantee. The utility column in Table 1 compares the algorithms based on the achievable α_{mul} and α_{add} in order to $F(\mathcal{A}_n(\mathbf{X}^{(n)}); \mathbf{X}^{(n)}) \leq (1+\alpha_{\text{mul}}) F(\theta^\star; \mathbf{X}^{(n)}) + \alpha_{\text{add}}$ with a high probability.

1.2 Related Work

DP convex optimization is a well-studied problem [CMS11; KST12; BST14; ACGMMTZ16; STU17; FKT20]. There has been significant interest in developing new algorithms that offer improved guarantees compared to DP-(S)GD for specific problem classes or by leveraging additional information. For instance, [LUZ20; SSTT21; ABGMU22; BMS22] demonstrate that for linear models the dependency of the excess error on the dimension can be improved, [GHST24; ABL23] study the impact of the second-order information on the convergence, [KDRT21; AGMRSSSTT22; GHNOSTTW23] explore the impact of public data, etc. The current paper addresses a drawback of DP-(S)GD, namely, the linear dependence of the excess error on the distance from the initializer to the optimal point in non-strongly convex settings.

Another related line of work to our warm-up strategy is private averaging of [NSV16; NS18; CKMST21; TCKMS22]. The advantage of the algorithm proposed in this work is its simplicity while being optimal in terms of sample complexity: we exploit a structural property of the geometric median and show that running DPGD with the geometrically decaying stepsizes can yield a suitable initialization point without the need for preprocessing steps such as filtering [CKMST21; TCKMS22], coordinate-wise discretization [NSV16], hashing [NS18], etc. The proposed quantile radius can be seen as a robust notion of radius proposed in [BHI02].

In one dimension (i.e., $d = 1$), private versions of the median are well studied [DNPR10; BNS13; BNSV15; DNRR15; BDRS18; ALMM19; KLMNS20; ASSU23; CLNSS23]. In particular, these works improve the dependence on the a priori bound R to $\log^* R$, rather than $\log R$ in our results.

1.3 Notation

Let $d \in \mathbb{N}$. For a vector $x \in \mathbb{R}^d$, $\|x\|$ denotes the ℓ_2 norm of x. We use the following notation for the ball of radius R: $\mathcal{B}_d(a, R) = \{x \in \mathbb{R}^d : \|x - a\| \leq R\}$. Also, $\mathcal{B}_d^\infty(a, R)$ denotes $\{x \in \mathbb{R}^d : \|x - a\|_\infty \leq R\}$. We refer to $\mathcal{B}_d(0, R) = \mathcal{B}_d(R)$, similarly, it holds for $\mathcal{B}_d^\infty(0, R) = \mathcal{B}_d^\infty(R)$. Let $\langle \cdot, \cdot \rangle$ denote the standard inner product in $\mathbb{R}^d$. For a convex and closed subset $\Theta \subseteq \mathbb{R}^d$, let $\Pi_\Theta : \mathbb{R}^d \to \Theta$ be the Euclidean projection operator, given by $\Pi_\Theta(x) = \arg\min_{y \in \Theta} \|y - x\|_2$. For a (measurable) space $\mathcal{R}$, $\mathcal{M}_1(\mathcal{R})$ denotes the set of all probability measures on $\mathcal{R}$. Let $\mathcal{Z}$ be the data space and let $\Theta \subseteq \mathbb{R}^d$ be the parameter space. Let $f : \Theta \times \mathcal{Z} \to \mathbb{R}$ be a loss function. We say f is *L-Lipschitz* iff there exists $L \in \mathbb{R}$ such that $\forall z \in \mathcal{Z}, \forall w, v \in \Theta : |f(w, z) - f(v, z)| \leq L\|w - v\|$.

1.4 Notions of DP

Definition 1.2. Let $\varepsilon > 0$ and $\delta \in [0,1)$. A randomized mechanism $\mathcal{A}_n : \mathcal{Z}^n \to \mathcal{M}_1(\Theta)$ is (ε, δ)-DP, iff, for every neighbouring dataset (i.e., replacement) $\mathbf{X} \in \mathcal{Z}^n$ and $\mathbf{X}' \in \mathcal{Z}^n$, and for every measurable subset $M \subseteq \Theta$, it holds $\mathbb{P}_{\theta \sim \mathcal{A}_n(\mathbf{X})}(\theta \in M) \leq e^{\varepsilon} \cdot \mathbb{P}_{\theta \sim \mathcal{A}_n(\mathbf{X}')}(\theta \in M) + \delta$.

For some of our privacy analysis, we use concentrated differential privacy [DR16; BS16], as it provides a simpler composition theorem – the privacy parameter ρ adds up when we compose.

Definition 1.3 ([BS16, Def. 1.1])**.** A randomized mechanism $\mathcal{A} : \mathcal{Z}^n \to \mathcal{M}_1(\mathcal{R})$ is ρ-zCDP, iff, for every neighbouring dataset (i.e., replacement) $\mathbf{X} \in \mathcal{Z}^n$ and $\mathbf{X}' \in \mathcal{Z}^n$, and for every $\alpha \in (1, \infty)$, it holds $\mathrm{D}_\alpha(\mathcal{A}_n(\mathbf{X}) \| \mathcal{A}_n(\mathbf{X}')) \leq \rho\alpha$, where $\mathrm{D}_\alpha(\mathcal{A}_n(\mathbf{X}) \| \mathcal{A}_n(\mathbf{X}'))$ is the α-Renyi divergence between $\mathcal{A}_n(\mathbf{X})$ and $\mathcal{A}_n(\mathbf{X}')$.

We should think of $\rho \approx \varepsilon^2$: to attain (ε, δ)-DP, it suffices to set $\rho = \frac{\varepsilon^2}{4\log(1/\delta)+4\varepsilon}$ [BS16, Lem. 3.5].

Lemma 1.4 ([BS16, Prop. 1.3])**.** *Assume we have a randomized mechanism $\mathcal{A} : \mathcal{Z} \to \mathcal{M}_1(\mathcal{R})$ that satisfies ρ-zCDP, then for every $\delta > 0$, $\mathcal{A}$ is $(\rho + 2\sqrt{\rho \log(1/\delta)}, \delta)$-DP.*

2 Private Localization

In this section, we present the proposed algorithm for the warm-up stage; it has two steps: *Private Estimation of Quantile Radius* and *Private Localization*.

2.1 Step 1: Private Estimation of Quantile Radius

Algorithm 1 describes our private algorithm `RadiusFinder` for quantile radius estimation.

Algorithm 1 $\texttt{RadiusFinder}_n$

1: Inputs: data set $\mathbf{X}^{(n)} \in (\mathcal{B}_d(R))^n$, fraction $\gamma \in (1/2, 1]$, privacy budget ρ-zCDP, failure probability β, discretization error $0 < r < R$.
2: $m = \lceil \gamma n \rceil$.
3: For every $\nu \geq 0$ and $i \in [n]$, let

$$N_i(\nu) \triangleq |\mathbf{X}^{(n)} \cap \mathcal{B}_d(x_i, \nu)|.$$

▷ Number of datapoints within distance of ν from x_i

4: For every $\nu \geq 0$, define

$$N(\nu) \triangleq \frac{1}{m} \max_{\text{distinct}\{i_1,\ldots,i_m\} \subseteq [n]} \{N_{i_1}(\nu) + \cdots + N_{i_m}(\nu)\}.$$

5: Grid $= \{r, 2r, 4r \cdots, 2^{\lceil \log(\frac{2R}{r}) \rceil} r\}$.
6: Queries $= \{N(v) : v \in \text{Grid}\}$
7:

$$\hat{i} = \texttt{AboveThreshold}\left(\text{Queries}, \rho, m + \frac{18}{\sqrt{2\rho}} \log\left(\frac{2}{\beta} \cdot \left\lceil \log\left(\frac{2R}{r}\right) \right\rceil\right)\right)$$

▷ Algorithm 7

8: Output $\hat{\Delta} = 2^{\hat{i}} r$ if $\hat{i} \neq \texttt{Fail}$; else Output `Fail`.

Remark 2.1. The runtime of `RadiusFinder` is $\Theta((n^2 + n\log(n))\log(\lceil R/r \rceil))$: First, we need to compute the pairwise distances which take n^2 time. Then, for a fixed ν, we can compute $N(\nu)$ using the pairwise distances in time $\Theta(n^2)$. To compute $N(\nu)$, we need to sort $\{N_i(\nu)\}_{i \in [n]}$, in $\Theta(n \log(n))$ time, and pick top m. Finally, we need to repeat this for each $\nu \in [r, \ldots, 2^{\lceil \log(\frac{2R}{r}) \rceil} r]$. ◁

Notice that Algorithm 1 uses the datapoints as centers for computing the number of the datapoints in a given distance. The privacy proof of Algorithm 1 is based on the following lemma.

Lemma 2.2. *Fix $n \in \mathbb{N}$. For every dataset $\mathbf{X}^{(n)}$, for every $1/2 \leq \gamma \leq 1$ and for every fixed ν, the query $N(\nu) \triangleq \frac{1}{m} \max_{\{i_1,\ldots,i_m\} \subseteq [n]} \{N_{i_1}(\nu) + \cdots + N_{i_m}(\nu)\}$, has a sensitivity upper-bounded by 3 where $m = \lceil \gamma n \rceil$ and $N_i(\nu) \triangleq |\mathbf{X}^{(n)} \cap \mathcal{B}_d(x_i, \nu)|$. Here $\mathcal{B}_d(x, \nu) := \{y \in \mathbb{R}^d : \|y - x\| \leq \nu\}$.*

The objective of Algorithm 1 is to privately approximate $\Delta_{\gamma n}(\theta^\star)$. Nonetheless, Algorithm 1 relies on computing the pairwise distances between datapoints. The following lemma elucidates why computing these pairwise distances serves as an effective proxy for computing $\Delta_{\gamma n}(\theta^\star)$.

Lemma 2.3. *Fix $n \in \mathbb{N}$, $1 \leq m \leq n$, $\gamma_1, \gamma_2 \in (1/2, 1]$ such that $\gamma_2 \geq \gamma_1$, and dataset $\mathbf{X}^{(n)}$. For every $\nu \geq 0$, define $N(\nu) \triangleq \frac{1}{m} \max_{\{i_1,\ldots,i_m\}\subseteq[n]}\{N_{i_1}(\nu) + \cdots + N_{i_m}(\nu)\}$, where $N_i(\nu) \triangleq |\mathbf{X}^{(n)} \cap \mathcal{B}_d(x_i, \nu)|$. Let $\theta^\star = \mathrm{GM}(\mathbf{X}^{(n)})$. For every $\hat{\nu}$ such that $N(\hat{\nu}) \geq \lceil \gamma_1 n \rceil$ and $N(\hat{\nu}/2) < \lceil \gamma_2 n \rceil$, we have*

$$\Delta_{\gamma_1 n}(\theta^\star) \cdot \frac{2\gamma_1 - 1}{4\gamma_1 - 1} \leq \hat{\nu} \leq 4\Delta_{\gamma_2 n}(\theta^\star).$$

Using these two lemmas, in the next theorem, we present the privacy and utility guarantees of Algorithm 1. As we are interested in finding the smallest radius, we use the standard `AboveThreshold` from [DNRRV09; DR+14] as a subroutine in Algorithm 1. The algorithmic description of `AboveThreshold` is provided in Appendix B for completeness.

Theorem 2.4. *Let $\texttt{RadiusFinder}_n$ denote Algorithm 1. Fix $d \in \mathbb{N}$, $R > 0$, $r > 0$, $\beta \in (0, 1]$, and $\rho > 0$. Then, for every $n \in \mathbb{N}$ and every dataset $\mathbf{X}^{(n)} \in (\mathcal{B}_d(R))^n$ the output of $\texttt{RadiusFinder}_n$ satisfies ρ-zCDP. Also, the output of $\texttt{RadiusFinder}_n$ satisfies the following utility guarantees:*

1. *Given $n > \frac{18}{(1-\gamma)\sqrt{2\rho}} \log(4/\beta)$, then $\mathbb{P}\Big(\Delta_{\gamma n}(\theta^\star)\frac{2\gamma-1}{4\gamma-1} \leq \hat{\Delta}\Big) \geq 1 - \beta$.*

2. *Assume that the data points satisfies $N(r) < m$. Let $\tilde{\gamma} \triangleq \min\{\gamma + \frac{1}{n}\frac{36}{\sqrt{2\rho}} \log\big(2(\lceil \log(\frac{2R}{r}) \rceil + 1)/\beta\big), 1\}$, then, given $n > \frac{18}{(1-\gamma)\sqrt{2\rho}} \log(4/\beta)$, we have*

$$\mathbb{P}\left(\Delta_{\gamma n}(\theta^\star)\frac{2\gamma - 1}{4\gamma - 1} \leq \hat{\Delta} \leq 4\Delta_{\tilde{\gamma} n}(\theta^\star)\right) \geq 1 - \frac{5}{2}\beta.$$

3. *Let $\tilde{\gamma} \triangleq \min\{\gamma + \frac{1}{n}\frac{36}{\sqrt{2\rho}} \log\big(2(\lceil \log(\frac{2R}{r}) \rceil + 1)/\beta\big), 1\}$. Given $n > \frac{18}{(1-\gamma)\sqrt{2\rho}} \log(4/\beta)$, we have*

$$\mathbb{P}\left(\Delta_{\gamma n}(\theta^\star)\frac{2\gamma - 1}{4\gamma - 1} \leq \hat{\Delta} \text{ and } \left\{\hat{\Delta} \leq 4\Delta_{\tilde{\gamma} n}(\theta^\star) \text{ or } \hat{\Delta} = r\right\}\right) \geq 1 - 2\beta.$$

Remark 2.5. A sufficient condition for $N(r) < m$ in Item 2 is that $\max_{i\in[n]} |\mathbf{X}^{(n)} \cap \mathcal{B}_d(x_i, r)| < m = \lceil \gamma n \rceil$. Intuitively, this means that no data point should have a significant portion of other data points within a ball of radius r centered on it. ◁

2.2 Step 2: Fast Localization

In the second step of the warm-up phase, we develop a fast algorithm for finding a good initialization point using the private estimate of the quantile radius. The main structural result that we use for the algorithm design is stated in the next lemma.

Lemma 2.6. *Fix $n \in \mathbb{N}$, $\mathbf{X}^{(n)} \in (\mathbb{R}^d)^n$ and $\theta_1, \theta_0 \in \mathbb{R}^d$. For every $\gamma \in [0, 1]$, define $\Delta_{\gamma n}(\theta_0) \triangleq \min\{r \geq 0 : |i \in [n] : \|x_i - \theta_0\| \leq r| \geq \gamma n\}$. Assume there exists $\zeta \geq 0$ such that $F(\theta_1; \mathbf{X}^{(n)}) - F(\theta_0; \mathbf{X}^{(n)}) \leq \zeta n$. Then, for every $\gamma \in (1/2, 1]$, we have*

$$(2\gamma - 1)\|\theta_1 - \theta_0\| - 2\gamma\Delta_{\gamma n}(\theta_0) \leq \zeta$$

To gain some intuition behind Lemma 2.6, let us instantiate $\theta_0 = \theta^\star$. This result implies that for a $\theta \in \mathbb{R}^d$ such that $\|\theta - \theta^\star\| \gtrsim \Delta_{\gamma n}(\theta^\star)$, the loss function of the geometric median satisfies $F(\theta; \mathbf{X}^{(n)}) - F(\theta^\star; \mathbf{X}^{(n)}) \gtrsim \|\theta - \theta^\star\|$. Using this result, we propose Algorithm 2 for finding a good initialization. The next theorem states the privacy and utility guarantees of Algorithm 2.

Theorem 2.7. *Let $\texttt{Localization}_n$ denote Algorithm 2. Fix $d \in \mathbb{N}$, $R > 0$, $r > 0$, $\rho > 0$, and $\beta \in (0, 1)$. Then for every dataset $\mathbf{X}^{(n)} \in (\mathcal{B}_d(R))^n$ the outputs of $\texttt{Localization}_n$ satisfies ρ-zCDP. Moreover, let $(\hat{\theta}, \hat{\Delta}) = \texttt{Localization}_n(\mathbf{X}^{(n)}, \rho, r, \beta)$ and define random set $\Theta_{loc} = \{\theta \in \mathcal{B}_d(R) : \left\|\theta - \hat{\theta}\right\| \leq 25\hat{\Delta}\}$. Then, given*

$$n \geq \Omega\left(\max\left\{\frac{\sqrt{d\log(\lceil R/r \rceil)}}{\sqrt{\rho}}\sqrt{\log\left(\frac{\log(\lceil R/r \rceil)}{\beta}\right)}, \frac{1}{\sqrt{\rho}}\log\left(\frac{\lceil R/r \rceil}{\beta}\right)\right\}\right),$$

Algorithm 2 $\texttt{Localization}_n$

Inputs: dataset $\mathbf{X}^{(n)} \in (\mathcal{B}_d(R))^n$, privacy parameters ρ-zCDP, discretization error r, failure probability β
$\gamma = 3/4$
$\hat{\Delta} = \texttt{RadiusFinder}_n\left(\mathbf{X}^{(n)}, \gamma, \frac{\rho}{2}, \frac{\beta}{2}, r\right)$ ▷ Algorithm 1
if $\hat{\Delta} = \texttt{Fail}$ **then**
    Output $\texttt{Fail}$ and Halt.
$k_{\text{wu}} = \frac{1}{\log(2)} \log\left(R/\hat{\Delta}\right)$ ▷ $k_{\text{wu}} \le \frac{1}{\log(2)} \log(R/r)$ with probability one
$\theta_0 = 0 \in \mathbb{R}^d$, $T_{\text{wu}} = 500$, $\text{rad}_0 = R$
for $t \in \{0, \dots, k_{\text{wu}} - 1\}$ **do**
    $\Theta_t = \{\theta \in \mathcal{B}_d(R) : \|\theta - \theta_t\| \le \text{rad}_t\}$
    $\eta_t = \text{rad}_t \sqrt{\frac{2dk_{\text{wu}}}{3\rho n^2}}$
    $\theta_{t+1} = \texttt{DPGD}\left(\theta_t, \mathbf{X}^{(n)}, \frac{\rho}{2k_{\text{wu}}}, \Theta_t, \eta_t, T_{\text{wu}}\right)$ ▷ Algorithm 6
    $\text{rad}_{t+1} = \frac{1}{2}\text{rad}_t + 12\hat{\Delta}$
Output $\theta_{k_{\text{wu}}}$ and $\hat{\Delta}$.

we have $\mathbb{P}\left(\theta^\star \in \Theta_{loc} \text{ and } \Delta_{0.75n}(\theta^\star) \le 4\hat{\Delta} \text{ and } \left\{\hat{\Delta} \le 4\Delta_{0.8n}(\theta^\star) \text{ or } \hat{\Delta} = r\right\}\right) \ge 1 - 2\beta$. *Also, assuming that the datapoints satisfies* $\max_{i\in[n]} |\mathbf{X}^{(n)} \cap \mathcal{B}_d(x_i, r)| < 3n/4$, *we have*

$$\mathbb{P}\left(\theta^\star \in \Theta_{loc} \text{ and } \Delta_{0.75n}(\theta^\star) \le 4\hat{\Delta} \text{ and } \hat{\Delta} \le 4\Delta_{0.8n}(\theta^\star)\right) \ge 1 - 2\beta.$$

3 Private Fine-tuning

In Section 2, we developed an algorithm for the warm-up stage. The output of the warm-up stage is θ_0 and radius $\hat{\Delta}$ such that $\|\theta_0 - \theta^\star\| \le O(\hat{\Delta})$ and $\hat{\Delta} = \widetilde{O}(\Delta_{4n/5}(\theta^\star))$ as formalized in Theorem 2.7. In this section, we build upon the output of the warm-up algorithm to develop two polynomial-time algorithms for the fine-tuning stage.

3.1 Fine-tuning Using DPGD

Our first algorithm is based on DP-GD [BST14]. The main ideas behind Algorithm 3 is as follows: 1) from the utility guarantee of the warm-up phase in Theorem 2.7, the distance of the initialization and $\theta^\star$ only depends on $\hat{\Delta}$, i.e., it does not depend on R, 2) By definition of the quantile radius in Definition 1.1 and Equation (1), we have that $F(\theta^\star; \mathbf{X}^{(n)}) \ge (1-\gamma)n\Delta_{\gamma n}(\theta^\star)$, 3) in the case that the data satisfies some regularity conditions, we have $\hat{\Delta} \le 4\Delta_{0.8n}(\theta^\star)$ from Theorem 2.7. The next theorem summarizes the utility and privacy guarantees of this algorithm.

Algorithm 3 $\texttt{LocDPGD}_n$

Inputs: dataset $\mathbf{X}^{(n)} \in (\mathcal{B}_d(R))^n$, privacy parameters ρ-zCDP, discretization error r, failure probability β.
$\theta_0, \hat{\Delta} = \texttt{Localization}_n\left(\mathbf{X}^{(n)}, \frac{\rho}{2}, r, \frac{\beta}{2}\right)$ ▷ Algorithm 2
$\Theta_0 = \{\theta \in \mathcal{B}_d(R) : \|\theta - \theta_0\| \le 25\hat{\Delta}\}$
$\eta_{\text{ft}} = 50\hat{\Delta}\sqrt{\frac{d}{6\rho n^2}}$ and $T_{\text{ft}} = \frac{n^2\rho}{256d}$
$\hat{\theta} = \texttt{DPGD}\left(\theta_0, \mathbf{X}^{(n)}, \frac{\rho}{2}, \Theta_0, \eta_{\text{ft}}, T_{\text{ft}}\right)$ ▷ Algorithm 6
Output $\hat{\theta}$

Theorem 3.1. *Let* $\texttt{LocalizedDPGD}_n$ *denote Algorithm 3. For every* $d \in \mathbb{N}$, $R > 0$, $r > 0$, $\rho > 0$, *and* $\beta \in (0, 1]$, $\mathcal{A} = \{\texttt{LocalizedDPGD}_n\}_{n\ge1}$ *satisfies the following: for every* $n \in \mathbb{N}$ *and every*

dataset $\mathbf{X}^{(n)} \in (\mathcal{B}_d(R))^n$ *the output of* $\texttt{LocalizedDPGD}_n$ *satisfies* ρ*-zCDP. Also, given*

$$n \geq \Omega\left(\max\left\{\frac{\sqrt{d\log(\lceil R/r\rceil)}}{\sqrt{\rho}}\sqrt{\log\left(\frac{\log(\lceil R/r\rceil)}{\beta}\right)}, \frac{1}{\sqrt{\rho}}\log\left(\frac{\lceil R/r\rceil}{\beta}\right)\right\}\right),$$

we have

$$\mathbb{P}\left(F\left(\hat{\theta};\mathbf{X}^{(n)}\right) \leq \left(1+O\left(\frac{\sqrt{d}}{n\sqrt{\rho}}\sqrt{\log(1/\beta)}\right)\right)F\left(\theta^\star;\mathbf{X}^{(n)}\right) + O\left(\sqrt{\frac{d\log(1/\beta)}{\rho}}\right)r\right) \geq 1-2\beta.$$

Moreover, given that the datapoints satisfies $\max_{i\in[n]}|\mathbf{X}^{(n)}\cap\mathcal{B}_d(x_i,r)| < 3n/4$*, we have*

$$\mathbb{P}\left(F\left(\hat{\theta};\mathbf{X}^{(n)}\right) \leq \left(1+O\left(\frac{\sqrt{d}}{n\sqrt{\rho}}\sqrt{\log(1/\beta)}\right)\right)F\left(\theta^\star;\mathbf{X}^{(n)}\right)\right) \geq 1-2\beta,$$

where $\hat{\theta}$ *is the output of Algorithm 3.*

3.2 Fine-tuning Using Noisy Cutting Plane Method

In this section, we present the second fine-tuning algorithm: `LocDPCuttingPlane` of Algorithm 4. This algorithm is based on the well-known cutting plane method [New65; Lev65; Nes98].

Algorithm 4 $\texttt{LocDPCuttingPlane}_n$

Inputs: dataset $\mathbf{X}^{(n)} \in (\mathcal{B}_d(R))^n$, privacy parameters (ε,δ)-DP, discretization error r, failure probability β
$\rho = \frac{\varepsilon^2}{16\log(2/\delta)+8\varepsilon}$
$\theta_0, \hat{\Delta} = \texttt{Localization}_n\left(\mathbf{X}^{(n)}, \frac{\rho}{2}, r, \min\{\frac{\beta}{3}, \frac{\delta}{2}\}\right)$ ▷ Algorithm 2
$\Theta_0 = \{\theta \in \mathcal{B}_d(R) : \|\theta-\theta_0\| \leq 25\hat{\Delta}\}$
$k_{\text{ft}} = \Theta\left(\frac{d}{\tau}\log\left(\frac{n\sqrt{\tau\cdot\rho}}{\sqrt{d}}+\sqrt{d}\right)\right)$ ▷ See Assumption 1 for definition of τ
for $t \in \{0,\dots,k_{\text{ft}}-1\}$ **do**
  $\theta_t = \texttt{Centre}(\Theta_t)$ ▷ See Assumption 1
  $\xi_{\text{dir},t} \sim \mathcal{N}\left(0, \frac{k_{\text{ft}}}{\rho}\mathbb{I}_d\right)$
  $\Theta_{t+1} = \left\{\theta \in \Theta_t \middle| \left\langle \nabla F(\theta_t;\mathbf{X}^{(n)}) + \xi_{\text{dir},t}, \theta-\theta_t\right\rangle < 0\right\}$
Define Probability Measure: $\pi(t) \propto \exp\left(-\frac{\varepsilon}{448\hat{\Delta}}F\left(\theta_t;\mathbf{X}^{(n)}\right)\right)$ for $t \in \{0,\dots,k_{\text{ft}}-1\}$
Output $\theta_{\hat{t}}$ where $\hat{t} \sim \pi$

Similar to non-private cutting plane method, `LocDPCuttingPlane` is not a descent algorithm. As a result, we need to devise a mechanism for selecting an iterate with minimal loss. In the next lemma, we provide a bespoke analysis of the exponential mechanism with the score function $F(\theta;\mathbf{X}^{(n)})$ defined in Equation (1). Note that the sensitivity of $F(\theta;\mathbf{X}^{(n)})$ is R. However, the next result demonstrates that through a novel analysis of the sensitivity of $F(\theta;\mathbf{X}^{(n)})$, the noise scale due to privacy can be significantly reduced. Proof can be found in Appendix E.

Lemma 3.2. *Let* $\varepsilon \in \mathbb{R}$*,* $k \in \mathbb{N}$*, and* $d \in \mathbb{N}$ *be constants. Let* $\Theta \subseteq \mathbb{R}^d$ *be a set with a bounded diameter of* diam*. Let* $\mathbf{X}^{(n)} \in (\mathbb{R}^d)^n$ *be a dataset and* $\theta^\star \in \mathrm{GM}(\mathbf{X}^{(n)})$*. Let* $\{\theta_1,\dots,\theta_k\} \subseteq \Theta$ *be* k *fixed vectors. Also, assume that* $\theta^\star \in \Theta$*. Let* Δ *be such that* $3\Delta_{3n/4}(\theta^\star) + 2\mathsf{diam} \leq \Delta$*. Consider the following probability measure over* $\{1,\dots,k\}$*:*

$$\pi(i;\mathbf{X}^{(n)}) = \frac{\exp\left(-\frac{\varepsilon}{2\Delta}F(\theta_i;\mathbf{X}^{(n)})\right)}{\sum_{j\in[k]}\exp\left(-\frac{\varepsilon}{2\Delta}F(\theta_j;\mathbf{X}^{(n)})\right)}, \quad i \in [k].$$

1. *Let* $\hat{i} \sim \pi(\cdot;\mathbf{X}^{(n)})$ *and* $OPT \triangleq \min_{i\in[k]}\{F(\theta_i;\mathbf{X}^{(n)}) - F(\theta^\star;\mathbf{X}^{(n)})\}$*. Then, for every* $\beta \in (0,1]$*, we have*

$$\mathbb{P}\left(F(\theta_{\hat{i}};\mathbf{X}^{(n)}) - F(\theta^\star;\mathbf{X}^{(n)}) \leq OPT + \frac{2\Delta}{\varepsilon}\log(k/\beta)\right) \geq 1-\beta.$$

2. *Let $\tilde{\mathbf{X}}^{(n)}$ be a dataset of size n that differs in one sample from $\mathbf{X}^{(n)}$. Then, for every $i \in [k]$, we have*

$$\exp(-\varepsilon)\pi(i;\tilde{\mathbf{X}}^{(n)}) \le \pi(i;\mathbf{X}^{(n)}) \le \exp(\varepsilon)\pi(i;\tilde{\mathbf{X}}^{(n)}).$$

The next theorem provides the privacy guarantee of Algorithm 4. The privacy analysis differs from the rest of the algorithms in the paper. This deviation arises from the fact that for analyzing the privacy guarantee of Line 10 of Algorithm 4, we use Lemma 3.2. Notice that the guarantee in Lemma 3.2 holds provided that Θ_0, defined in Line 4 of Algorithm 4, satisfies $\theta^\star \in \Theta_0$. Ergo, the privacy guarantee of Algorithm 4 only satisfies *approximate-DP*.

Theorem 3.3. *Let* $\texttt{LocDPCuttingPlane}_n$ *denote Algorithm 4. Fix $d \in \mathbb{N}$, $R > 0$, $r > 0$, $\varepsilon > 0$, $\delta \in (0,1]$, and $\beta \in (0,1]$. Then, for every $n \in \mathbb{N}$ and every dataset $\mathbf{X}^{(n)} \in (\mathcal{B}_d(R))^n$ the output of* $\texttt{LocDPCuttingPlane}_n$ *satisfies (ε,δ)-DP.*

We also make the following assumption about the performance of `Centre` subroutine in Algorithm 4.

Assumption 1. *There exists some $\tau \in (0,1]$ such that for all $t \in \{0,\dots,k_{ft}-1\}$, the subroutine of* `Centre` *in Algorithm 4 satisfies $vol(\Theta_{t+1}) \le (1-\tau)vol(\Theta_t)$. Furthermore, the time for calling the routine* `Centre` *is T_c.*

Using the John Ellipsoid [Joh14] as the `Centre` makes τ a dimension independent constant and $T_c = \tilde{O}(d^{1+\omega})$ (by [LSW15]). Now we are ready to state the utility guarantee of Algorithm 4.

Theorem 3.4. *Let* $\texttt{LocDPCuttingPlane}_n$ *denote Algorithm 4. For every $d \in \mathbb{N}$, $R > 0$, $r > 0$, $\varepsilon > 0$, $\delta \in (0,1]$, and $\beta \in (0,1]$, $\mathcal{A} = \{\texttt{LocDPCuttingPlane}_n\}_{n\ge 1}$ satisfies the following: for every $n \in \mathbb{N}$ and every dataset $\mathbf{X}^{(n)} \in (\mathcal{B}_d(R))^n$, given*

$$n \ge \Omega\left(\max\left\{\frac{\sqrt{d\log(\lceil R/r\rceil)}}{\sqrt{\rho}}\sqrt{\log\left(\frac{\log(\lceil R/r\rceil)}{\beta}\right)}, \frac{1}{\sqrt{\rho}}\log\left(\frac{\lceil R/r\rceil}{\beta}\right)\right\}\right),$$

where $\rho = \frac{\varepsilon^2}{16\log(2/\delta)+8\varepsilon}$, we have the following: Let $\kappa \triangleq \frac{n\sqrt{\rho}}{\sqrt{d}} + \sqrt{d}$ and $\alpha = O\left(\sqrt{\frac{d\log(\kappa)}{\tau\rho}\cdot\log\left(\frac{d\log(\kappa)}{\tau\beta}\right)}\right)$. Then,

$$\mathbb{P}\left(F\left(\hat{\theta};\mathbf{X}^{(n)}\right) \le \left(1+\frac{\alpha}{n}\right)F(\theta^\star;\mathbf{X}^{(n)}) + r\alpha\right) \ge 1-3\beta,$$

Moreover, assuming that the datapoints satisfies $\max_{i\in[n]}|\mathbf{X}^{(n)} \cap \mathcal{B}_d(x_i,r)| < 3n/4$, we have

$$\mathbb{P}\left(F\left(\hat{\theta};\mathbf{X}^{(n)}\right) \le \left(1+\frac{\alpha}{n}\right)F(\theta^\star;\mathbf{X}^{(n)})\right) \ge 1-3\beta,$$

where $\hat{\theta}$ is the output of Algorithm 4.

4 Pure-DP Algorithm for Geometric Median

In this section, we propose an algorithm based on the assumption that we have an access to an oracle that outputs an *exact* $\mathrm{GM}(\mathbf{X}^{(n)})$. Before presenting the algorithm, we need a definition: For two sequences of $\boldsymbol{a} = (a_1,\dots,a_n) \in (\mathbb{R}^d)^n$ and $\boldsymbol{b} = (b_1,\dots,b_n) \in (\mathbb{R}^d)^n$, we define the hamming distance as $\mathrm{d_H}(\boldsymbol{a},\boldsymbol{b}) = \sum_{i=1}^n \mathbb{1}[a_i \ne b_i]$. The proposed algorithm is shown in Algorithm 5, and its utility and privacy guarantees are presented in the following theorem.

Theorem 4.1. *Let* $\texttt{SInvS}_n$ *denote the algorithm in Algorithm 5. Fix $d \in \mathbb{N}$, $R > 0$, $r > 0$, and $\varepsilon > 0$. Then, for every $n \in \mathbb{N}$ and every dataset $\mathbf{X}^{(n)} \in (\mathcal{B}_d(R))^n$ the output of* $\texttt{SInvS}_n$ *satisfies ε-DP. Also, for every $\beta \in (0,1)$ and for every $n > 2k^\star \triangleq 2\left\lfloor\frac{2}{\varepsilon}(\log(1/\beta)+d\log(R/r))\right\rfloor$, with probability at least $1-\beta$, we have:*

1. *The value of the cost function satisfies*

$$F(\hat{\theta};\mathbf{X}^{(n)}) \le \left(1+\frac{4k^\star}{n-2k^\star}\right)F(\theta^\star;\mathbf{X}^{(n)}) + nr.$$

Algorithm 5 $\texttt{SInvS}_n$

1: Input: dataset $\mathbf{X}^{(n)} \in (\mathcal{B}_d(R))^n$, privacy parameter ε-DP, discretization error r.
2: For every $y \in \mathcal{B}_d(R)$

$$\mathrm{len}_r(\mathbf{X}, y) \triangleq \min_{\tilde{\mathbf{X}} \in (\mathbb{R}^d)^n} \{\mathrm{d_H}(\mathbf{X}^{(n)}, \tilde{\mathbf{X}}^{(n)}) \text{ such that } \exists z \in \mathcal{B}_d(y, r) \text{ with } \mathrm{GM}(\tilde{\mathbf{X}}^{(n)}) = z\}$$

3: Define density: $d\pi(y) = \dfrac{\exp\left(-\frac{\varepsilon}{2} \cdot \mathrm{len}_r(\mathbf{X}, y)\right)}{\int_{y \in \mathcal{B}_d(R)} \exp\left(-\frac{\varepsilon}{2} \cdot \mathrm{len}_r(\mathbf{X}, y)\right) dy} \mathbb{1}[y \in \mathcal{B}_d(R)]$
4: Output $\hat{\theta} \sim \pi$

2. In terms of distance,

$$\left\|\hat{\theta} - \theta^\star\right\| \le r + \min_{\gamma \in (1/2, 1] : \gamma > \frac{k^\star}{n} + \frac{1}{2}} \frac{\Delta_{\gamma n}(\theta^\star)}{\sqrt{2(\gamma - k^\star/n) - (\gamma - k^\star/n)^2}}.$$

The proof of Theorem 4.1 is provided in Appendix F. The proof is based on showing that the output $\hat{\theta} = \mathrm{GM}\left(\tilde{\mathbf{X}}^{(n)}\right)$ is such that $\tilde{\mathbf{X}}^{(n)}$ and $\mathbf{X}^{(n)}$ differ in at most $k^\star = O(d \log(R)/\varepsilon)$ datapoints with a high probability. Then, we use the properties of the geometric median to show that the sensitivity of GM to changing $k < n/2$ points can be bounded by the value of the optimal loss at $\theta^\star = \mathrm{GM}(\mathbf{X}^{(n)})$.

Lemma 4.2. *For every $n \in \mathbb{N}$ and for every $k < \frac{n}{2}$, and for every $(x_1, \ldots, x_n, y_1, \ldots, y_k) \in (\mathbb{R}^d)^{n+k}$, define $\theta_0 = \mathrm{GM}((x_1, \ldots, x_n))$ and $\theta_k = \mathrm{GM}((x_1, \ldots, x_{n-k}, y_1, \ldots, y_k))$. Then, $\|\theta_k - \theta_0\| \le \dfrac{2}{n - 2k} F(\theta_0; (x_1, \ldots, x_n))$.*

5 Lower Bound on the Sample Complexity

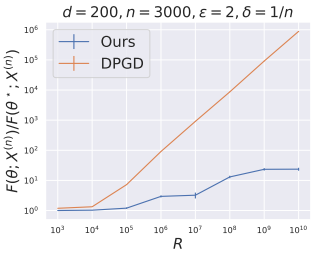
In this section we prove a lower bound on the sample complexity of any (ε, δ)-DP algorithm for the task of private geometric median with a multiplicative error.

Theorem 5.1. *Let $\varepsilon_0, \alpha_0, d_0$ be universal constants. Then, for every $\varepsilon \le \varepsilon_0$, $\alpha \le \alpha_0$, and $d \ge d_0$ and every (ε, δ)-DP algorithm $\mathcal{A}_n : (\mathbb{R}^d)^n \to \mathcal{M}_1(\mathbb{R}^d)$ (with $\delta = \tilde{O}(\sqrt{d}/n)$) such that for every dataset $\mathbf{X}^{(n)} \in (\mathbb{R}^d)^n$ its output satisfies $\mathbb{E}_{\hat{\theta} \sim \mathcal{A}_n(\mathbf{X}^{(n)})}\left[F\left(\hat{\theta}; \mathbf{X}^{(n)}\right)\right] \le (1 + \alpha) \min_{\theta \in \mathcal{B}_d^\infty(1)} F(\theta; \mathbf{X}^{(n)})$, we require $n = \tilde{\Omega}\left(\frac{\sqrt{d}}{\varepsilon}\right)$.*

This result, whose proof can be found in Appendix G, shows that the sample complexity of the proposed polynomial time algorithms is tight in terms of the dependence on ε and d.

6 Numerical Example

In this section, we numerically compare $\texttt{LocDPGD}_n$ (Algorithm 3) and DPGD on a synthetic dataset. The dataset consists of two subsets: one tightly clustered at a random location on $\mathcal{B}_d(R)$, and the other uniformly distributed over $\mathcal{B}_d(R)$. We plot $F(\theta; \mathbf{X}^{(n)})/F(\theta^\star; \mathbf{X}^{(n)})$ for both algorithms as R varies. The results show that $\texttt{LocDPGD}_n$'s performance degrades more gracefully than DP-GD with increasing R. See Appendix H for experimental details and more results.



7 Conclusion and Limitations

In this paper, we presented three private algorithms for the geometric median task, ensuring an excess error guarantee that scales with the effective data scale. Our results open up many directions: we believe our warm-up algorithm has broader applications, and finding other problems where it can be used as a subroutine is interesting. Another direction is to characterize the optimal run-time: is it possible to develop a linear time algorithm, i.e. $\tilde{\Theta}(nd)$, with an optimal excess error?

Acknowledgments

The authors would like to thank Jad Silbak, Eliad Tsfadia, and Mohammad Yaghini for helpful discussions.

References

[ACGMMTZ16] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. "Deep learning with differential privacy". In: *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*. 2016, pp. 308–318.

[AHJSDT22] A. Acharya, A. Hashemi, P. Jain, S. Sanghavi, I. S. Dhillon, and U. Topcu. "Robust training in high dimensions via block coordinate geometric median descent". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2022, pp. 11145–11168.

[ASSU23] M. Aliakbarpour, R. Silver, T. Steinke, and J. Ullman. "Differentially Private Medians and Interior Points for Non-Pathological Data". *arXiv preprint arXiv:2305.13440* (2023).

[ALMM19] N. Alon, R. Livni, M. Malliaris, and S. Moran. "Private PAC learning implies finite Littlestone dimension". In: *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*. 2019, pp. 852–860.

[AGMRSSSTT22] E. Amid, A. Ganesh, R. Mathews, S. Ramaswamy, S. Song, T. Steinke, V. M. Suriyakumar, O. Thakkar, and A. Thakurta. "Public data-assisted mirror descent for private model training". In: *International Conference on Machine Learning*. PMLR. 2022, pp. 517–535.

[ABGMU22] R. Arora, R. Bassily, C. Guzmán, M. Menart, and E. Ullah. "Differentially private generalized linear models revisited". *Advances in Neural Information Processing Systems* 35 (2022), pp. 22505–22517.

[AD20] H. Asi and J. C. Duchi. "Instance-optimality in differential privacy via approximate inverse sensitivity mechanisms". *Advances in neural information processing systems* 33 (2020), pp. 14106–14117.

[ABL23] M. Avella-Medina, C. Bradshaw, and P.-L. Loh. "Differentially private inference via noisy optimization". *The Annals of Statistics* 51.5 (2023), pp. 2067–2092.

[BHI02] M. Bādoiu, S. Har-Peled, and P. Indyk. "Approximate clustering via coresets". In: *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*. 2002, pp. 250–257.

[BMS22] R. Bassily, M. Mohri, and A. T. Suresh. "Differentially private learning with margin guarantees". *Advances in Neural Information Processing Systems* 35 (2022), pp. 32127–32141.

[BST14] R. Bassily, A. Smith, and A. Thakurta. "Private empirical risk minimization: Efficient algorithms and tight error bounds". In: *2014 IEEE 55th annual symposium on foundations of computer science*. IEEE. 2014, pp. 464–473.

[BNS13] A. Beimel, K. Nissim, and U. Stemmer. "Private learning and sanitization: Pure vs. approximate differential privacy". In: *International Workshop on Approximation Algorithms for Combinatorial Optimization*. Springer. 2013, pp. 363–378.

[BMM03] P. Bose, A. Maheshwari, and P. Morin. "Fast approximations for sums of distances, clustering and the Fermat–Weber problem". *Computational Geometry* 24.3 (2003), pp. 135–146.

[BDRS18] M. Bun, C. Dwork, G. N. Rothblum, and T. Steinke. "Composable and versatile privacy via truncated cdp". In: *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*. 2018, pp. 74–86.

[BNSV15] M. Bun, K. Nissim, U. Stemmer, and S. Vadhan. "Differentially private release and learning of threshold functions". In: *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*. IEEE. 2015, pp. 634–649.

[BS16] M. Bun and T. Steinke. "Concentrated differential privacy: Simplifications, extensions, and lower bounds". In: *Theory of Cryptography Conference*. Springer. 2016, pp. 635–658.

[CMS11] K. Chaudhuri, C. Monteleoni, and A. D. Sarwate. "Differentially private empirical risk minimization". *Journal of Machine Learning Research* 12.Mar (2011), pp. 1069–1109.

[CSX17] Y. Chen, L. Su, and J. Xu. "Distributed statistical machine learning in adversarial settings: Byzantine gradient descent". *Proceedings of the ACM on Measurement and Analysis of Computing Systems* 1.2 (2017), pp. 1–25.

[CKMST21] E. Cohen, H. Kaplan, Y. Mansour, U. Stemmer, and E. Tsfadia. "Differentially-private clustering of easy instances". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 2049–2059.

[CLNSS23] E. Cohen, X. Lyu, J. Nelson, T. Sarlós, and U. Stemmer. "Optimal differentially private learning of thresholds and quasi-concave optimization". In: *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*. 2023, pp. 472–482.

[CLMPS16] M. B. Cohen, Y. T. Lee, G. Miller, J. Pachocki, and A. Sidford. "Geometric median in nearly linear time". In: *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*. 2016, pp. 9–21.

[DNPR10] C. Dwork, M. Naor, T. Pitassi, and G. N. Rothblum. "Differential privacy under continual observation". In: *Proceedings of the forty-second ACM symposium on Theory of computing*. 2010, pp. 715–724.

[DNRR15] C. Dwork, M. Naor, O. Reingold, and G. N. Rothblum. "Pure differential privacy for rectangle queries via private partitions". In: *International Conference on the Theory and Application of Cryptology and Information Security*. Springer. 2015, pp. 735–751.

[DNRRV09] C. Dwork, M. Naor, O. Reingold, G. N. Rothblum, and S. Vadhan. "On the complexity of differentially private data release: efficient algorithms and hardness results". In: *Proceedings of the forty-first annual ACM symposium on Theory of computing*. 2009, pp. 381–390.

[DR+14] C. Dwork, A. Roth, et al. "The algorithmic foundations of differential privacy". *Foundations and Trends® in Theoretical Computer Science* 9.3–4 (2014), pp. 211–407.

[DR16] C. Dwork and G. N. Rothblum. "Concentrated differential privacy". *arXiv preprint arXiv:1603.01887* (2016).

[FGGPS22] S. Farhadkhani, R. Guerraoui, N. Gupta, R. Pinot, and J. Stephan. "Byzantine machine learning made easy by resilient averaging of momentums". In: *International Conference on Machine Learning*. PMLR. 2022, pp. 6246–6283.

[FKT20] V. Feldman, T. Koren, and K. Talwar. "Private stochastic convex optimization: optimal rates in linear time". In: *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*. 2020, pp. 439–449.

[GHNOSTTW23] A. Ganesh, M. Haghifam, M. Nasr, S. Oh, T. Steinke, O. Thakkar, A. Thakurta, and L. Wang. "Why is public pretraining necessary for private model training?" In: *International Conference on Machine Learning*. PMLR. 2023, pp. 10611–10627.

[GHST24] A. Ganesh, M. Haghifam, T. Steinke, and A. Thakurta. "Faster differentially private convex optimization via second-order methods". *Advances in Neural Information Processing Systems* 36 (2024).

[Joh14] F. John. "Extremum problems with inequalities as subsidiary conditions". *Traces and emergence of nonlinear programming* (2014), pp. 197–215.

[KDRT21] P. Kairouz, M. R. Diaz, K. Rush, and A. Thakurta. "(Nearly) Dimension Independent Private ERM with AdaGrad Rates
via Publicly Estimated Subspaces". In: *Proceedings of Thirty Fourth Conference on Learning Theory*. Ed. by M. Belkin and S. Kpotufe. Vol. 134. Proceedings of Machine Learning Research. PMLR, 15–19 Aug 2021, pp. 2717–2746.

[KLSU19] G. Kamath, J. Li, V. Singhal, and J. Ullman. "Privately learning high-dimensional distributions". In: *Conference on Learning Theory*. PMLR. 2019, pp. 1853–1902.

[KLMNS20] H. Kaplan, K. Ligett, Y. Mansour, M. Naor, and U. Stemmer. "Privately learning thresholds: Closing the exponential gap". In: *Conference on Learning Theory*. PMLR. 2020, pp. 2263–2285.

[KST12] D. Kifer, A. Smith, and A. Thakurta. "Private convex empirical risk minimization and high-dimensional regression". In: *Conference on Learning Theory*. 2012, pp. 25–1.

[Ksc17] F. R. Kschischang. "The complementary error function". *Online, April* (2017).

[LM00] B. Laurent and P. Massart. "Adaptive estimation of a quadratic functional by model selection". *Annals of Statistics* (2000), pp. 1302–1338.

[LUZ20] H. Lê Nguyen, J. Ullman, and L. Zakynthinou. "Efficient private algorithms for learning large-margin halfspaces". In: *Algorithmic Learning Theory*. PMLR. 2020, pp. 704–724.

[LSW15] Y. T. Lee, A. Sidford, and S. C.-w. Wong. "A faster cutting plane method and its implications for combinatorial and convex optimization". In: *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*. IEEE. 2015, pp. 1049–1065.

[Lev65] A. J. Levin. "An algorithm for minimizing convex functions". *Dokl. Akad. Nauk SSSR* 160 (1965), pp. 1244–1247. ISSN: 0002-3264.

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

[EFGH23] E.-M. El-Mhamdi, S. Farhadkhani, R. Guerraoui, and L.-N. Hoang. "On the strategyproofness of the geometric median". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2023, pp. 2603–2640.

[Min15] S. Minsker. "Geometric median and robust estimation in Banach spaces" (2015).

[Nes98] Y. Nesterov. "Introductory lectures on convex programming volume i: Basic course". *Lecture notes* 3.4 (1998), p. 5.

[New65] D. J. Newman. "Location of the maximum on unimodal surfaces". *Journal of the ACM (JACM)* 12.3 (1965), pp. 395–398.

[NS18] K. Nissim and U. Stemmer. "Clustering algorithms for the centralized and local models". In: *Algorithmic Learning Theory*. PMLR. 2018, pp. 619–653.

[NSV16] K. Nissim, U. Stemmer, and S. Vadhan. "Locating a small cluster privately". In: *Proceedings of the 35th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*. 2016, pp. 413–427.

[PKH22] K. Pillutla, S. M. Kakade, and Z. Harchaoui. "Robust aggregation for federated learning". *IEEE Transactions on Signal Processing* 70 (2022), pp. 1142–1154.

[Sha11] O. Shamir. "A variant of azuma's inequality for martingales with subgaussian tails". *arXiv preprint arXiv:1110.2392* (2011).

[STU17] A. Smith, A. Thakurta, and J. Upadhyay. "Is interaction necessary for distributed private learning?" In: *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE. 2017, pp. 58–77.

[SCS13] S. Song, K. Chaudhuri, and A. D. Sarwate. "Stochastic gradient descent with differentially private updates". In: *2013 IEEE global conference on signal and information processing*. IEEE. 2013, pp. 245–248.

[SSTT21] S. Song, T. Steinke, O. Thakkar, and A. Thakurta. "Evading the curse of dimensionality in unconstrained private glms". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2021, pp. 2638–2646.

[TCKMS22] E. Tsfadia, E. Cohen, H. Kaplan, Y. Mansour, and U. Stemmer. "Friendlycore: Practical differentially private aggregation". In: *International Conference on Machine Learning*. PMLR. 2022, pp. 21828–21863.

[WLCG20] Z. Wu, Q. Ling, T. Chen, and G. B. Giannakis. "Federated variance-reduced stochastic gradient descent with robustness to byzantine attacks". *IEEE Transactions on Signal Processing* 68 (2020), pp. 4583–4596.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and the introduction completely summarize our findings.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: In Section 7, we discussed two limitations of our work.

   Guidelines:

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

3. **Theory Assumptions and Proofs**

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

   Answer: [Yes]

Justification: In our problem setup, we completely discussed all the assumptions. Also, a complete proof of every claim is presented in the appendix.

Guidelines:

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

4. **Experimental Result Reproducibility**

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

   Answer: [Yes]

   Justification: In Appendix H, we discussed all the details behind our implementation. Also, we release the code. Since the dataset considered is synthetic, there is no concern regarding the dataset.

   Guidelines:

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

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

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

Answer: [Yes]

Justification: We have released the code along with a Colab notebook.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

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

Answer: [Yes]

Justification: In our experiments, we compare our proposed algorithm with a well-known baseline. We implemented our algorithm from scratch. Also, all the details are included in the code and Appendix H.

Guidelines:

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

7. **Experiment Statistical Significance**

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

Answer: [Yes]

Justification: We have included the error bars in the plots.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

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

   Answer:[Yes]

   Justification: Our results can be produced using public Google Colab.

   Guidelines:

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

9. **Code Of Ethics**

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

   Answer:[NA]

   Justification: This question is not applicable to our paper.

   Guidelines:

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

10. **Broader Impacts**

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

    Answer: [NA]

    Justification: Our work does not have any societal impact.

    Guidelines:

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

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

11. **Safeguards**

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

    Answer: [NA]

    Justification: It is not applicable to our work.

    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: Not applicable to our work.

    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: It is not applicable to our work.

    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: It is not applicable to our work.

    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: It is not applicable to our work.

    Guidelines:

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

A Preliminaries

A.1 Gradient of the Geometric Loss

$$\nabla_\theta(\|\theta - x\|) = \begin{cases} \frac{\theta - x}{\|\theta - x\|} & \theta \neq x \\ 0 & \theta = x \end{cases}. \tag{3}$$

A.2 DP Gradient Descent (DPGD)

In this section, we provide the algorithmic description of DPGD and its privacy and utility analysis for completeness.

Algorithm 6 DPGD

Inputs: initialization point $\theta_1 \in \mathbb{R}^d$, dataset $\mathbf{X}^{(n)} \in (\mathbb{R}^d)^{(n)}$, privacy budget ρ, feasible set Θ, stepsize η, number of iterations T.
$\sigma^2 = \frac{T}{2\rho n^2}$
for $t \in \{1, \dots, T\}$ **do**

$$\theta_{t+1} = \Pi_\Theta(\theta_t - \eta(\nabla F(\theta_t; \mathbf{X}^{(n)}) + \xi_t)),$$

where $\xi_t \sim \mathcal{N}(0, \sigma^2 I_d)$.
Output $\frac{1}{T}\sum_{t=1}^T \theta_t$

Lemma A.1. *Let $\Theta \subseteq \mathbb{R}^d$ be a closed and convex set with a finite diameter* diam. *Let $\ell : \Theta \times \mathcal{Z} \to \mathbb{R}$ be a loss function such that for every $z \in \mathcal{Z}$, $\ell(\cdot, z)$ is convex and L-Lipschitz. Let $\mathbf{X}^{(n)} = (z_1, \dots, z_n) \in \mathcal{Z}^n$ and $\hat{L}_n(\theta) = \frac{1}{n}\sum_{i\in[n]} \ell(\theta, z_i)$. Consider DP-Gradient descent algorithm $\theta_{t+1} = \Pi_\Theta(\theta_t - \eta(\nabla \hat{L}_n(\theta_t) + \xi_t))$, where $\xi_t \sim \mathcal{N}(0, \sigma^2 I_d)$. Then, for every $T \in \mathbb{N}$, by setting $\eta = \mathsf{diam}\sqrt{\frac{d}{12L^2\rho n^2}}$, and $\sigma^2 = \frac{L^2 T}{2\rho n^2}$, we have the following: $\{\theta_t\}_{t\in[T]}$ satisfies ρ-zCDP. Also, for every $\beta > 0$, given $Td \geq \log(4/\beta)$ and $1 \leq \sqrt{56\log(2/\beta)}$, with probability at least $1 - \beta$, we have*

$$\hat{L}_n\left(\frac{1}{T}\sum_{t\in[T]}\theta_t\right) - \min_{\theta\in\Theta}\hat{L}_n(\theta) \leq L \cdot \mathsf{diam}\left[\frac{16\sqrt{d}}{n\sqrt{\rho}}\sqrt{\log(2/\beta)} + \frac{\sqrt{2}}{\sqrt{T}}\right].$$

Proof. The privacy proof is based on the zCDP analysis of the Gaussian mechanism and the composition property of zCDP [BS16].

Let $g_t \triangleq \nabla \hat{L}_n(\theta_t) + \xi_t$ and $\theta^\star \in \arg\min_{\theta\in\Theta} \hat{L}_n(\theta)$. Note that we can replace $\nabla \hat{L}_n(\theta_t)$ by any subgradient at θ_t. By the convexity of ℓ and the first-order convexity condition we can write

$$\begin{aligned}\hat{L}_n\left(\frac{1}{T}\sum_{t\in[T]}\theta_t\right) - \hat{L}_n(\theta^\star) &\leq \frac{1}{T}\sum_{i\in[T]}\hat{L}_n(\theta_t) - \hat{L}_n(\theta^\star) \\ &\leq \frac{1}{T}\sum_{t\in[T]}\left\langle \nabla\hat{L}_n(\theta_t), \theta_t - \theta^\star\right\rangle.\end{aligned}$$

Then, by the contraction property of the projection, we can write

$$\begin{aligned}\|\theta_{t+1} - \theta^\star\|^2 &= \|\Pi_\Theta(\theta_t - \eta g_t) - \theta^\star\|^2 \\ &\leq \|\theta_t - \theta^\star - \eta g_t\|^2 \\ &= \|\theta_t - \theta^\star\|^2 + \eta^2\|g_t\|^2 - 2\eta\langle g_t, \theta_t - \theta^\star\rangle \\ &\leq \|\theta_t - \theta^\star\|^2 + 2\eta^2\left(\left\|\nabla\hat{L}_n(\theta_t)\right\|^2 + \|\xi_t\|^2\right) - 2\eta\langle g_t, \theta_t - \theta^\star\rangle \\ &\leq \|\theta_t - \theta^\star\|^2 + 2\eta^2 L^2 + 2\eta^2\|\xi_t\|^2 - 2\eta\langle g_t, \theta_t - \theta^\star\rangle.\end{aligned}$$

Here, we have used for every $a, b \in \mathbb{R}^d$, $\|a + b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$, and $\left\|\nabla\hat{L}_n(\theta)\right\| \leq L$ for every θ. Therefore, we conclude that

$$\langle g_t, \theta_t - \theta^\star\rangle \leq \frac{1}{2\eta}\left(\|\theta_t - \theta^\star\|^2 - \|\theta_{t+1} - \theta^\star\|^2\right) + \eta\|\xi_t\|^2 + \eta L^2. \tag{4}$$

Define the following random variable for every $t \in [T]$

$$Y_t = \left\langle \nabla \hat{L}_n(\theta_t), \theta_t - \theta^\star \right\rangle - \langle g_t, \theta_t - \theta^\star \rangle. \tag{5}$$

Also, define the following filtration

$$\mathcal{F}_t = \sigma(\theta_0, \dots, \theta_t), \tag{6}$$

which is the sigma-field generated by $\theta_0, \dots, \theta_t$.

Lemma A.2. *$\{Y_t\}_{t\in[T]}$ is a martingale difference sequence adapted to $\{\mathcal{F}_t\}_{t\in[T]}$.*

Proof. Notice that $\nabla \hat{L}_n(\theta_t)$, θ_t, and $\theta^\star$ are $\mathcal{F}_t$-measurable. Therefore, we can write

$$\begin{aligned}
\mathbb{E}[Y_t|\mathcal{F}_t] &= \mathbb{E}[\langle g_t, \theta_t - \theta^\star \rangle - \left\langle \nabla \hat{L}_n(\theta_t), \theta_t - \theta^\star \right\rangle | \mathcal{F}_t] \\
&= \langle \mathbb{E}[g_t|\mathcal{F}_t], \theta_t - \theta^\star \rangle - \left\langle \nabla \hat{L}_n(\theta_t), \theta_t - \theta^\star \right\rangle.
\end{aligned}$$

By definition ξ_t is independent of the history up to time t. Therefore, $\mathbb{E}[\xi_t|\mathcal{F}_t] = 0$ since $\mathbb{E}[\xi_t] = 0$ which gives

$$\mathbb{E}[g_t|\mathcal{F}_t] = \mathbb{E}[\nabla \hat{L}_n(\theta_t) + \xi_t|\mathcal{F}_t] = \nabla \hat{L}_n(\theta_t), \tag{7}$$

Therefore, $\mathbb{E}[Y_t|\mathcal{F}_t] = 0$. Moreover, by Cuachy-Schwartz inequality and the boundedness of Θ we can write

$$\mathbb{E}[|Y_t|] = \mathbb{E}[|\langle \xi_t, \theta_t - \theta^\star \rangle|] \le \mathbb{E}[\|\xi_t\| \|\theta_t - \theta^\star\|] \le R\mathbb{E}[\|\xi_t\|] < \infty. \tag{8}$$

Therefore, $\{Y_t\}_{t\in[T]}$ is a martingale difference sequence as was to be shown. □

Using Equation (4) and by the definition of Y_t in Equation (5), we can write

$$\left\langle \nabla \hat{L}_n(\theta_t), \theta_t - \theta^\star \right\rangle \le \frac{1}{2\eta}\left(\|\theta_t - \theta^\star\|^2 - \|\theta_{t+1} - \theta^\star\|^2\right) + \eta\|\xi_t\|^2 + \eta L^2 + Y_t.$$

Summing it from 0 to $T-1$ gives

$$\begin{aligned}
\frac{1}{T}\sum_{t\in[T]} \left\langle \nabla \hat{L}_n(\theta_t), \theta_t - \theta^\star \right\rangle &\le \frac{1}{2\eta T}\|\theta_0 - \theta^\star\|^2 + \frac{\eta}{T}\sum_{t\in[T]}\|\xi_t\|^2 + \eta L^2 + \frac{1}{T}\sum_{t\in[T]} Y_t \\
&\le \frac{R^2}{2\eta T} + \eta L^2 + \underbrace{\frac{\eta}{T}\sum_{t\in[T]}\|\xi_t\|^2}_{(A)} + \underbrace{\frac{1}{T}\sum_{t\in[T]} Y_t}_{(B)}.
\end{aligned} \tag{9}$$

Analyzing (A) in Equation (9)

Notice that $\sum_{t\in[T]}\|\xi_t\|^2 \overset{d}{=} \sigma^2\|Y\|^2$. Therefore, for every $\beta \in (0,1)$ provided that $Td \ge \log(4/\beta)$, with probability at least $1 - \beta/2$, we have

$$\frac{\eta}{T}\sum_{t\in[T]}\|\xi_t\|^2 \le \eta\sigma^2 d\left(1 + 4\sqrt{\frac{\log(2/\beta)}{Td}}\right). \tag{10}$$

Analyzing (B) in Equation (9)

Lemma A.3 (Shamir [Sha11])**.** *Let $m \in \mathbb{N}$. Let $\{Z_m\}_{m\in[M]}$ be a martingale difference sequence adapted to a filtration $\{\mathcal{F}_m\}_{m\in[M]}$, and suppose there are constants $b > 1$ and $c > 0$ such that for any m and any $\alpha > 0$, it holds that*

$$\mathbb{P}(|Z_t| \ge \alpha|\mathcal{F}_t) \le b\exp(-c\alpha^2).$$

Then for any $\beta > 0$, it holds with probability at least $1 - \beta$ that

$$\frac{1}{M}\sum_{m\in[M]} Z_m \le \sqrt{\frac{28b\log(1/\beta)}{cM}}.$$

We can rephrase $Equation$ (5) as

$$Y_t = \left\langle \nabla \hat{L}_n(\theta_t), \theta_t - \theta^\star \right\rangle - \langle g_t, \theta_t - \theta^\star \rangle = \langle \xi_t, \theta^\star - \theta_t \rangle.$$

Notice that condition on $\mathcal{F}_t$, $\langle \xi_t, \theta^\star - \theta_t \rangle | \mathcal{F}_t \sim \mathcal{N}(0, \sigma^2 \|\theta_t - \theta^\star\|^2)$. Therefore,

$$\mathbb{P}(|\langle \xi_t, \theta^\star - \theta_t \rangle| \geq \alpha | \mathcal{F}_t) \leq 2 \exp\left(-\frac{\alpha^2}{2\sigma^2 \|\theta_t - \theta^\star\|^2} \right) \leq 2 \exp\left(-\frac{\alpha^2}{2\sigma^2 R^2} \right).$$

Note that the bound holds for every $t \in [T]$. Therefore, using Lemma A.3, with probability at least $1 - \beta/2$, we have

$$\frac{1}{T} \sum_{t \in [T]} \langle \xi_t, \theta^\star - \theta_t \rangle \leq 2\sigma R \sqrt{\frac{28 \log(2/\beta)}{T}}. \tag{11}$$

From Equation (9), Equation (10), and Equation (11), we have with probability at least $1 - \beta$

$$\hat{L}_n\left(\frac{1}{T} \sum_{t \in [T]} \theta_t \right) - \min_{\theta \in \Theta} \hat{L}_n(\theta) \leq \frac{R^2}{2\eta T} + \eta L^2 + \eta \sigma^2 d \left(1 + 4\sqrt{\frac{\log(4/\beta)}{Td}} \right) + 2\sigma R \sqrt{\frac{28 \log(2/\beta)}{T}}, \tag{12}$$

provided that $Td \geq \log(4/\beta)$. Let

$$\sigma^2 = \frac{L^2 T}{2\rho n^2} \quad , \quad \eta = \frac{R}{L\sqrt{T}} \cdot \frac{1}{\sqrt{2 + \frac{5dT}{\rho n^2}}}.$$

Using these parameters, we obtain that

$$\begin{aligned} \hat{L}_n\left(\frac{1}{T} \sum_{t \in [T]} \theta_t \right) - \min_{\theta \in \Theta} \hat{L}_n(\theta) &\leq \frac{RL\sqrt{d}}{n\sqrt{\rho}} \left[\sqrt{1 + \frac{2\rho n^2}{Td}} + \sqrt{56 \log(2/\beta)} \right] \\ &\leq \frac{2RL\sqrt{d}}{n\sqrt{\rho}} \sqrt{56 \log(2/\beta)} + \frac{RL\sqrt{2}}{\sqrt{T}}, \end{aligned} \tag{13}$$

where the last step is by assuming that $1 \leq \sqrt{56 \log(2/\beta)}$.

□

B Above Threshold Algorithm

Algorithm 7 `AboveThreshold`

Inputs: Queries $\{f_0, \ldots, f_{k-1}\}$, Privacy Budget ρ-zCDP, Threshold T.
$\xi_{\text{tresh}} \sim \text{Lap}\left(\frac{6}{\sqrt{2\rho}}\right)$
$\hat{T} = T + \xi_{\text{tresh}}$
for $i \in [k]$ **do**
    $\xi_i \sim \text{Lap}\left(\frac{12}{\sqrt{2\rho}}\right)$
    if $f_i + \xi_i > \hat{T}$: **then**
        Output $\hat{\Delta} = i$.
        Halt
Output `Fail`.

C Technical Lemma

Lemma C.1. *Let $\sigma > 0$. Let Y be a random variable with the distribution $\mathcal{N}(0, \sigma^2)$. Then, for every $\beta \in (0, 1]$, we have $\mathbb{P}\left(|Y| > \sigma\sqrt{2 \log(2/\beta)} \right) \leq \beta$.*

Lemma C.2 (Laurent and Massart [LM00]). *Let $m \in \mathbb{N}$. Consider random vector $Y \sim \mathcal{N}(0, \mathbb{I}_m)$. Then, for every $t \geq 0$,*

$$\mathbb{P}\left(\|Y\|^2 \geq m + 2\sqrt{tm} + 2t \right) \leq \exp(-t)$$

Corollary C.3. *Let $\beta \in (0,1)$, $m \in \mathbb{N}$, and $m \geq \log \frac{2}{\beta}$. Consider $Y \sim \mathcal{N}(0, \mathbb{I}_m)$, then*

$$\mathbb{P}\left(m\left(1 - 2\sqrt{\frac{\log(2/\beta)}{m}}\right) \leq \|Y\|^2 \leq m\left(1 + 4\sqrt{\frac{\log(2/\beta)}{m}}\right) \right) \geq 1 - \beta,$$

Lemma C.4. *Let $n \in \mathbb{N}$ and $n \geq 4$. Let $\mathbf{X}^{(n)} \in (\mathbb{R}^d)^n$ and $\tilde{\mathbf{X}}^{(n)} \in (\mathbb{R}^d)^n$ be two datasets that differ in one sample. Let $\theta^\star \in \mathrm{GM}(\mathbf{X}^{(n)})$ and $\theta^\circledast \in \mathrm{GM}(\tilde{\mathbf{X}}^{(n)})$. Let $\Delta_{3n/4}(\theta^\star)$ be the radius of the ball around $\theta^\star$ that contains at least $3n/4$ of $\mathbf{X}^{(n)}$. Then,*

$$\left\|\theta^\circledast - \theta^\star\right\| \leq \frac{3}{2}\Delta_{3n/4}(\theta^\star).$$

Proof. The proof is by contrapositive. In particular, we show that for every $\theta \in \mathbb{R}^d$ such that $\|\theta - \theta^\star\| > \frac{3}{2}\Delta_{3n/4}(\theta^\star)$, we have, $\theta \notin \mathrm{GM}(\tilde{\mathbf{X}}^{(n)})$. Let $\mathcal{I} = \{i \in [n] : x_i \in \mathcal{B}_d(\theta^\star, \Delta_{3n/4}(\theta^\star))$ and $x_i \in \tilde{\mathbf{X}}^{(n)}\}$. Using the variational representation of $\|\cdot\|_2$, we can write

$$\begin{aligned}
\left\|\nabla F(\theta; \tilde{\mathbf{X}}^{(n)})\right\| &\geq \left\langle \nabla F(\theta; \tilde{\mathbf{X}}^{(n)}), \frac{\theta - \theta^\star}{\|\theta - \theta^\star\|} \right\rangle \\
&= \sum_{i \in \mathcal{I}} \left\langle \frac{\theta - x_i}{\|\theta - x_i\|}, \frac{\theta - \theta^\star}{\|\theta - \theta^\star\|} \right\rangle + \sum_{i \in [n] \setminus \mathcal{I}} \left\langle \frac{\theta - x_i}{\|\theta - x_i\|}, \frac{\theta - \theta^\star}{\|\theta - \theta^\star\|} \right\rangle \\
&\geq \sum_{i \in \mathcal{I}} \left\langle \frac{\theta - x_i}{\|\theta - x_i\|}, \frac{\theta - \theta^\star}{\|\theta - \theta^\star\|} \right\rangle - (n - |\mathcal{I}|)
\end{aligned}$$

where the last step follows from Cauchy–Schwarz inequality. Then, we can write

$$\begin{aligned}
\left\|\nabla F(\theta; \tilde{\mathbf{X}}^{(n)})\right\| &\geq |\mathcal{I}| \sqrt{1 - \left(\frac{\Delta_{3n/4}(\theta^\star)}{\|\theta - \theta^\star\|}\right)^2} - (n - |\mathcal{I}|) \\
&= |\mathcal{I}| \left(1 + \sqrt{1 - \left(\frac{\Delta_{3n/4}(\theta^\star)}{\|\theta - \theta^\star\|}\right)^2}\right) - n \\
&\geq (3n/4) \left(1 + \sqrt{1 - \left(\frac{\Delta_{3n/4}(\theta^\star)}{\|\theta - \theta^\star\|}\right)^2}\right) - n \\
&\geq (3n/4)\left(1 + \sqrt{1 - 4/9}\right) - n \\
&> 0.
\end{aligned} \tag{14}$$

Therefore $\|\theta^\circledast - \theta^\star\| \leq 3/2\Delta_{3n/4}(\theta^\star)$. □

Lemma C.5. *For every $n \in \mathbb{N}$ and for every $\mathbf{X}^{(n)} = (x_1, \ldots, x_n)$, we have $GM(\mathbf{X}^{(n)})$ lies in the convex hull of $\{x_1, \ldots, x_n\}$.*

Lemma C.6. *Let $(x_1, \ldots, x_n) \in (\mathbb{R}^d)^n$ be a dataset and $\theta^\star = \mathrm{GM}((x_1, \ldots, x_n))$. Let $\mathcal{B} \subseteq [n]$ such that $|\mathcal{B}| < n/2$. Then, for every θ, we have*

$$\|\theta - \theta^\star\| \leq \frac{2n - 2|\mathcal{B}|}{n - 2|\mathcal{B}|} \max_{i \notin \mathcal{B}} \|\theta - x_i\|.$$

Proof. It is a simple modification of [CLMPS16, Lemma. 24]. □

D Proof of Section 2

Proof of Lemma 2.2. The proof follows closely [NSV16, Lemma 4.5]. Let $\mathbf{X}$ and $\mathbf{X}'$ are two neighboring datasets of size n that differ in the first sample. For a fixed ν and $i \in [n]$, if $i \neq 1$, $N_i(\nu)$ can change by one. Also, in the worst-case the new datapoint can be close to the rest of the datapoints. Therefore, the sensitivity is bounded by $\frac{1}{m}((m-1)+n) \leq 1 + \frac{n}{m} \leq 1 + \frac{1}{\gamma} \leq 3$ where the last step follows from $\gamma \geq 1/2$. □

Proof of Lemma 2.3. Let $\hat{\nu}$ be such that $N(\hat{\nu}) \geq \lceil \gamma_1 n \rceil$, by definition of $N(\cdot)$ it means that there exists a datapoint x_i such that the ball of radius $\hat{\nu}$ around it contains at least $\lceil \gamma_1 n \rceil$ datapoints. Let $\mathcal{B} = \{j \in [n] : \|x_i - x_j\| > \hat{\nu}\}$. By the described argument, we have $|\mathcal{B}| \leq (1-\gamma_1)n$. Then, we invoke Lemma C.6 with the described $\mathcal{B}$ and $\theta = x_i$ to obtain

$$\begin{aligned}\|\theta^\star - x_i\| &\leq \frac{2n - 2(1-\gamma_1)n}{n - 2(1-\gamma_1)n}\hat{\nu} \\ &= \frac{2\gamma_1}{2\gamma_1 - 1}\hat{\nu}.\end{aligned}$$

The first step follows because function $h : \mathbb{R} \to \mathbb{R}, h(z) = \frac{2n-2z}{n-2z}$ is increasing for $z < n/2$. In the next step, we use the triangle inequality to write

$$\begin{aligned}\Delta_{\gamma_1 n}(\theta^\star) &\leq \Delta_{\gamma_1 n}(x_i) + \|x_i - \theta^\star\| \\ &\leq \hat{\nu} + \frac{2\gamma_1}{2\gamma_1 - 1}\hat{\nu} \\ &= \frac{4\gamma_1 - 1}{2\gamma_1 - 1}\hat{\nu},\end{aligned}$$

where $\Delta_\cdot(\cdot)$ is defined in Definition 1.1.

Next, we turn into proving the upperbound on $\hat{\nu}$. By assumption $N(\hat{\nu}/2) < \lceil \gamma_2 n \rceil$. For the sake of contradiction, assume that $\hat{\nu} > 4\Delta_{\gamma_2 n}(\theta^\star)$. Then, consider the set $\mathcal{G} = \{i \in [n] : \|\theta^\star - x_i\| \leq \Delta_{\gamma_2 n}(\theta^\star)\}$. By definition, $|\mathcal{G}| \geq \lceil \gamma_2 n \rceil$. Consider an arbitrary subset of $\mathcal{G}$ with the size of $\lceil \gamma_2 n \rceil$. The main observation, which follows from the triangle inequality, is that a ball of radius $2\Delta_{\gamma_2 n}(\theta^\star)$ around every point in $\mathcal{G}$ contains at least $\lceil \gamma_2 n \rceil$ datapoint. Therefore, $N(\hat{\nu}/2) \geq N(2\Delta_{\gamma_2 n}(\theta^\star)) \geq \lceil \gamma_2 n \rceil$ which contradicts with the assumption that $N(\hat{\nu}/2) < \lceil \gamma_2 n \rceil$. □

Proof of Theorem 2.4. The privacy proof simply follows from the privacy analysis in [DR+14, Sec. 3.6]. We focus here on the utility guarantees.

Part 1: Let $k = \lceil \log(\frac{2R}{r}) \rceil$. It is simple to see that

$$\begin{aligned}\mathbb{P}\Big(\hat{\Delta} = \texttt{Fail}\Big) &\leq \mathbb{P}\big(N(2^k r) + \xi_k \leq m + \xi_{\text{thresh}}\big) \\ &= \mathbb{P}(n - m \leq \xi_{\text{thresh}} - \xi_k) \\ &\leq \mathbb{P}((1-\gamma)n \leq \xi_{\text{thresh}} - \xi_k)\end{aligned}$$

where the last step follows from the assumption that $\max_{x_i, x_j \in \mathbf{X}^{(n)}} \|x_i - x_j\| \leq 2R$, which gives us $N(2^k r) = n$ by the definition of $N(\cdot)$. By a simple tail bound on the Laplace distribution, we have $\mathbb{P}\Big(|\xi_{\text{thresh}}| \geq \frac{6}{\sqrt{2\rho}}\log(4/\beta)\Big) \leq \beta/4$ and $\mathbb{P}\Big(|\xi_{\text{k}}| \geq \frac{12}{\sqrt{2\rho}}\log(4/\beta)\Big) \leq \beta/4$. Therefore, given $n > \frac{1}{1-\gamma}\frac{18}{\sqrt{2\rho}}\log(4/\beta)$, $\mathbb{P}(n - m \leq \xi_{\text{thresh}} - \xi_k) \leq \beta/2$.

Part 2: Lemma C.6 implies that for every ν such that $N(\nu) \geq \lceil \gamma n \rceil$, we have, $\Delta_{\gamma n}(\theta^\star) \cdot \frac{2\gamma - 1}{4\gamma - 1} \leq \nu$. Therefore, we can write

$$\mathbb{P}\Big(\Delta_{\gamma n}(\theta^\star)\frac{2\gamma-1}{4\gamma-1} \leq \hat{\Delta}\Big) \geq \mathbb{P}\Big(N\Big(\hat{\Delta}\Big) \geq \lceil \gamma n \rceil\Big) \Leftrightarrow \mathbb{P}\Big(\Delta_{\gamma n}(\theta^\star)\frac{2\gamma-1}{4\gamma-1} > \hat{\Delta}\Big) \leq \mathbb{P}\Big(N\Big(\hat{\Delta}\Big) < \lceil \gamma n \rceil\Big).$$

Consider

$$\mathbb{P}\Big(N\Big(\hat{\Delta}\Big) < \lceil \gamma n \rceil\Big) \leq \mathbb{P}\Big(N\Big(\hat{\Delta}\Big) < \lceil \gamma n \rceil \text{ and } \hat{\Delta} \neq \texttt{Fail}\Big) + \mathbb{P}\Big(\hat{\Delta} = \texttt{Fail}\Big). \tag{15}$$

Under the event that $\hat{\Delta} \neq \texttt{Fail}$, there exists $i \in \{0, \ldots, k-1\}$, such that

$$N(\hat{\Delta}) + \xi_i \geq m + \frac{18}{\sqrt{2\rho}} \log(2(k+1)/\beta) + \xi_{\text{thresh}}$$

By a simple tail bound and union bound, we have

$$\begin{aligned} \mathbb{P}(\mathcal{B}) &\triangleq \mathbb{P}\left(|\xi_{\text{thresh}}| \geq \frac{6}{\sqrt{2\rho}} \log(2(k+1)/\beta) \text{ and } \{\max_i |\xi_i| \geq \frac{12}{\sqrt{2\rho}} \log(2(k+1)/\beta)\} \right) \\ &\leq \beta/2, \end{aligned} \tag{16}$$

where $k = \lceil \log(\frac{2R}{r}) \rceil$. We further upperbound the first term in Equation (15) as follows

$$\mathbb{P}\Big(N\Big(\hat{\Delta}\Big) < \lceil \gamma n \rceil \text{ and } \hat{\Delta} \neq \texttt{Fail}\Big) \leq \mathbb{P}\Big(N\Big(\hat{\Delta}\Big) < \lceil \gamma n \rceil \text{ and } \hat{\Delta} \neq \texttt{Fail} \text{ and } \mathcal{B}^c\Big) + \mathbb{P}(\mathcal{B}).$$

We claim that the first term in this equation is zero. Recall that $m = \lceil \gamma n \rceil$. Under the event $\mathcal{B}^c$, $\xi_{\text{thresh}} - \xi_i \geq -\frac{18}{\sqrt{2\rho}} \log(2(k+1)/\beta)$ and as a result, $m + \frac{18}{\sqrt{2\rho}} \log(2(k+1)/\beta) + \xi_{\text{thresh}} - \xi_i \geq m$. Therefore, it shows that the probability of the first term is zero. Also, as showed above, $\mathbb{P}(\mathcal{B}) \leq \beta/2$. Therefore, $\mathbb{P}\Big(N\Big(\hat{\Delta}\Big) < \lceil \gamma n \rceil\Big) \leq \mathbb{P}(\mathcal{B}) + \mathbb{P}\Big(\hat{\Delta} = \texttt{Fail}\Big)$. Combining it with $\mathbb{P}\Big(\hat{\Delta} = \texttt{Fail}\Big) \leq \beta/2$ concludes the proof.

Part 3: Assume that $N(r) < m$. Let $k = \lceil \log(\frac{2R}{r}) \rceil$. Let $\tilde{\gamma} = \gamma + \frac{1}{n} \frac{18}{\sqrt{2\rho}} \log(2(k+1)/\beta)$. In Part 2, we showed that given $n > \frac{18}{\sqrt{2\rho}} \log(4/\beta)$, we have

$$\mathbb{P}\left(\Delta_{\gamma n}(\theta^\star) \frac{2\gamma - 1}{4\gamma - 1} \leq \hat{\Delta} \right) \geq 1 - \beta.$$

We only focus on the upperbound. From Lemma 2.3, we have

$$\mathbb{P}\Big(\hat{\Delta} \leq 4\Delta_{\tilde{\gamma} n}(\theta^\star)\Big) \geq \mathbb{P}\Big(N\Big(\hat{\Delta}/2\Big) \leq \lceil \tilde{\gamma} n \rceil\Big) \Leftrightarrow \mathbb{P}\Big(\hat{\Delta} > 4\Delta_{\tilde{\gamma} n}(\theta^\star)\Big) \leq \mathbb{P}\Big(N\Big(\hat{\Delta}/2\Big) > \lceil \tilde{\gamma} n \rceil\Big). \tag{17}$$

We can write

$$\begin{aligned} \mathbb{P}\Big(N\Big(\hat{\Delta}/2\Big) > \lceil \tilde{\gamma} n \rceil\Big) &\leq \mathbb{P}\Big(N\Big(\hat{\Delta}/2\Big) > \lceil \tilde{\gamma} n \rceil \text{ and } \hat{\Delta} \notin \{r, \texttt{Fail}\}\Big) + \mathbb{P}\Big(\hat{\Delta} \in \{r, \texttt{Fail}\}\Big) \\ &\leq \mathbb{P}\Big(N\Big(\hat{\Delta}/2\Big) > \lceil \tilde{\gamma} n \rceil \text{ and } \hat{\Delta} \notin \{r, \texttt{Fail}\}\Big) + \mathbb{P}\Big(\hat{\Delta} = r\Big) + \mathbb{P}\Big(\hat{\Delta} = \texttt{Fail}\Big), \end{aligned}$$

where the last step follows from the union bound. For the first term, we have

$$\begin{aligned} &\mathbb{P}\Big(N\Big(\hat{\Delta}/2\Big) > \lceil \tilde{\gamma} n \rceil \text{ and } \hat{\Delta} \notin \{r, \texttt{Fail}\}\Big) \\ &= \mathbb{P}\left(N\Big(\hat{\Delta}/2\Big) > \lceil \tilde{\gamma} n \rceil \text{ and } \hat{\Delta} \notin \{r, \texttt{Fail}\} \text{ and } N(\hat{\Delta}/2) + \xi_{\hat{i}} < m + \frac{18}{\sqrt{\rho}} \log\left(\frac{2}{\beta} \cdot \left\lceil \log\left(\frac{2R}{r}\right) \right\rceil\right) + \xi_{\text{thresh}}\right), \end{aligned} \tag{18}$$

where $\hat{i} = \log\Big(\hat{\Delta}/2r\Big) - 1$. The last step follows from the following observation: under the event that $\hat{\Delta} \notin \{r, \texttt{Fail}\}$, during the execution of Algorithm 7, both $N(\hat{\Delta})$ and $N(\hat{\Delta}/2)$ are compared to the noisy threshold. Using the tail bounds in Equation (16), we have under the event $\mathcal{B}^c$, with probability at least $1 - \beta/2$,

$$\begin{aligned} m + \frac{18}{\sqrt{2\rho}} \log\left(\frac{2}{\beta} \cdot \left\lceil \log\left(\frac{2R}{r}\right) \right\rceil\right) + \xi_{\text{thresh}} - \xi_{\hat{i}} &\leq m + \frac{18}{\sqrt{2\rho}} \log\left(\frac{2}{\beta} \cdot \left\lceil \log\left(\frac{2R}{r}\right) \right\rceil\right) + \frac{18}{\sqrt{2\rho}} \log(2(k+1)/\beta) \\ &\leq m + \frac{36}{\sqrt{2\rho}} \log(2(k+1)/\beta). \end{aligned}$$

This shows that we have

$$\mathbb{P}\left(N\Big(\hat{\Delta}/2\Big) > \lceil \tilde{\gamma} n \rceil \text{ and } \hat{\Delta} \notin \{r, \texttt{Fail}\} \text{ and } N(\hat{\Delta}/2) + \xi_{\hat{i}} < m + \frac{18}{\sqrt{\rho}} \log\left(\frac{2}{\beta} \cdot \left\lceil \log\left(\frac{2R}{r}\right) \right\rceil\right) + \xi_{\text{thresh}}\right) \leq \beta/2.$$

In the next step, we bound $\mathbb{P}\Big(\hat{\Delta} = r\Big)$. Notice that

$$\mathbb{P}\Big(\hat{\Delta} = r\Big) = \mathbb{P}\Bigg(N(r) + \xi_1 > m + \frac{18}{\sqrt{2\rho}} \log\Bigg(\frac{2}{\beta} \cdot \Bigg\lceil \log\Big(\frac{2R}{r}\Big)\Bigg\rceil\Bigg) + \xi_{\text{thresh}}\Bigg).$$

Using simple tail bound, we have $\mathbb{P}\Big(\xi_{\text{thresh}} - \xi_1 \leq -\frac{18}{\sqrt{2\rho}} \log\Big(\frac{2}{\beta} \cdot \lceil \log\big(\frac{2R}{r}\big)\rceil\Big)\Big) \leq \beta/2$ which shows that $\mathbb{P}\Big(\hat{\Delta} = r\Big) \leq \beta/2$ since we assume that $N(r) < m$. Therefore, combining all the pieces together, we proved

$$\mathbb{P}\Bigg(\Delta_{\gamma n}(\theta^\star)\frac{2\gamma - 1}{4\gamma - 1} \leq \hat{\Delta} \leq 4\Delta_{\tilde{\gamma} n}(\theta^\star)\Bigg) \geq 1 - \frac{5\beta}{2}.$$

Part 4: In Part 2, we showed that given $n > \frac{18}{(1-\gamma)\sqrt{2\rho}} \log(4/\beta)$, we have

$$\mathbb{P}\Bigg(\hat{\Delta} \leq \Delta_{\gamma n}(\theta^\star)\frac{2\gamma - 1}{4\gamma - 1}\Bigg) \leq \beta. \tag{19}$$

Consider the following event $\mathcal{E} = \Big\{\hat{\Delta} \leq 4\Delta_{\tilde{\gamma} n}(\theta^\star) \text{ or } \hat{\Delta} = r\Big\}$. We have

$$\begin{aligned}
\mathbb{P}(\mathcal{E}^c) &= \mathbb{P}\Big(\hat{\Delta} > 4\Delta_{\tilde{\gamma} n}(\theta^\star) \text{ and } \hat{\Delta} \neq r\Big) \\
&\leq \mathbb{P}\Big(\hat{\Delta} > 4\Delta_{\tilde{\gamma} n}(\theta^\star) \text{ and } \hat{\Delta} \neq \{r, \mathsf{Fail}\}\Big) + \mathbb{P}\Big(\hat{\Delta} = \mathsf{Fail}\Big) \\
&\leq \mathbb{P}\Big(N\Big(\hat{\Delta}/2\Big) > \lceil \tilde{\gamma} n \rceil \text{ and } \hat{\Delta} \neq \{r, \mathsf{Fail}\}\Big) + \mathbb{P}\Big(\hat{\Delta} = \mathsf{Fail}\Big).
\end{aligned}$$

Here, the last step follows from Equation (17). Notice that in Equation (18), we analyzed the probability of the first term and we showed that it is as most $\beta/2$. We also have that $\mathbb{P}\Big(\hat{\Delta} = \mathsf{Fail}\Big) \leq \beta/2$ from Part 1. Therefore, $\mathbb{P}(\mathcal{E}^c) \leq \beta$. Combining it with Equation (19) concludes the proof. □

Proof of Lemma 2.6. Let $\mathcal{I} = \{i \in [n] : \|\theta_0 - x_i\| \leq \Delta_{\gamma n}(\theta_0)\}$. For every $i \in \mathcal{I}$, we have $\|x_i - \theta_0\| \leq \Delta_{\gamma n}(\theta_0)$. Using the triangle inequality, for every $i \in \mathcal{I}$, we can write

$$\begin{aligned}
\|\theta_1 - x_i\| &\geq \|\theta_1 - \theta_0\| - \|\theta_0 - x_i\| \\
&\geq \|\theta_1 - \theta_0\| - (2\Delta_{\gamma n}(\theta_0) - \|\theta_0 - x_i\|).
\end{aligned}$$

The last equation is equivalent to

$$\|\theta_1 - x_i\| - \|\theta_0 - x_i\| \geq \|\theta_1 - \theta_0\| - 2\Delta_{\gamma n}(\theta_0). \tag{20}$$

Then, for every $i \notin \mathcal{I}$, by an application of the triangle inequality

$$\begin{aligned}
&\|\theta_1 - x_i\| + \|\theta_1 - \theta_0\| \geq \|\theta_0 - x_i\| \\
(\Leftrightarrow)&\|\theta_1 - x_i\| - \|\theta_0 - x_i\| \geq -\|\theta_1 - \theta_0\|.
\end{aligned} \tag{21}$$

Then, by adding both sides of Equation (20) and Equation (21), we have

$$F(\theta_1; \mathbf{X}^{(n)}) - F(\theta_0; \mathbf{X}^{(n)}) \geq |\mathcal{I}|\|\theta_1 - \theta_0\| - (n - |\mathcal{I}|)\|\theta_1 - \theta_0\| - 2|\mathcal{I}|\Delta_{\gamma n}(\theta_0).$$

This equation can be represented as

$$\begin{aligned}
\|\theta_1 - \theta_0\| &\leq \frac{F(\theta_1; \mathbf{X}^{(n)}) - F(\theta_0; \mathbf{X}^{(n)}) + 2|\mathcal{I}|\Delta_{\gamma n}(\theta_0)}{2|\mathcal{I}| - n} \\
&\leq \frac{\zeta n + 2|\mathcal{I}|\Delta_{\gamma n}(\theta_0)}{2|\mathcal{I}| - n}.
\end{aligned} \tag{22}$$

Let $\gamma' n = |\mathcal{I}|$. We know that $\gamma' \geq \gamma$. Using this representation we can write

$$\|\theta_1 - \theta_0\| \leq \frac{\zeta + 2\gamma'\Delta_{\gamma n}(\theta_0)}{2\gamma' - 1}.$$

For a fixed $a, b > 0$ define $h(x) \triangleq \frac{a+2xb}{2x-1}$. For $x > 1/2$, $\frac{\mathrm{d}h(x)}{\mathrm{d}x} = -\frac{2(a+b)}{(2x-1)^2}$. This shows that $h(x)$ is decreasing for $x > 1/2$. Therefore using this observation

$$\frac{\zeta + 2\gamma'\Delta_{\gamma n}(\theta_0)}{2\gamma' - 1} \leq \frac{\zeta + 2\gamma\Delta_{\gamma n}(\theta_0)}{2\gamma - 1},$$

as was to be shown. □

Proof of Theorem 2.7. The privacy proof is straightforward. Algorithm 2 uses the data in Line 3 and Line 11. Based on the privacy budget allocation and the composition properties of zCDP, we can show that the output satisfies ρ-zCDP.

For the claim regarding utility, in the first step, consider the recursion in Line 12 of Algorithm 2, i.e., $\mathrm{rad}_{t+1} = \frac{1}{2}\mathrm{rad}_t + 12\hat{\Delta}$ initialized at $\mathrm{rad}_0 = R$. It can be easily shown that $\mathrm{rad}_m = \frac{1}{2^m}\mathrm{rad}_0 + 12\hat{\Delta}\sum_{i=0}^{m-1}(1/2)^i$ for $m \geq 1$. In particular, let $k_{\mathrm{wu}} = \frac{1}{\log(2)}\log\Big(R/\hat{\Delta}\Big)$, then, we obtain that $\mathrm{rad}_{k_{\mathrm{wu}}} \leq 25\hat{\Delta}$.

Let $\gamma = 3/4$ and

$$\tilde{\gamma} = \gamma + \frac{1}{n}\frac{36\sqrt{2}}{\sqrt{2\rho}}\log\left(2\left(\left\lceil\log\left(\frac{2R}{r}\right)\right\rceil + 1\right)\frac{2}{\beta}\right) \leq 0.75 + 0.05 = 0.8,$$

where the last step follows because $n \geq \Omega\Big(\frac{1}{\sqrt{\rho}}\log((\lceil\log(R/r)\rceil + 1)/\beta)\Big)$. Then, define the following event

$$\mathcal{G}_1 = \Big\{\Delta_{0.75n}(\theta^\star) \leq 4\hat{\Delta} \text{ and } \Big\{\hat{\Delta} \leq 4\Delta_{0.8n}(\theta^\star) \text{ or } \hat{\Delta} = r\Big\}\Big\}.$$

In the next step, we analyze the probability that $\theta^\star \in \Theta_{\mathrm{loc}}$. We claim that

$$\mathbb{P}\big(\theta^\star \in \Theta_{\mathrm{loc}}\big|\mathcal{G}_1\big) \geq (1 - \beta/(2k_{\mathrm{wu}}))^{k_{\mathrm{wu}}}.$$

We prove this by induction. In particular, we claim that for every $m \in \{0, \ldots, k_{\mathrm{wu}}\}$ we have $\mathbb{P}\big(\theta^\star \in \Theta_m\big|\mathcal{G}_1\big) \geq (1 - \beta/(2k_{\mathrm{wu}}))^m$. Note that we $\Theta_{\mathrm{loc}} = \Theta_{k_{\mathrm{wu}}}$.

For the base case, by the assumption that the datapoints are in $\mathcal{B}_d(R)$, we have $\mathbb{P}(\theta^\star \in \Theta_0|\mathcal{G}_1) = \mathbb{P}(\theta^\star \in \Theta_0) = 1$ since Θ_0 is trivially independent of every random variable. Then, we show the claim for $m \in \{1, \ldots, k_{\mathrm{wu}}\}$ assuming the claim holds for $m-1$. We can write

$$\begin{aligned}\mathbb{P}\big(\theta^\star \in \Theta_m\big|\mathcal{G}_1\big) &= \mathbb{P}\big(\|\theta^\star - \theta_m\| \leq \mathrm{rad}_m\big|\mathcal{G}_1\big)\\ &\geq \mathbb{P}\big(\|\theta^\star - \theta_m\| \leq \mathrm{rad}_m\big|\theta^\star \in \Theta_{m-1} \text{ and } \mathcal{G}_1\big)\mathbb{P}\big(\theta^\star \in \Theta_{m-1}\big|\mathcal{G}_1\big).\end{aligned} \tag{23}$$

We claim that

$$\mathbb{P}\big(\|\theta^\star - \theta_m\| \leq \mathrm{rad}_m\big|\theta^\star \in \Theta_{m-1} \text{ and } \mathcal{G}_1\big) \geq \mathbb{P}\left(\frac{2\big(F(\theta_m; \mathbf{X}^{(n)}) - F(\theta^\star; \mathbf{X}^{(n)})\big)}{n} \leq \frac{1}{2}\mathrm{rad}_{m-1}\big|\theta^\star \in \Theta_{m-1} \text{ and } \mathcal{G}_1\right).$$

To show this lets instantiate Lemma 2.6 with $\theta_0 = \theta^\star$ and $\gamma = 3/4$ to obtain that for every $\theta \in \mathbb{R}^d$,

$$\|\theta^\star - \theta\| \leq \frac{2\big(F(\theta; \mathbf{X}^{(n)}) - F(\theta^\star; \mathbf{X}^{(n)})\big)}{n} + 3\Delta_{\gamma n}(\theta^\star).$$

Notice that conditioned on $\mathcal{G}_1$, we have $3\Delta_{\gamma n}(\theta^\star) \leq 12\hat{\Delta}$. This shows that $\frac{2\big(F(\theta_m; \mathbf{X}^{(n)}) - F(\theta^\star; \mathbf{X}^{(n)})\big)}{n} \leq \frac{1}{2}\mathrm{rad}_{m-1}$ implies that $\|\theta^\star - \theta_m\| \leq \mathrm{rad}_m$ conditioned on $\mathcal{G}_1$ by the definition of rad_m in Line 12. In the next step, we invoke Lemma A.1. Conditioned on $\theta^\star \in \Theta_{m-1}$ and $\mathcal{G}_1$, with probability at least $1 - \frac{\beta}{2k_{\mathrm{wu}}}$, we have

$$F(\theta_m; \mathbf{X}^{(n)}) - F(\theta^\star; \mathbf{X}^{(n)}) \leq 2\mathrm{rad}_{m-1} \cdot \left[\frac{16\sqrt{2dk_{\mathrm{wu}}}}{n\sqrt{\rho}}\sqrt{\log(4k_{\mathrm{wu}}/\beta)} + \frac{\sqrt{2}}{\sqrt{T_{\mathrm{wu}}}}\right].$$

Notice $k_{\mathrm{wu}} \leq \frac{1}{\log(2)}\log(R/r)$ a.s. By setting $T_{\mathrm{wu}} = 128$ and the bound on the sample size, we have $F(\theta_m; \mathbf{X}^{(n)}) - F(\theta^\star; \mathbf{X}^{(n)}) \leq \frac{\mathrm{rad}_{m-1}}{2}$. Also, notice that the randomness in DPGD is independent of history. Therefore,

$$\mathbb{P}\left(\frac{2\big(F(\theta_m; \mathbf{X}^{(n)}) - F(\theta^\star; \mathbf{X}^{(n)})\big)}{n} \leq \frac{1}{2}\mathrm{rad}_{m-1}\big|\theta^\star \in \Theta_{m-1} \text{ and } \mathcal{G}_1\right) \geq 1 - \frac{\beta}{2k_{\mathrm{wu}}}, \tag{24}$$

Therefore, combining Equations (23) and (24), we obtain

$$\mathbb{P}\big(\theta^\star \in \Theta_m\big|\mathcal{G}_1\big) \geq \left(1 - \frac{\beta}{2k_{\mathrm{wu}}}\right)^m,$$

as was to be shown. From Theorem 2.4, given $n \geq \Omega\Big(\frac{1}{\sqrt{\rho}}\log((\lceil\log(R/r)\rceil+1)/\beta)\Big)$, we have

$$\mathbb{P}(\mathcal{G}_1) \geq 1-\beta.$$

Therefore,

$$\mathbb{P}(\theta^\star \in \Theta_{\text{loc}} \text{ and } \mathcal{G}_1) = \mathbb{P}\big(\theta^\star \in \Theta_{\text{loc}}\big|\mathcal{G}_1\big)\mathbb{P}(\mathcal{G}_1) \geq \left(1-\frac{\beta}{2k_{\text{wu}}}\right)^{k_{\text{wu}}} \cdot (1-\beta) \geq (1-2\beta). \tag{25}$$

This proves the first claim.

Regarding the second claim, define the following event

$$\mathcal{G}_2 = \left\{\Delta_{0.75n}(\theta^\star)\frac{1}{4} \leq \hat{\Delta} \leq 4\Delta_{0.8n}(\theta^\star)\right\}.$$

Notice that in the proof of $\mathbb{P}(\theta^\star \in \Theta_{\text{loc}})$ we only used the fact that with a high probability $\Delta_{\gamma n}(\theta^\star)\frac{1}{4} \leq \hat{\Delta}$. Since $\mathcal{G}_2 \subseteq \mathcal{G}_1$, we can write

$$\mathbb{P}(\theta^\star \in \Theta_{\text{loc}} \text{ and } \mathcal{G}_2) = \mathbb{P}\big(\theta^\star \in \Theta_{\text{loc}}\big|\mathcal{G}_2\big)\mathbb{P}(\mathcal{G}_2) \geq (1-\frac{\beta}{2k_{\text{wu}}})^{k_{\text{wu}}} \cdot (1-5\beta/4) \geq 1-2\beta,$$

where the last step follows from Part 3 of Theorem 2.4 which states that $\mathbb{P}(\mathcal{G}_2) \geq 1-5\beta/4$. □

E Proof of Section 3

Proof of Theorem 3.1. For the privacy proof, notice that Algorithm 3 uses the training set in Line 2 and Line 5. By the privacy budget allocation and the composition properties of zCDP in [BS16], it is immediate to see that the output satisfies ρ-zCDP.

Next, we prove the utility properties. Define the following event

$$\mathcal{G}_1 = \left\{\theta^\star \in \Theta_{\text{loc}} \text{ and } \Delta_{0.75n}(\theta^\star) \leq 4\hat{\Delta} \text{ and } \left\{\hat{\Delta} \leq 4\Delta_{0.8n}(\theta^\star) \text{ or } \hat{\Delta} = r\right\}\right\}.$$

Also, by the non-negativity of $\|\cdot\|_2$, we have

$$F(\theta^\star; \mathbf{X}^{(n)}) = \sum_{i=1}^{n}\|\theta^\star - x_i\| \geq 0.2n\Delta_{0.8n}(\theta^\star).$$

Using this inequality, for every θ, we can write

$$\begin{aligned} &F\Big(\theta; \mathbf{X}^{(n)}\Big) - F\Big(\theta^\star; \mathbf{X}^{(n)}\Big) \leq O\left(\frac{\sqrt{d}}{\sqrt{\rho}}\sqrt{\log(1/\beta)}\right)\Delta_{0.8n}(\theta^\star) \\ &\Rightarrow F\Big(\hat{\theta}; \mathbf{X}^{(n)}\Big) - F\Big(\theta; \mathbf{X}^{(n)}\Big) \leq O\left(\frac{\sqrt{d}}{n\sqrt{\rho}}\sqrt{\log(1/\beta)}\right)F\Big(\theta^\star; \mathbf{X}^{(n)}\Big). \end{aligned} \tag{26}$$

Under the event that $\theta^\star \in \Theta_0$, we can invoke Lemma A.1 to write

$$\mathbb{P}\left(F\Big(\hat{\theta}; \mathbf{X}^{(n)}\Big) - F\Big(\theta^\star; \mathbf{X}^{(n)}\Big) \leq O\left(\frac{\sqrt{d}}{\sqrt{\rho}}\sqrt{\log(1/\beta)}\right)\cdot\hat{\Delta}\,\Big|\,\mathcal{G}_1\right) \geq 1-\beta/2,$$

where it follows because the internal randomness of DPGD is independent of the randomness in Localization step.

By the definition of event $\mathcal{G}_1$, either $\hat{\Delta} = r$ or $\hat{\Delta} \leq 4\Delta_{0.8n}(\theta^\star)$. Note that if $\hat{\Delta} \leq 4\Delta_{0.8n}(\theta^\star)$, we can use Equation (26) to provide a multiplicative guarantee. Therefore, conditioned on the event $\mathcal{G}_1$, we have

$$\begin{aligned} &\mathbb{P}\Big(F\Big(\hat{\theta}; \mathbf{X}^{(n)}\Big) - F\Big(\theta^\star; \mathbf{X}^{(n)}\Big) \leq O\left(\frac{\sqrt{d}}{\sqrt{\rho}}\sqrt{\log(1/\beta)}\right)\cdot r \text{ or} \\ &F\Big(\hat{\theta}; \mathbf{X}^{(n)}\Big) \leq \left(1+O\left(\frac{\sqrt{d}}{n\sqrt{\rho}}\sqrt{\log(1/\beta)}\right)\right)\min_{\theta\in\mathbb{R}^d}F\Big(\theta; \mathbf{X}^{(n)}\Big)\Big|\mathcal{G}_1\Big) \geq 1-\beta/2. \end{aligned} \tag{27}$$

The first statement then follows because, from Theorem 2.7, we have $\mathbb{P}(\mathcal{G}_1) \geq 1 - \beta$.

For the second statement, under the condition that $\max_{i\in[n]}|\mathbf{X}^{(n)} \cap \mathcal{B}_d(x_i, r)| < 3n/4$, we can define the following high probability event:

$$\mathcal{G}_2 = \Big\{\theta^\star \in \Theta_{\text{loc}} \text{ and } \Delta_{0.75n}(\theta^\star) \leq 4\hat{\Delta} \text{ and } \hat{\Delta} \leq 4\Delta_{0.8n}(\theta^\star)\Big\}.$$

The argument then proceeds in the same way as the argument for the first claim.

□

Proof of Lemma 3.2. We can write $\pi(\cdot; \mathbf{X}^{(n)})$ as

$$\pi(i; \mathbf{X}^{(n)}) = \frac{\exp\big(-\frac{\varepsilon}{2\Delta}\big[F(\theta_i; \mathbf{X}^{(n)}) - F(\theta^\star; \mathbf{X}^{(n)})\big]\big)}{\sum_{j\in[k]} \exp\big(-\frac{\varepsilon}{2\Delta}\big[F(\theta_j; \mathbf{X}^{(n)}) - F(\theta^\star; \mathbf{X}^{(n)})\big]\big)}, \quad \forall i \in [k]. \tag{28}$$

It follows because $F(\theta^\star; \mathbf{X}^{(n)})$ is a constant independent of i. Then, the first claim follows from the standard utility analysis of the exponential mechanism in [MT07].

In the next step, we provide the proof for the second claim. To this end, because of Equation (28), we analyze the sensitivity of $F(\theta_i; \mathbf{X}^{(n)}) - F(\theta^\star; \mathbf{X}^{(n)})$ for every $i \in [k]$. Note that $\theta^\star$ is a data dependent quantity. Let $\tilde{\mathbf{X}}^{(n)}$ be a dataset that differ in one sample from $\mathbf{X}^{(n)}$. Also, let $\theta^\circledast \in \text{GM}(\tilde{\mathbf{X}}^{(n)})$ and assume, without loss of generality, that $\tilde{\mathbf{X}}^{(n)} = (x_1, \ldots, x'_n)$ and $\mathbf{X}^{(n)} = (x_1, \ldots, x_n)$. For a fixed $\theta \in \{\theta_1, \ldots, \theta_k\}$, we can write

$$\begin{aligned}
&\Big[F(\theta; \mathbf{X}^{(n)}) - F(\theta^\star; \mathbf{X}^{(n)})\Big] - \Big[F(\theta; \tilde{\mathbf{X}}^{(n)}) - F(\theta^\circledast; \tilde{\mathbf{X}}^{(n)})\Big] \\
&= \|\theta - x_n\| - \|\theta - x'_n\| - \|\theta^\star - x_n\| + \big\|\theta^\circledast - x_{n+1}\big\| + \sum_{i=1}^{n-1}\big(\big\|\theta^\circledast - x_i\big\| - \|\theta^\star - x_i\|\big) \\
&\leq \|\theta - \theta^\star\| + \big\|\theta - \theta^\circledast\big\| + \sum_{i=1}^{n}\big(\big\|\theta^\circledast - x_i\big\| - \|\theta^\star - x_i\|\big) \\
&\leq 2\|\theta - \theta^\star\| + \big\|\theta^\star - \theta^\circledast\big\| + \sum_{i=1}^{n}\big(\big\|\theta^\circledast - x_i\big\| - \|\theta^\star - x_i\|\big).
\end{aligned} \tag{29}$$

Here, the last two steps follow from the triangle inequality. Note that $\theta^\circledast$ is the geometric median of $\tilde{\mathbf{X}}^n$. Therefore by the first-order optimally condition, we have

$$\sum_{i=1}^{n-1} \nabla\big(\big\|\theta^\circledast - x_i\big\|\big) = -\nabla\big(\big\|\theta^\circledast - x'_n\big\|\big). \tag{30}$$

Using the first-order convexity condition applied to the function $h(\theta) = \|\theta - x\|$ for a fixed x, we can write

$$\begin{aligned}
\sum_{i=1}^{n-1}\big(\big\|\theta^\circledast - x_i\big\| - \|\theta^\star - x_i\|\big) &\leq \sum_{i=1}^{n-1}\big\langle\nabla\big(\big\|\theta^\circledast - x_i\big\|\big), \theta^\circledast - \theta^\star\big\rangle \\
&= -\big\langle\nabla\big(\big\|\theta^\circledast - x'_n\big\|\big), \theta^\circledast - \theta^\star\big\rangle \\
&\leq \big\|\theta^\circledast - \theta^\star\big\|,
\end{aligned} \tag{31}$$

where the second step follows from Equation (30) and the last step follows because $\|\nabla(\|\theta^\circledast - x_{n+1}\|)\| \leq 1$. Therefore, using Equation (29) and Equation (31), we have

$$\Big[F(\theta; \mathbf{X}^{(n)}) - F(\theta^\star; \mathbf{X}^{(n)})\Big] - \Big[F(\theta; \tilde{\mathbf{X}}^{(n)}) - F(\theta^\circledast; \tilde{\mathbf{X}}^{(n)})\Big] \leq 2\|\theta - \theta^\star\| + 2\big\|\theta^\star - \theta^\circledast\big\|.$$

In the next step of the proof, we invoke Lemma C.4 to upperbound the sensitivity as follows

$$\begin{aligned}
2\|\theta - \theta^\star\| + 2\big\|\theta^\star - \theta^\circledast\big\| &\leq 2\mathsf{diam} + 3\Delta_{3n/4}(\theta^\star) \\
&\leq \Delta,
\end{aligned}$$

where the last step follows because $\|\theta - \theta^\star\| \leq \mathsf{diam}$ by the assumption. Notice that the sensitivity analysis in the reverse direction is also the same. Therefore, the second claim follows from the standard analysis of the privacy of the exponential mechanism. □

Proof of Theorem 3.3. The privacy proof of Algorithm 4 is relatively non-standard. Let $\mathcal{A}_1\left(\mathbf{X}^{(n)}\right) = \left(\hat{\Delta}, \{\theta_i\}_{i\in\{0,\dots,k_{\text{ft}}-1\}}\right)$. Also, let $\mathcal{A}_2\left(\mathbf{X}^{(n)}; \left(\hat{\Delta}, \{\theta_i\}_{i\in\{0,\dots,k_{\text{ft}}-1\}}\right)\right) = \hat{t}$. In particular, $\mathcal{A}_1\left(\mathbf{X}^{(n)}\right)$ can be viewed as the first part of Algorithm 4 before Line 10. Also, $\mathcal{A}_2(\cdot;\cdot)$ denotes the exponential mechanism in Line 10 of Algorithm 4. Using the conversion between zCDP and DP, the privacy budget allocation, and the composition properties of zCDP, we have that $\mathcal{A}_1(\cdot)$ satisfies $(\varepsilon/2, \delta/2)$-DP.

Define the following event

$$\mathcal{G} = \{\theta^\star \in \Theta_0 \text{ and } \Delta_{0.75n}(\theta^\star) \le 4\hat{\Delta}\}.$$

Let μ be a measure on $\mathcal{M}_1(\mathbb{R} \times (\mathbb{R}^d)^{k_{\text{ft}}})$ that satisfies the following: for every dataset $\mathbf{X}^{(n)}$, $\mathbb{P}(\mathcal{A}_1\left(\mathbf{X}^{(n)}\right) \in \cdot) \ll \mu(\cdot)$. Let P_1 denote the density. Since $\mathcal{A}_1$ satisfies approximate-DP, we assume for every $z \in \mathbb{R} \times (\mathbb{R}^d)^{k_{\text{ft}}}$, we have $P_1(z; \mathbf{X}^{(n)}) \le \exp(\varepsilon/2) P_1(z; \tilde{\mathbf{X}}^{(n)}) + \delta/2$.

To prove the requirement of privacy, let $\mathcal{S} \subseteq \mathbb{R} \times (\mathbb{R}^d)^{k_{\text{ft}}} \times \{0, \dots, k_{\text{ft}} - 1\}$. Also, let $\tilde{\mathbf{X}}^{(n)}$ be a dataset of size n that differs in one sample from $\mathbf{X}^{(n)}$. Then, we can write

$$\begin{aligned}
&\mathbb{P}_{\mathcal{A}_1(\mathbf{X}^{(n)}), \mathcal{A}_2(\cdot;\mathbf{X}^{(n)})}\left(\left(\hat{\Delta}, \{\theta_i\}_{i\in\{0,\dots,k_{\text{ft}}-1\}}, \hat{t}\right) \in \mathcal{S}\right) \\
&= \sum_{\hat{t}} \int \mathbb{1}[\mathcal{S}] P_1\left(\hat{\Delta}, \{\theta_i\}_{i\in\{0,\dots,k_{\text{ft}}-1\}} | \mathbf{X}^{(n)}\right) \cdot \pi\left(\hat{t} | \hat{\Delta}, \{\theta_i\}_{i\in\{0,\dots,k_{\text{ft}}-1\}}, \mathbf{X}^{(n)}\right) d\mu \\
&= \sum_{\hat{t}} \int \mathbb{1}[\mathcal{S}] P_1\left(\hat{\Delta}, \{\theta_i\}_{i\in\{0,\dots,k_{\text{ft}}-1\}} | \mathbf{X}^{(n)}\right) \cdot \pi\left(\hat{t} | \hat{\Delta}, \{\theta_i\}_{i\in\{0,\dots,k_{\text{ft}}-1\}}, \mathbf{X}^{(n)}\right) \mathbb{1}\left[(\hat{\Delta}, \theta_0) \in \mathcal{G}\right] d\mu \\
&+ \sum_{\hat{t}} \int \mathbb{1}[\mathcal{S}] P_1\left(\hat{\Delta}, \{\theta_i\}_{i\in\{0,\dots,k_{\text{ft}}-1\}} | \mathbf{X}^{(n)}\right) \cdot \pi\left(\hat{t} | \hat{\Delta}, \{\theta_i\}_{i\in\{0,\dots,k_{\text{ft}}-1\}}, \mathbf{X}^{(n)}\right) \mathbb{1}\left[(\hat{\Delta}, \theta_0) \in \mathcal{G}^c\right] d\mu \\
&= \sum_{\hat{t}} \int \mathbb{1}[\mathcal{S}] P_1\left(\hat{\Delta}, \{\theta_i\}_{i\in\{0,\dots,k_{\text{ft}}-1\}} | \mathbf{X}^{(n)}\right) \cdot \pi\left(\hat{t} | \hat{\Delta}, \{\theta_i\}_{i\in\{0,\dots,k_{\text{ft}}-1\}}, \mathbf{X}^{(n)}\right) \mathbb{1}\left[(\hat{\Delta}, \theta_0) \in \mathcal{G}\right] d\mu + \mathbb{P}(\mathcal{G}^c).
\end{aligned}$$

Notice that under the event that $(\hat{\Theta}, \theta_0) \in \mathcal{G}$, we can invoke Lemma 3.2 to reason about the privacy properties of $\mathcal{A}_2$. Under the event $\mathcal{G}$, we can see that $3\Delta_{3n/4}(\theta^\star) + 2\text{diam}(\Theta_0) \le 112\hat{\Delta}$. Therefore, by Lemma 3.2, we have

$$\pi\left(\hat{t} | \hat{\Delta}, \{\theta_i\}_{i\in\{0,\dots,k_{\text{ft}}-1\}}, \mathbf{X}^{(n)}\right) \le \exp(\varepsilon/2) \cdot \pi\left(\hat{t} | \hat{\Delta}, \{\theta_i\}_{i\in\{0,\dots,k_{\text{ft}}-1\}}, \tilde{\mathbf{X}}^{(n)}\right).$$

Moreover, by Theorem 2.7, we have $\mathbb{P}(\mathcal{G}^c) \le \delta/2$. Therefore,

$$\begin{aligned}
&\sum_{\hat{t}} \int \mathbb{1}[\mathcal{S}] P_1\left(\hat{\Delta}, \{\theta_i\}_{i\in\{0,\dots,k_{\text{ft}}-1\}} | \mathbf{X}^{(n)}\right) \cdot \pi\left(\hat{t} | \hat{\Delta}, \{\theta_i\}_{i\in\{0,\dots,k_{\text{ft}}-1\}}, \mathbf{X}^{(n)}\right) d\mu \\
&\le \exp(\varepsilon/2) \sum_{\hat{t}} \int \mathbb{1}[\mathcal{S}] P_1\left(\hat{\Delta}, \{\theta_i\}_{i\in\{0,\dots,k_{\text{ft}}-1\}} | \mathbf{X}^{(n)}\right) \cdot \pi\left(\hat{t} | \hat{\Delta}, \{\theta_i\}_{i\in\{0,\dots,k_{\text{ft}}-1\}}, \tilde{\mathbf{X}}^{(n)}\right) d\mu + \delta/2.
\end{aligned}$$

Then, we use the fact that $\mathcal{A}_1\left(\mathbf{X}^{(n)}\right)$ satisfies $(\varepsilon/2, \delta/2)$:

$$\begin{aligned}
&\exp(\varepsilon/2) \sum_{\hat{t}} \int \mathbb{1}[\mathcal{S}] P_1\left(\hat{\Delta}, \{\theta_i\}_{i\in\{0,\dots,k_{\text{ft}}-1\}} | \mathbf{X}^{(n)}\right) \cdot \pi\left(\hat{t} | \hat{\Delta}, \{\theta_i\}_{i\in\{0,\dots,k_{\text{ft}}-1\}}, \tilde{\mathbf{X}}^{(n)}\right) d\mu + \delta/2 \\
&\le \exp(\varepsilon) \sum_{\hat{t}} \int \mathbb{1}[\mathcal{S}] P_1\left(\hat{\Delta}, \{\theta_i\}_{i\in\{0,\dots,k_{\text{ft}}-1\}} | \tilde{\mathbf{X}}^{(n)}\right) \cdot \pi\left(\hat{t} | \hat{\Delta}, \{\theta_i\}_{i\in\{0,\dots,k_{\text{ft}}-1s\}}, \tilde{\mathbf{X}}^{(n)}\right) d\mu + \delta.
\end{aligned}$$

It concludes the proof. □

Proof of Theorem 3.4. Define the following event

$$\mathcal{G}_1 = \left\{\theta^\star \in \Theta_{\text{loc}} \text{ and } \Delta_{0.75n}(\theta^\star) \le 4\hat{\Delta} \text{ and } \left\{\hat{\Delta} \le 4\Delta_{0.8n}(\theta^\star) \text{ or } \hat{\Delta} = r\right\}\right\}. \tag{32}$$

By Theorem 2.7, and the assumption on the minimum number of samples, we have $\mathbb{P}(\mathcal{G}) \ge 1 - \beta$.

By applying the standard tail bound on the norm of a Gaussian random vector (as outlined in Corollary C.3), we have

$$\begin{aligned}\mathbb{P}(\mathcal{G}_{n,1}) &\triangleq \mathbb{P}\left(\forall t \in \{0,\dots,k_{\text{ft}}-1\} : \|\xi_{\text{dir},t}\|^2 \leq \frac{3dk_{\text{ft}}}{2\rho}\left(1+4\sqrt{\frac{\log(10k_{\text{ft}}/\beta)}{d}}\right)\right) \\ &\geq 1-\beta/5.\end{aligned} \tag{33}$$

Also, we can write

$$\begin{aligned}&\mathbb{P}\left(\exists t \in \{0,\dots,k_{\text{ft}}-1\} : \langle \xi_{\text{dir},t}, \theta^\star - \theta_t\rangle > 50\hat{\Delta}\cdot\sqrt{\frac{3k_{\text{ft}}}{\rho}\log(10k_{\text{ft}}/\beta)}\right) \\ &\leq \mathbb{P}\left(\exists t \in \{0,\dots,k_{\text{ft}}-1\} : \langle \xi_{\text{dir},t}, \theta^\star - \theta_t\rangle > 50\hat{\Delta}\cdot\sqrt{\frac{3k_{\text{ft}}}{\rho}\log(10k_{\text{ft}}/\beta)}\Big|\mathcal{G}_1\right) + \mathbb{P}(\mathcal{G}_1^c) \\ &\leq \mathbb{P}\left(\exists t \in \{0,\dots,k_{\text{ft}}-1\} : \langle \xi_{\text{dir},t}, \theta^\star - \theta_t\rangle > 50\hat{\Delta}\cdot\sqrt{\frac{3k_{\text{ft}}}{\rho}\log(10k_{\text{ft}}/\beta)}\Big|\mathcal{G}_1\right) + \beta,\end{aligned}$$

where the last step follows because $\mathbb{P}(\mathcal{G}_1^c) \leq \beta$. Conditioned on $\mathcal{G}_1$, for all $t \in \{0,\dots,k_{\text{ft}-1}\}$, we have $\|\theta_t - \theta^\star\| \leq 50\hat{\Delta}$ since $\theta^\star \in \Theta_0$, $\theta_t \in \Theta_0$, and the diameter of Θ_0 is $50\hat{\Delta}$. Also, notice that $\xi_{\text{dir},t} \perp\!\!\!\perp (\theta_t, \Theta_0)$. Using these observations, conditioned on the event $\mathcal{G}_1$, using the standard tail bound on Gaussian random variable (as outlined in Lemma C.1), we can write

$$\mathbb{P}\left(\exists t \in \{0,\dots,k_{\text{ft}}-1\} : \langle \xi_{\text{dir},t}, \theta^\star - \theta_t\rangle > 50\hat{\Delta}\cdot\sqrt{\frac{3k_{\text{ft}}}{\rho}\log(10k_{\text{ft}}/\beta)}\Big|\mathcal{G}_1\right) \leq \beta/5.$$

Therefore, we conclude

$$\begin{aligned}\mathbb{P}(\mathcal{G}_{n,2}) &\triangleq \mathbb{P}\left(\forall t \in \{0,\dots,k_{\text{ft}}-1\} : \langle \xi_{\text{dir},t}, \theta^\star - \theta_t\rangle \leq 50\hat{\Delta}\cdot\sqrt{\frac{3k_{\text{ft}}}{\rho}\log(10k_{\text{ft}}/\beta)}\right) \\ &\geq 1-6\beta/5.\end{aligned} \tag{34}$$

To prove the claim regarding the suboptimality gap, we consider two cases:

1. There exists $t \in \{0,\dots,k_{\text{ft}}-1\}$ such that $\theta^\star \in \Theta_t$ and $\theta^\star \notin \Theta_{t+1}$,
2. $\theta^\star \in \Theta_{k_{\text{ft}}}$.

Note that these two events are mutually exclusive and their union covers all the space. In what follows, we show that in both cases there exists $t \in \{0,\dots,k_{\text{ft}}-1\}$ such that $F(\theta_t; \mathbf{X}^{(n)})$ has a small excess loss.

For the first case, suppose t be such that $\theta^\star \in \Theta_t$ and $\theta^\star \notin \Theta_{t+1}$. Therefore, we can write

$$\begin{aligned}\theta^\star \notin \Theta_{t+1} &\Leftrightarrow \left\langle \nabla F(\theta_t; \mathbf{X}^{(n)}) + \xi_{\text{dir},t}, \theta^\star - \theta_t\right\rangle \geq 0 \\ &\Leftrightarrow \left\langle \nabla F(\theta_t; \mathbf{X}^{(n)}), \theta^\star - \theta_t\right\rangle \geq -\langle \xi_{\text{dir},t}, \theta^\star - \theta_t\rangle.\end{aligned} \tag{35}$$

Notice that using the first-order convexity condition, we have $F(\theta_t; \mathbf{X}^{(n)}) - F(\theta^\star; \mathbf{X}^{(n)}) \leq \left\langle \nabla F(\theta_t; \mathbf{X}^{(n)}), \theta_t - \theta^\star\right\rangle$. Therefore, by Equation (35), we have

$$\begin{aligned}F(\theta_t; \mathbf{X}^{(n)}) - F(\theta^\star; \mathbf{X}^{(n)}) &\leq \left\langle \nabla F(\theta_t; \mathbf{X}^{(n)}), \theta_t - \theta^\star\right\rangle \\ &\leq \langle \xi_{\text{dir},t}, \theta^\star - \theta_t\rangle.\end{aligned} \tag{36}$$

Under the events $\mathcal{G}_1$ and $\mathcal{G}_{n,2}$, defined in Equations (32) and (34), we have

$$F(\theta_t; \mathbf{X}^{(n)}) - F(\theta^\star; \mathbf{X}^{(n)}) \leq \langle \xi_{\text{dir},t}, \theta^\star - \theta_t\rangle \leq \hat{\Delta}\cdot O\left(\sqrt{\frac{k_{\text{ft}}}{\rho}\log(k_{\text{ft}}/\beta)}\right). \tag{37}$$

For the second case, i.e., $\theta^\star \in \Theta_{k_{\text{ft}}}$, we have the following geometric fact [Nes98]: there exists $t \in \{0, \dots, k_{\text{ft}} - 1\}$ such that the distance of $\theta^\star$ and the separating hyperplane at time t satisfies

$$-\nu \leq \left\langle \frac{\nabla F(\theta_t; \mathbf{X}^{(n)}) + \xi_{\text{dir},t}}{\left\|\nabla F(\theta_t; \mathbf{X}^{(n)}) + \xi_{\text{dir},t}\right\|}, \theta^\star - \theta_t \right\rangle \leq 0. \tag{38}$$

Here ν is a constant such that $\nu^d \geq \exp(-\tau k_{\text{ft}})(25\hat{\Delta})^d$. The values of ν and k_{ft} will be determined later. Using the first order convexity, we can write

$$\begin{aligned}
&F(\theta^\star; \mathbf{X}^{(n)}) - F(\theta_t; \mathbf{X}^{(n)}) \\
&\geq \left\|\nabla F(\theta_t; \mathbf{X}^{(n)}) + \xi_{\text{dir},t}\right\| \left\langle \frac{\nabla F(\theta_t; \mathbf{X}^{(n)}) + \xi_{\text{dir},t}}{\left\|\nabla F(\theta_t; \mathbf{X}^{(n)}) + \xi_{\text{dir},t}\right\|}, \theta^\star - \theta_t \right\rangle - \langle \xi_{\text{dir},t}, \theta^\star - \theta_t \rangle \\
&\geq -\nu \left\|\nabla F(\theta_t; \mathbf{X}^{(n)}) + \xi_{\text{dir},t}\right\| - \langle \xi_{\text{dir},t}, \theta^\star - \theta_t \rangle \\
&\geq -\nu \left(2\left\|\nabla F(\theta_t; \mathbf{X}^{(n)})\right\| + 2\|\xi_{\text{dir},t}\|\right) - \langle \xi_{\text{dir},t}, \theta^\star - \theta_t \rangle \\
&\geq -\nu(2n + 2\|\xi_{\text{dir},t}\|) - \langle \xi_{\text{dir},t}, \theta^\star - \theta_t \rangle,
\end{aligned}$$

where the second step follows from the well-known inequality $\|a + b\| \leq 2\|a\| + 2\|b\|$ for every $a, b \in \mathbb{R}^d$. Then, the last step follows because for every $\theta \in \mathbb{R}^d$, $\left\|\nabla F(\theta; \mathbf{X}^{(n)}\right\| \leq n$. Therefore, under the events $\mathcal{G}_1$, $\mathcal{G}_{n,1}$, and $\mathcal{G}_{n,2}$, defined in Equations (32) to (34), we have the following bound on the suboptimality gap

$$\begin{aligned}
F(\theta_t; \mathbf{X}^{(n)}) - F(\theta^\star; \mathbf{X}^{(n)}) &\leq \nu(2n + 2\|\xi_{\text{dir},t}\|) + \langle \xi_{\text{dir},t}, \theta^\star - \theta_t \rangle \\
&\leq \nu(2n + 2\|\xi_{\text{dir},t}\|) + O\left(\hat{\Delta}\sqrt{\frac{k_{\text{ft}}}{\rho}\log(k_{\text{ft}}/\beta)}\right) \\
&\leq \nu \cdot O\left(n + \sqrt{\frac{dk_{\text{ft}}}{\rho}\left(1 + \sqrt{\frac{\log(k_{\text{ft}}/\beta)}{d}}\right)}\right) + O\left(\hat{\Delta}\sqrt{\frac{k_{\text{ft}}}{\rho}\log(k_{\text{ft}}/\beta)}\right).
\end{aligned} \tag{39}$$

Recall that ν satisfies $\nu^d \geq \exp(-\tau k_{\text{ft}})(25\hat{\Delta})^d$. It can be easily seen that by setting

$$k_{\text{ft}} = \Theta\left(\frac{d}{\tau}\log\left(\frac{n\sqrt{\rho}}{\sqrt{d}} + \sqrt{d}\right)\right),$$

under the events $\mathcal{G}_1$, $\mathcal{G}_{n,1}$, and $\mathcal{G}_{n,2}$, we can further upperbound Equation (39) as follows

$$F(\theta_t; \mathbf{X}^{(n)}) - F(\theta^\star; \mathbf{X}^{(n)}) \leq O\left(\hat{\Delta}\sqrt{\frac{k_{\text{ft}}}{\rho}\log(k_{\text{ft}}/\beta)}\right). \tag{40}$$

Therefore, from Equations (37) and (40), under the event $\mathcal{G}_1 \cap \mathcal{G}_{n,1} \cap \mathcal{G}_{n,2}$, for both cases we showed that there exists t such that

$$\begin{aligned}
&F(\theta_t; \mathbf{X}^{(n)}) - F(\theta^\star; \mathbf{X}^{(n)}) \leq \Delta_{0.8n}(\theta^\star) \cdot O\left(\sqrt{\frac{k_{\text{ft}}}{\rho}\log(k_{\text{ft}}/\beta)}\right) \\
&\text{or } F(\theta_t; \mathbf{X}^{(n)}) - F(\theta^\star; \mathbf{X}^{(n)}) \leq r \cdot O\left(\sqrt{\frac{k_{\text{ft}}}{\rho}\log(k_{\text{ft}}/\beta)}\right).
\end{aligned} \tag{41}$$

By the non-negativity of $\|\cdot\|_2$, we have

$$F(\theta^\star; \mathbf{X}^{(n)}) = \sum_{i=1}^{n} \|\theta^\star - x_i\| \geq 0.2n\Delta_{0.8n}(\theta^\star). \tag{42}$$

Therefore, we conclude that for both cases there exists t such that

$$\begin{aligned}
&F(\theta_t; \mathbf{X}^{(n)}) \leq \left(1 + O\left(\frac{1}{n}\sqrt{\frac{d\log(\kappa)}{\tau\rho} \cdot \log\left(\frac{d}{\tau\beta}\log(\kappa)\right)}\right)\right) F(\theta^\star; \mathbf{X}^{(n)}) \\
&\text{or } F(\theta_t; \mathbf{X}^{(n)}) - F(\theta^\star; \mathbf{X}^{(n)}) \leq r \cdot O\left(\sqrt{\frac{d\log(\kappa)}{\tau\rho} \cdot \log\left(\frac{d}{\tau\beta}\log(\kappa)\right)}\right)
\end{aligned} \tag{43}$$

where $\kappa \triangleq \frac{n\sqrt{\rho}}{\sqrt{d}} + \sqrt{d}$.

Let us define OPT $\triangleq \min_{t\in\{0,\ldots,k_{\text{ft}}-1\}}\{F(\theta_t;\mathbf{X}^{(n)}) - F(\theta^\star;\mathbf{X}^{(n)})\}$. In the next step of the proof, we show that the exponential mechanism in Line 10 with high probability can identify an iterate whose suboptimality gap is close to OPT. Using the properties of the exponential mechanism in Line 10 as outlied in Lemma 3.2, we have with probability at least $1-\beta/3$ over the randomness of the exponential mechanism

$$F(\theta_{\hat{t}};\mathbf{X}^{(n)}) - F(\theta^\star;\mathbf{X}^{(n)}) \leq \text{OPT} + \frac{448\hat{\Delta}}{\varepsilon}(\log(3k_{\text{ft}}/\beta)). \tag{44}$$

Notice that under the event $\mathcal{G}_1$ and using Equation (42), we have

$$\frac{448\hat{\Delta}}{\varepsilon}\log(3k_{\text{ft}}/\beta) \leq \frac{F(\theta^\star;\mathbf{X}^{(n)})}{n\varepsilon}\cdot O(\log(k_{\text{ft}}/\beta)) \text{ or } \frac{448\hat{\Delta}}{\varepsilon}\log(3k_{\text{ft}}/\beta) \leq \frac{r}{\varepsilon}O(\log(k_{\text{ft}}/\beta)) \tag{45}$$

Moreover, under the event $\mathcal{G}_1 \cap \mathcal{G}_{n,1} \cap \mathcal{G}_{n,2}$, we provided an upperbound on OPT in Equation (43). Combining Equation (43), Equation (44), and Equation (45), proves the first claim.

For the second statement, under the condition that $\max_{i\in[n]}|\mathbf{X}^{(n)}\cap\mathcal{B}_d(x_i,r)| < 3n/4$, we can define the following high probability event:

$$\mathcal{G}_2 = \Big\{\theta^\star \in \Theta_{\text{loc}} \text{ and } \Delta_{0.75n}(\theta^\star) \leq 4\hat{\Delta} \text{ and } \hat{\Delta} \leq 4\Delta_{0.8n}(\theta^\star)\Big\}.$$

The argument then proceeds in the same way as the argument for the first claim. □

F Proof of Section 4

Proof of Lemma 4.2. For $i \in [n-k]$, we can write

$$\begin{aligned}\|\theta_k - x_i\| &\geq \|\theta_k - \theta_0\| - \|\theta_0 - x_i\| \\ &= \|\theta_k - \theta_0\| - 2\|\theta_0 - x_i\| + \|\theta_0 - x_i\|.\end{aligned}$$

Also for every $j \in [k]$, we have

$$\|\theta_k - y_j\| \geq \|\theta_0 - y_j\| - \|\theta_0 - \theta_k\|.$$

Summing both sides of these inequalities, we obtain

$$\begin{aligned}&\sum_{i=1}^{n-k}\|\theta_k - x_i\| + \sum_{j=1}^{k}\|\theta_k - y_j\| \\ &\geq (n-2k)\|\theta_0 - \theta_k\| - 2\sum_{i=1}^{n-k}\|\theta_0 - x_i\| + \sum_{i=1}^{n-k}\|\theta_0 - x_i\| + \sum_{j=1}^{k}\|\theta_0 - y_j\| \\ &(\Leftrightarrow) \sum_{i=1}^{n-k}\|\theta_k - x_i\| + \sum_{j=1}^{k}\|\theta_k - y_j\| - \left(\sum_{i=1}^{n-k}\|\theta_0 - x_i\| + \sum_{j=1}^{k}\|\theta_0 - y_j\|\right) \\ &\geq (n-2k)\|\theta_0 - \theta_k\| - 2\sum_{i=1}^{n-k}\|\theta_0 - x_i\|\end{aligned}$$

Since $\sum_{i=1}^{n-k}\|\theta_k - x_i\| + \sum_{j=1}^{k}\|\theta_k - y_j\| - \left[\sum_{i=1}^{n-k}\|\theta_0 - x_i\| + \sum_{j=1}^{k}\|\theta_0 - y_j\|\right] \leq 0$ by the assumption that $\theta_k = \text{GM}(x_1,\ldots,x_{n-k},y_1,\ldots,y_k)$, we obtain

$$\begin{aligned}&(n-2k)\|\theta_0 - \theta_k\| - 2\sum_{i=1}^{n-k}\|\theta_0 - x_i\| \leq 0 \\ &\Leftrightarrow \|\theta_0 - \theta_k\| \leq \frac{2}{n-2k}\cdot\sum_{i=1}^{n-k}\|\theta_0 - x_i\| \\ &\Rightarrow \|\theta_0 - \theta_k\| \leq \frac{2}{n-2k}\cdot\left(\sum_{i=1}^{n-k}\|\theta_0 - x_i\| + \sum_{i=n-k+1}^{n}\|\theta_0 - x_i\|\right) \\ &\Leftrightarrow \|\theta_0 - \theta_k\| \leq \frac{2}{n-2k}\cdot F(\theta_0;(x_1,\ldots,x_n)).\end{aligned}$$

□

Proof of Theorem 4.1. We claim that for every neighbouring datasets $\mathbf{X} \in (\mathcal{B}_d(R))^n$ and $\mathbf{X}' \in (\mathcal{B}_d(R))^n$ and for every $y \in \mathcal{B}_d(R)$, we have

$$|\mathrm{len}_r(\mathbf{X}, y) - \mathrm{len}_r(\mathbf{X}', y)| \leq 1.$$

This follows from the fact that for every $\tilde{X}$, we have $\mathrm{d_H}(\mathbf{X}, \tilde{\mathbf{X}}) \leq \mathrm{d_H}(\mathbf{X}', \tilde{\mathbf{X}}) + 1$. Then, the proof of privacy follows from the privacy proof of the exponential mechanism [MT07].

Next, we present the utility proof. Let $k \in \mathbb{N}$ be a constant that determined later. Define the following two sets:

$$A_1 = \{y \in \mathcal{B}_d(R) : \mathrm{len}_r(\mathbf{X}, y) \geq k\}$$
$$A_2 = \{y \in \mathcal{B}_d(R) : \mathrm{len}_r(\mathbf{X}, y) = 0\}$$

Then,

$$\begin{aligned} \frac{\mathbb{P}_{\hat{\theta}\sim\pi}(\hat{\theta} \in A_1)}{\mathbb{P}_{\hat{\theta}\sim\pi}(\hat{\theta} \in A_2)} &= \frac{\int_{y\in A_1} \exp\big(-\frac{\varepsilon}{2} \cdot \mathrm{len}_r(\mathbf{X}, y)\big) dy}{\int_{y\in A_2} \exp\big(-\frac{\varepsilon}{2} \cdot \mathrm{len}_r(\mathbf{X}, y)\big) dy} \\ &\leq \frac{\exp(-\frac{\varepsilon}{2}k) \int_{y\in A_1} dy}{\int_{y\in A_2} dy}. \end{aligned} \tag{46}$$

We can use the following simple facts: $\displaystyle\int_{y\in A_1} dy \leq \int_{y\in\mathcal{B}_d(R)} dy = V_1 R^d$ where V_1 is the volume of the ball of radius one in $\mathbb{R}^d$. For A_2 notice that, for all $y \in \mathcal{B}_d(\mathrm{GM}(\mathbf{X}), r)$, we have $\mathrm{len}_r(\mathbf{X}, y) = 0$. Thus, $\int_{y\in A_2} dy \geq V_1 r^d$. Putting these two pieces together,

$$\frac{\mathbb{P}_{\hat{\theta}\sim\pi}(\hat{\theta} \in A_1)}{\mathbb{P}_{\hat{\theta}\sim\pi}(\hat{\theta} \in A_2)} \leq \exp\Big(-\frac{\varepsilon}{2}k\Big)\left(\frac{R}{r}\right)^d \Rightarrow \mathbb{P}_{\hat{\theta}\sim\pi}(\hat{\theta} \in A_1) \leq \exp\Big(-\frac{\varepsilon}{2}k\Big)\left(\frac{R}{r}\right)^d, \tag{47}$$

where the last step follows from the fact that $\mathbb{P}_{\hat{\theta}\sim\pi}(\hat{\theta} \in A_2) \leq 1$. Therefore, we obtain that for every $\beta \in (0, 1)$ with probability at least $1 - \beta$ we have

$$\mathrm{len}_r(\mathbf{X}, \hat{\theta}) \leq \left\lfloor \frac{2}{\varepsilon}\left(\log\left(\frac{1}{\beta}\right) + d\log\left(\frac{R}{r}\right)\right)\right\rfloor \triangleq k^\star, \tag{48}$$

where $\hat{\theta} \sim \pi$.

Under the above event, let $\hat{\theta} \in \mathcal{B}_d(R)$ be such that $\mathrm{len}_r(\mathbf{X}, \hat{\theta}) \leq k^\star$. This is equivalent to the following: there exists $z \in \mathcal{B}_d(\hat{\theta}, r)$ and $\tilde{\mathbf{X}} \in (\mathbb{R}^d)^n$ such that $z = \mathrm{GM}(\tilde{\mathbf{X}})$ and $\mathrm{d_H}(\mathbf{X}, \tilde{\mathbf{X}}) \leq k^\star$. Using this observation, we can write

$$\begin{aligned} \left\|\hat{\theta} - \mathrm{GM}(\mathbf{X})\right\| &\leq \|z - \mathrm{GM}(\mathbf{X})\| + \left\|\hat{\theta} - z\right\| \\ &\leq \|z - \mathrm{GM}(\mathbf{X})\| + r \\ &= \left\|\mathrm{GM}(\tilde{\mathbf{X}}) - \mathrm{GM}(\mathbf{X})\right\| + r. \end{aligned} \tag{49}$$

Suboptimality Gap: Let $\theta^\star \in \mathrm{GM}(\mathbf{X})$, then

$$\begin{aligned} F(\hat{\theta}; \mathbf{X}) - F(\theta^\star; \mathbf{X}) &= \sum_{i=1}^n \Big(\left\|\hat{\theta} - x_i\right\| - \|\theta^\star - x_i\|\Big) \\ &\leq \sum_{i=1}^n (\|z - x_i\| + r - \|\theta^\star - x_i\|) \\ &= nr + \sum_{i=1}^n \Big(\left\|\mathrm{GM}(\tilde{\mathbf{X}}) - x_i\right\| - \|\theta^\star - x_i\|\Big). \end{aligned}$$

Define $\mathcal{I} \subseteq [n]$ be the indices of the points that $\mathbf{X}$ and $\tilde{\mathbf{X}}$ differs. We know that $|\mathcal{I}| \leq k^\star$. Then, we can write

$$\begin{aligned}
&\sum_{i=1}^{n}\Big(\Big\|\mathrm{GM}(\tilde{\mathbf{X}}) - x_i\Big\| - \|\theta^\star - x_i\|\Big) \\
&= \underbrace{\sum_{i\in\mathcal{I}}\Big(\Big\|\mathrm{GM}(\tilde{\mathbf{X}}) - x_i\Big\| - \|\theta^\star - x_i\|\Big)}_{A_1} + \underbrace{\sum_{i\in[n]/\mathcal{I}}\Big(\Big\|\mathrm{GM}(\tilde{\mathbf{X}}) - x_i\Big\| - \|\theta^\star - x_i\|\Big)}_{A_2}.
\end{aligned} \tag{50}$$

By triangle inequality, we can write

$$\begin{aligned}
A_1 &= \sum_{i\in\mathcal{I}}\Big(\Big\|\mathrm{GM}(\tilde{\mathbf{X}}) - x_i\Big\| - \|\theta^\star - x_i\|\Big) \\
&\leq |\mathcal{I}|\Big\|\theta^\star - \mathrm{GM}(\tilde{\mathbf{X}})\Big\| \\
&\leq k^\star \cdot \Big\|\theta^\star - \mathrm{GM}(\tilde{\mathbf{X}})\Big\|.
\end{aligned} \tag{51}$$

For $i \in \mathcal{I}$, let $(\tilde{\mathbf{X}})_i = x_i'$ where $(\tilde{\mathbf{X}})_i$ denote the i-th data point in $\tilde{\mathbf{X}}$. Since $\mathrm{GM}(\tilde{\mathbf{X}})$ is a geometric median of $\tilde{\mathbf{X}}$, by the first-order optimality condition

$$\nabla_\theta F(\mathrm{GM}(\tilde{\mathbf{X}}); \tilde{\mathbf{X}}) = 0 \Leftrightarrow \sum_{i\in[n]/\mathcal{I}} \nabla_\theta\Big(\Big\|\mathrm{GM}(\tilde{\mathbf{X}}) - x_i\Big\|\Big) = -\sum_{i\in\mathcal{I}} \nabla_\theta\Big(\Big\|\mathrm{GM}(\tilde{\mathbf{X}}) - x_i'\Big\|\Big). \tag{52}$$

To control A_2, notice that $\|\theta - x_i\|$ is a convex function in θ for every x_i. By the first-order convexity condition, for every θ_1 and θ_2, we have $\|\theta_1 - x_i\| - \|\theta_2 - x_i\| \leq \langle \nabla(\|\theta_1 - x_i\|), \theta_1 - \theta_2\rangle$. Therefore, we can write

$$\sum_{i\in[n]/\mathcal{I}}\Big(\Big\|\mathrm{GM}(\tilde{\mathbf{X}}) - x_i\Big\| - \|\theta^\star - x_i\|\Big) \leq \sum_{i\in[n]/\mathcal{I}}\Big\langle \nabla\Big(\Big\|\mathrm{GM}(\tilde{\mathbf{X}}) - x_i\Big\|\Big), \mathrm{GM}(\tilde{\mathbf{X}}) - \theta^\star\Big\rangle. \tag{53}$$

Then, by Equation (52),

$$\begin{aligned}
A_2 &= \sum_{i\in[n]/\mathcal{I}}\Big\langle \nabla\Big(\Big\|\mathrm{GM}(\tilde{\mathbf{X}}) - x_i\Big\|\Big), \mathrm{GM}(\tilde{\mathbf{X}}) - \theta^\star\Big\rangle \\
&= -\sum_{i\in\mathcal{I}}\Big\langle \nabla_\theta\Big(\Big\|\mathrm{GM}(\tilde{\mathbf{X}}) - x_i'\Big\|\Big), \mathrm{GM}(\tilde{\mathbf{X}}) - \theta^\star\Big\rangle.
\end{aligned}$$

Finally notice that by Equation (3), for every x_i', $\Big\|\nabla_\theta\Big(\Big\|\mathrm{GM}(\tilde{\mathbf{X}}) - x_i'\Big\|\Big)\Big\| \leq 1$. Therefore, by Cauchy–Schwarz inequality

$$\begin{aligned}
A_2 &= -\sum_{i\in\mathcal{I}}\Big\langle \nabla_\theta\Big(\Big\|\mathrm{GM}(\tilde{\mathbf{X}}) - x_i'\Big\|\Big), \mathrm{GM}(\tilde{\mathbf{X}}) - \theta^\star\Big\rangle \\
&\leq |\mathcal{I}|\Big\|\mathrm{GM}(\tilde{\mathbf{X}}) - \theta^\star\Big\| \\
&\leq k^\star\Big\|\mathrm{GM}(\tilde{\mathbf{X}}) - \theta^\star\Big\|.
\end{aligned} \tag{54}$$

By Equations (51) and (54), we obtain

$$F(\hat{\theta}; \mathbf{X}) - F(\theta^\star; \mathbf{X}) \leq nr + 2k^\star\Big\|\mathrm{GM}(\tilde{\mathbf{X}}) - \theta^\star\Big\|.$$

Then, we invoke Lemma 4.2 which states that $\Big\|\mathrm{GM}(\tilde{\mathbf{X}}) - \mathrm{GM}(\mathbf{X})\Big\| \leq \dfrac{2}{n - 2k^\star} \cdot F(\mathrm{GM}(\mathbf{X}); \mathbf{X})$. Putting all the pieces together,

$$\begin{aligned}
F(\hat{\theta}; \mathbf{X}) - F(\theta^\star; \mathbf{X}) &\leq nr + 2k^\star\Big\|\mathrm{GM}(\tilde{\mathbf{X}}) - \theta^\star\Big\| \\
&\leq nr + \frac{4k^\star}{n - 2k^\star} \cdot F(\theta^\star; \mathbf{X}),
\end{aligned} \tag{55}$$

as was to be shown.

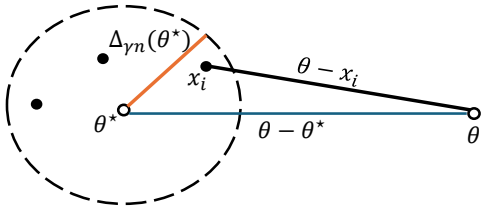


Figure 1: Graphical Intuition Behind Equation (57)

Distance to $\theta^\star$: From Equation (49), we know that $\left\|\hat{\theta} - \mathrm{GM}(\mathbf{X})\right\| \leq \left\|\mathrm{GM}(\tilde{\mathbf{X}}) - \mathrm{GM}(\mathbf{X})\right\| + r$ where $\tilde{\mathbf{X}}$ is a dataset of size n such that $\mathrm{d_H}(\mathbf{X}, \tilde{\mathbf{X}}) \leq k^\star$. The proof is based on characterizing the worst case distance between the geometric median of two datasets that differ in at most $k^\star$ points.

For the dataset $\mathbf{X}^{(n)}$, recall that $\theta^\star = \mathrm{GM}\big(\mathbf{X}^{(n)}\big)$. Also, recall the definition of $\Delta_{\gamma n}(\theta^\star)$ from Definition 1.1. Let $\theta \in \mathbb{R}^d$ be such that $\|\theta - \theta^\star\| > \Delta_{\gamma n}(\theta^\star)$. Define $m = |\tilde{\mathbf{X}} \cap \mathcal{B}_d(\theta^\star, \Delta_{\gamma n}(\theta^\star))|$. By the variational representation of $\|\cdot\|$, we can write

$$\begin{aligned}
\left\|\nabla F(\theta; \tilde{\mathbf{X}})\right\| &\geq \left\langle \nabla F(\theta; \tilde{\mathbf{X}}), \frac{\theta - \theta^\star}{\|\theta - \theta^\star\|} \right\rangle \\
&= \sum_{x \in \tilde{\mathbf{X}} \cap \mathcal{B}_d(\theta^\star, \Delta_{\gamma n}(\theta^\star))} \left\langle \frac{\theta - x}{\|\theta - x\|}, \frac{\theta - \theta^\star}{\|\theta - \theta^\star\|} \right\rangle + \sum_{x \in \tilde{\mathbf{X}} \setminus \{\tilde{\mathbf{X}} \cap \mathcal{B}_d(\theta^\star, \Delta_{\gamma n}(\theta^\star))\}} \left\langle \frac{\theta - x}{\|\theta - x\|}, \frac{\theta - \theta^\star}{\|\theta - \theta^\star\|} \right\rangle \\
&\geq \sum_{x \in \tilde{\mathbf{X}} \cap \mathcal{B}_d(\theta^\star, \Delta_{\gamma n}(\theta^\star))} \left\langle \frac{\theta - x}{\|\theta - x\|}, \frac{\theta - \theta^\star}{\|\theta - \theta^\star\|} \right\rangle - (n - m),
\end{aligned} \tag{56}$$

where the last step follows from Cauchy-Schwarz inequality. Then, we claim that for every $x \in \tilde{\mathbf{X}} \cap \mathcal{B}_d(\theta^\star, \Delta_{\gamma n}(\theta^\star))$, we have

$$\left\langle \frac{\theta - x}{\|\theta - x\|}, \frac{\theta - \theta^\star}{\|\theta - \theta^\star\|} \right\rangle \geq \sqrt{1 - \left(\frac{\Delta_{\gamma n}(\theta^\star)}{\|\theta - \theta^\star\|}\right)^2} \tag{57}$$

To gain the intuition behind it see Figure 1. Therefore, from Equation (56),

$$\left\|\nabla F(\theta; \tilde{\mathbf{X}})\right\| \geq m\sqrt{1 - \left(\frac{\Delta_{\gamma n}(\theta^\star)}{\|\theta - \theta^\star\|}\right)^2} - (n - m).$$

We are interested on characterizing the condition under which $\left\|\nabla F(\theta; \tilde{\mathbf{X}})\right\| > 0$. A sufficient condition is that given $n < 2m$

$$m\sqrt{1 - \left(\frac{\Delta_{\gamma n}(\theta^\star)}{\|\theta - \theta^\star\|}\right)^2} - (n - m) > 0 \quad (\Leftrightarrow) \quad \Delta_{\gamma n}(\theta^\star)\frac{1}{\sqrt{2\frac{m}{n} - \left(\frac{m}{n}\right)^2}} < \|\theta - \theta^\star\|.$$

This shows that the distance of $\mathrm{GM}(\tilde{\mathbf{X}})$ and $\theta^\star$ has to satisfy

$$\left\|\mathrm{GM}(\tilde{\mathbf{X}}) - \theta^\star\right\| \leq \frac{\Delta_{\gamma n}(\theta^\star)}{\sqrt{2\frac{m}{n} - \left(\frac{m}{n}\right)^2}}.$$

The function $h(x) = \frac{1}{\sqrt{2x - x^2}}$ is decreasing in the range of $x \in (0, 1]$. Also, notice that $m = |\tilde{\mathbf{X}} \cap \mathcal{B}_d(\theta^\star, \Delta_{\gamma n}(\theta^\star))| \geq \gamma n - k^\star$. Therefore,

$$\left\|\mathrm{GM}(\tilde{\mathbf{X}}) - \theta^\star\right\| \leq \frac{\Delta_{\gamma n}(\theta^\star)}{\sqrt{2\left(\gamma - \frac{k^\star}{n}\right) - \left(\gamma - \frac{k^\star}{n}\right)^2}},$$

as was to be shown.

□

G Proof of Section 5

Proof of Theorem 5.1. The proof is based on the reduction provided in Lemma G.1 and the lower-bound on the sample complexity of the mean estimation of Gaussian distribution with known covariance matrix in [KLSU19, Thm. 6.5]. □

Lemma G.1. *Let $\varepsilon \le 49 \times 10^{-5}$, $\alpha \le 49 \times 10^{-5}$, $\delta \le 10^{-4}$, and $d \ge 22500$ be constants. Let $\mathcal{A}_n$ be an arbitrary (ε, δ)-DP algorithm such that for every dataset $\mathbf{X}^{(n)}$, its output satisfies*

$$\mathbb{E}_{\hat{\theta}\sim\mathcal{A}_n(\mathbf{X}^{(n)})}\Big[F(\hat{\theta};\mathbf{X}^{(n)})\Big] \le (1+\alpha)\min_{\theta\in\mathcal{B}_d^\infty(1)} F(\theta;\mathbf{X}^{(n)}).$$

Let $\mu \in \mathcal{B}_d^\infty(1)$ and let $\mathbf{X}^{(n)} = (X_1,\dots,X_n) \sim \mathcal{N}(\mu,\mathbb{I}_d)^{\otimes n}$. Let $\hat{\theta} \sim \mathcal{A}_n(\mathbf{X}^{(n)})$. Then, with probability at least $2/3$ over $\mathbf{X}^{(n)}$ and the internal randomness of $\mathcal{A}_n$, we have

$$\Big\|\hat{\theta} - \mu\Big\| \le 0.2\sqrt{d}.$$

Proof. The proof consists of several steps:

Step 1: Bound on the Empirical Error. Let $\mathcal{A}_n$ be an arbitrary (ε,δ)-DP algorithm such that for every dataset $\mathbf{X}^{(n)}$, its output satisfies

$$F(\hat{\theta};\mathbf{X}^{(n)}) \le (1+\alpha)\min_{\theta\in\mathcal{B}_d^\infty(R)} F(\theta;\mathbf{X}^{(n)}).$$

Let $\mu \in \mathcal{B}_d^\infty(R)$ and $\mathbf{X}^{(n)} = (X_1,\dots,X_n) \sim \mathcal{N}(\mu,\mathbb{I}_d)^{\otimes n}$. The utility guarantee of the algorithm implies that

$$\begin{aligned}\mathbb{E}\Big[F(\hat{\theta};\mathbf{X}^{(n)})\Big] &\le (1+\alpha)\mathbb{E}\left[\min_{\theta\in\mathcal{B}_d^\infty(R)} F(\theta;\mathbf{X}^{(n)})\right]\\ &\le (1+\alpha)\mathbb{E}\Big[F(\mu;\mathbf{X}^{(n)})\Big].\end{aligned}$$

To further upperbound the last step, we can use Jensen's inequality to write

$$\begin{aligned}\frac{1}{n}\cdot\mathbb{E}\Big[F(\mu;\mathbf{X}^{(n)})\Big] &= \mathbb{E}[\|X_1-\mu\|]\\ &\le \sqrt{\mathbb{E}\Big[\|X_1-\mu\|^2\Big]}\\ &= \sqrt{d}\end{aligned}$$

Therefore, in-expectation over $\mathbf{X}^{(n)} \sim \mathcal{N}(\mu,\mathbb{I}_d)^{\otimes n}$ and the internal randomness of $\mathcal{A}_n$, we have

$$\mathbb{E}_{\mathbf{X}^{(n)}\sim\mathcal{N}(\mu,\mathbb{I}_d)^{\otimes n},\hat{\theta}\sim\mathcal{A}_n(\mathbf{X}^{(n)})}\Big[F\Big(\hat{\theta};\mathbf{X}^{(n)}\Big) - F\Big(\mu;\mathbf{X}^{(n)}\Big)\Big] \le n\alpha\sqrt{d}. \tag{58}$$

Step 2: Relating Empirical Error to Population Error. Let $(X_0,X_1,\dots,X_n) \sim \mathcal{N}(\mu,\mathbb{I}_d)^{\otimes(n+1)}$. With an abuse of notation, let $\theta = \mathcal{A}_n((X_1,\dots,X_n))$, and, for every $i \in [n]$, let $\theta^{(i)} = \mathcal{A}_n((X_1,\dots,X_{i-1},X_0,X_{i+1},\dots,X_n))$. Let T be a constant that will be determined later. We can write

$$\begin{aligned}\mathbb{E}\Big[\Big\|\theta^{(i)} - X_i\Big\|\Big] &= \int_{t=0}^{\infty}\mathbb{P}\Big(\Big\|\theta^{(i)} - X_i\Big\| \ge t\Big)\mathrm{d}t\\ &= \int_{t=0}^{T}\mathbb{P}\Big(\Big\|\theta^{(i)} - X_i\Big\| \ge t\Big)\mathrm{d}t + \int_{t=T}^{\infty}\mathbb{P}\Big(\Big\|\theta^{(i)} - X_i\Big\| \ge t\Big)\mathrm{d}t.\end{aligned} \tag{59}$$

Consider the first term in Equation (59). Since $\mathcal{A}_n$ satisfies (ε,δ)-DP,

$$\begin{aligned}\mathbb{P}\Big(\Big\|\theta^{(i)} - X_i\Big\| \ge t\Big) &= \mathbb{E}\Big[\mathbb{P}\Big(\Big\|\theta^{(i)} - X_i\Big\| \ge t\Big|(X_0,\dots,X_n)\Big)\Big]\\ &\le \mathbb{E}\Big[\exp(\varepsilon)\mathbb{P}\Big(\|\theta - X_i\| \ge t\Big|(X_0,\dots,X_n)\Big) + \delta\Big]\\ &= \exp(\varepsilon)\cdot\mathbb{P}(\|\theta - X_i\| \ge t) + \delta.\end{aligned} \tag{60}$$

Therefore, the first term can be upperbounded as

$$\begin{aligned}\int_{t=0}^{T} \mathbb{P}\Big(\Big\|\theta^{(i)} - X_i\Big\| \geq t\Big)\mathrm{d}t &\leq \exp(\varepsilon) \cdot \int_{t=0}^{T} \mathbb{P}(\|\theta - X_i\| \geq t)\mathrm{d}t + T\delta \\ &\leq \exp(\varepsilon) \cdot \mathbb{E}[\|\theta - X_i\|] + T\delta.\end{aligned}$$

In the next step, we upperbound the the second term in Equation (59). Notice that

$$\begin{aligned}\Big\{(\theta^{(i)}, X_i) : \Big\|\theta^{(i)} - \mu - (X_i - \mu)\Big\| \geq t\Big\} &\subseteq \Big\{(\theta^{(i)}, X_i) : \|X_i - \mu\| \geq t - \Big\|\theta^{(i)} - \mu\Big\|\Big\} \\ &\subseteq \Big\{X_i : \|X_i - \mu\| \geq t - 2R\sqrt{d}\Big\},\end{aligned} \tag{61}$$

where the first step follows from the triangle inequality and the last step follows because μ and $\theta^{(i)}$ are in $\mathcal{B}_d^\infty(R)$. Using this, we can write

$$\begin{aligned}\int_{t=T}^{\infty} \mathbb{P}\Big(\Big\|\theta^{(i)} - X_i\Big\| \geq t\Big)\mathrm{d}t &\leq \int_{t=T}^{\infty} \mathbb{P}\Big(\|X_i - \mu\| \geq t - 2R\sqrt{d}\Big)\mathrm{d}t \\ &= \int_{u=T-(2R+1)\sqrt{d}}^{\infty} \mathbb{P}\Big(\|X_i - \mu\| \geq u + \sqrt{d}\Big)\mathrm{d}u,\end{aligned} \tag{62}$$

where the last step follows from the change of variable $u = t - (2R+1)\sqrt{d}$. In the next step, we use the concentration bounds for the norm of multivariate Guassian random variable. Using Lemma C.2, we can write

$$\begin{aligned}\mathbb{P}\Big(\|X_i - \mu\| \geq u + \sqrt{d}\Big) &= \mathbb{P}\Big(\|X_i - \mu\|^2 \geq u^2 + d + 2u\sqrt{d}\Big) \\ &\leq \exp\Big(-\frac{u^2}{2}\Big).\end{aligned} \tag{63}$$

Let $T = 2(2R+1)\sqrt{d}$. Then, using standard bounds on the *complementary error function* [Ksc17], we can write

$$\begin{aligned}\int_{t=T}^{\infty} \mathbb{P}\Big(\|X_i - \mu\| \geq u + \sqrt{d}\Big) &\leq \int_{u=(2R+1)\sqrt{d}}^{\infty} \exp\Big(-\frac{u^2}{2}\Big)\mathrm{d}u \\ &\leq \frac{1}{(4R+2)\sqrt{d}} \exp\Big(-2(2R+1)^2 d\Big).\end{aligned} \tag{64}$$

In the last step, we claim that $\mathbb{E}[\|\theta - X_0\|] = \mathbb{E}\big[\big\|\theta^{(i)} - X_i\big\|\big]$ for every $i \in [n]$. It is because $\theta^{(i)} \stackrel{\mathrm{d}}{=} \theta$, $X_i \stackrel{\mathrm{d}}{=} X_0$, and $\theta^{(i)} \perp\!\!\!\perp X_i$. Ergo, combining and summing over $i \in [n]$, we obtain

$$\begin{aligned}&\mathbb{E}[\|\theta - X_0\|] \\ &\leq \exp(\varepsilon)\left(\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\|\theta - X_i\|]\right) + (4R+2)\sqrt{d}\delta + \frac{1}{(4R+2)\sqrt{d}} \exp\Big(-2(2R+1)^2 d\Big)\end{aligned}$$

This bound implies that

$$\begin{aligned}&\mathbb{E}[\|\theta - X_0\|] - \mathbb{E}[\|\mu - X_0\|] \\ &\leq \exp(\varepsilon)\left(\frac{1}{n}\sum_{i=1}^{n}(\mathbb{E}[\|\theta - X_i\|] - \mathbb{E}[\|\mu - X_0\|])\right) + (\exp(\varepsilon) - 1)\mathbb{E}[\|\mu - X_0\|] \\ &+ (4R+2)\sqrt{d}\delta + \frac{1}{(4R+2)\sqrt{d}} \exp\Big(-2(2R+1)^2 d\Big).\end{aligned}$$

This equation can be rephrased as follows

$$\mathbb{E}[\|\theta - X_0\|] - \mathbb{E}[\|\mu - X_0\|] \leq \beta\sqrt{d} \tag{65}$$

where

$$\beta = \exp(\varepsilon)\alpha + (\exp(\varepsilon) - 1) + (4R+2)\delta + \frac{1}{(4R+2)d} \exp\Big(-2(2R+1)^2 d\Big). \tag{66}$$

Step 3: Relating Population Error to Distance In Step 2, we showed that in-expectation over $(X_0, \dots, X_n) \sim \mathcal{N}(\mu, \mathbb{I}_d)^{\otimes(n+1)}$ and $\hat{\theta} \sim \mathcal{A}_n(\mathbf{X}^{(n)})$ where $\mathbf{X}^{(n)} = (X_1, \dots, X_n)$, we have

$$\mathbb{E}\Big[\Big\|\hat{\theta} - X_0\Big\|\Big] - \mathbb{E}[\|\mu - X_0\|] \le \beta\sqrt{d}. \tag{67}$$

For notiational convenience, let $h : \mathbb{R}^d \to \mathbb{R}$ be $h(\theta) \triangleq \mathbb{E}_{X\sim\mathcal{N}(\mu,\mathbb{I}_d)}[\|\theta - X\|]$. Equation (67) can be written as

$$\mathbb{E}_{\mathbf{X}^{(n)}\sim\mathcal{N}(\mu,\mathbb{I}_d)^{\otimes n},\hat{\theta}\sim\mathcal{A}_n(\mathbf{X}^{(n)})}\Big[h(\hat{\theta}) - h(\mu)\Big] \le \beta\sqrt{d}.$$

Since μ is the minimizer of $h(\theta)$, for every $\theta \in \mathbb{R}^d$ we have that $h(\theta) \ge h(\mu)$. Therefore, $h(\hat{\theta}) - h(\mu)$ is a non-negative random variable. We can invoke Markov's inequality to write

$$\mathbb{P}_{\mathbf{X}^{(n)}\sim\mathcal{N}(\mu,\mathbb{I}_d)^{\otimes n},\hat{\theta}\sim\mathcal{A}_n(\mathbf{X}^{(n)})}\Big(h(\hat{\theta}) - h(\mu) \le 3\beta\sqrt{d}\Big) \ge \frac{2}{3}. \tag{68}$$

In the next step, we provide a *deterministic* argument: For every $\theta \in \mathbb{R}^d$ such that $h(\theta) - h(\mu) \le 3\beta\sqrt{d}$, we provide an upperbound on $\|\theta - \mu\|$. By subtracting μ, we can write

$$\begin{aligned} h(\theta) - h(\mu) &= \mathbb{E}_{X\sim\mathcal{N}(\mu,\mathbb{I}_d)}[\|\theta - X\| - \|\mu - X\|] \\ &= \mathbb{E}_{Z\sim\mathcal{N}(0,\mathbb{I}_d)}[\|\theta - \mu + Z\| - \|Z\|]. \end{aligned} \tag{69}$$

Define the following events

$$\begin{aligned} \mathcal{E}_1 &\triangleq \left\{ Z : \sqrt{d\left(1 - 2\sqrt{\frac{\log(4/\gamma)}{d}}\right)} \le \|Z\| \le \sqrt{d\left(1 + 4\sqrt{\frac{\log(4/\gamma)}{d}}\right)} \right\}, \\ \mathcal{E}_2 &\triangleq \Big\{ Z : \langle \theta - \mu, Z\rangle \ge -\|\theta - \mu\|\sqrt{2\log(2/\gamma)} \Big\}. \end{aligned} \tag{70}$$

Using Corollary C.3 and simple concentration bound for Gaussian random variable we have that $\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) \ge 1 - \gamma$. Let $\mathcal{E} = \mathcal{E}_1 \cap \mathcal{E}_2$. By dropping the positive term, we can write

$$\begin{aligned} &\mathbb{E}[\|\theta - \mu + Z\| - \|Z\|] \\ &= \mathbb{E}[(\|\theta - \mu + Z\| - \|Z\|) \cdot \mathbb{1}[\mathcal{E}]] + \mathbb{E}[(\|\theta - \mu + Z\| - \|Z\|) \cdot \mathbb{1}[\mathcal{E}^c]] \\ &\ge \mathbb{E}[(\|\theta - \mu + Z\| - \|Z\|) \cdot \mathbb{1}[\mathcal{E}]] - \mathbb{E}[\|Z\| \cdot \mathbb{1}[\mathcal{E}^c]]. \end{aligned} \tag{71}$$

Using Cauchy-Schwarz inequality, $\mathbb{E}[\|Z\| \cdot \mathbb{1}[\mathcal{E}^c]] \le \sqrt{\mathbb{P}(\mathcal{E}^c)}\sqrt{\mathbb{E}[\|Z\|^2]} = \sqrt{\mathbb{P}(\mathcal{E}^c)}\sqrt{d} \le \sqrt{\gamma}\sqrt{d}$. In the next step, we analyze the first term.

$$\begin{aligned} &\mathbb{E}[(\|\theta - \mu + Z\| - \|Z\|) \cdot \mathbb{1}[\mathcal{E}]] \\ &= \mathbb{E}\left[\left(\sqrt{\|\theta - \mu\|^2 + \|Z\|^2 + 2\langle \theta - \mu, Z\rangle} - \|Z\|\right) \cdot \mathbb{1}[\mathcal{E}]\right] \\ &= \mathbb{E}\left[\|Z\|\left(\sqrt{1 + \frac{\|\theta - \mu\|^2}{\|Z\|^2} + 2\frac{\langle \theta - \mu, Z\rangle}{\|Z\|^2}} - 1\right) \cdot \mathbb{1}[\mathcal{E}]\right] \end{aligned} \tag{72}$$

The value of γ will be determined later. Let d be large enough such that

$$\left(1 - 2\sqrt{\frac{\log(4/\gamma)}{d}}\right) = 0.9 \;\text{ and }\; \left(1 + 4\sqrt{\frac{\log(4/\gamma)}{d}}\right) = 1.1. \tag{73}$$

Then, we can write

$$\begin{aligned} &\mathbb{E}\left[\|Z\|\left(\sqrt{1 + \frac{\|\theta - \mu\|^2}{\|Z\|^2} + 2\frac{\langle \theta - \mu, Z\rangle}{\|Z\|^2}} - 1\right) \cdot \mathbb{1}[\mathcal{E}]\right] \\ &\ge \sqrt{0.9d}\left(\sqrt{1 + \frac{\|\theta - \mu\|^2}{1.1d} - \frac{2\sqrt{2\log(2/\gamma)}\|\theta - \mu\|}{0.9d}} - 1\right). \end{aligned} \tag{74}$$

Notice that we assumed that $h(\theta) - h(\mu) \leq 3\beta\sqrt{d}$. Therefore, we have

$$\sqrt{0.9d}\left(\sqrt{1 + \frac{\|\theta-\mu\|^2}{1.1d} - \frac{2\sqrt{2\log(2/\gamma)}\|\theta-\mu\|}{0.9d}} - 1\right) - \sqrt{\gamma}\sqrt{d} \leq 3\beta\sqrt{d}$$
$$(\Leftrightarrow)\sqrt{1 + \frac{\|\theta-\mu\|^2}{1.1d} - \frac{2\sqrt{2\log(2/\gamma)}\|\theta-\mu\|}{0.9d}} \leq \frac{(3\beta+\sqrt{\gamma})}{\sqrt{0.9}} + 1. \tag{75}$$

Simple calculations show that this bound implies that

$$\begin{aligned}\|\theta-\mu\| &\leq 3.45\sqrt{\log(2/\gamma)} + \sqrt{1.1d}\sqrt{\left(\left(1 + \frac{3\beta+\sqrt{\gamma}}{\sqrt{0.9}}\right)^2 - 1\right)} \\ &\leq 3.45\sqrt{\log(2/\gamma)} + 0.1\sqrt{d} \\ &\leq 15 + 0.1\sqrt{d}.\end{aligned} \tag{76}$$

We would like to set the parameters such that $\sqrt{1.1d}\sqrt{\left(\left(1 + \frac{3\beta+\sqrt{\gamma}}{\sqrt{0.9}}\right)^2 - 1\right)} = 0.1\sqrt{d}$. We can easily see that it implies $3\beta + \sqrt{\gamma} = 0.0045$. For example, we can pick $\beta = 0.0014$ and $\gamma = 9 \times 10^{-8}$. Recall that

$$\beta = \exp(\varepsilon)\alpha + (\exp(\varepsilon) - 1) + (4R+2)\delta + \frac{1}{(4R+2)d}\exp\left(-2(2R+1)^2 d\right). \tag{77}$$

For instance, by setting $\varepsilon \leq 49 \times 10^{-5}$, $\alpha \leq 49 \times 10^{-5}$, $\delta \leq \frac{2}{3} \times 10^{-4}$ and $d \geq 2$, we obtain $\beta \leq 0.1$. Finally, we need to set d such that Equation (73) holds. We can see that $d \geq 7050$ satisfies this condition. □

H Details of the Numerical Experiment

Our goal in the experiments is to evaluate the impact of increasing the radius of the initial feasible set, i.e. R, on the performance of our proposed algorithm and compare it with DPGD. Also, we want to show that our method without any additional hyperparmeter tuning can achieve a good excess error.

Data Generation. Let n denote the number of samples. We assume that $0.9n$ of the data is distributed as follows: let $\mu \in \mathbb{R}^d$ be a uniformly random vector within $\mathcal{S}_{d-1}(50)$. We then sample $0.9n$ of the data points from $\mathcal{N}(\mu, (0.01)^2 \cdot \mathbb{I}_d)$. The remaining $0.1n$ of the points are sampled uniformly at random from $\mathcal{B}_d(100)$.

Hyperparameters. We set the discretization parameter to $r = 0.05$ in Algorithm 2 and failure probability to 5%. Additionally, we repeat each algorithm 10 times and report the mean. For the other hyperparameters, we used exactly the same hyperparameters as stated in Algorithm 3. For DPGD, we use the hyperparameters in Lemma A.1, and in particular, we choose T such that $\frac{\sqrt{2}}{\sqrt{T}} = \frac{16\sqrt{d}}{n\sqrt{\rho}}$.

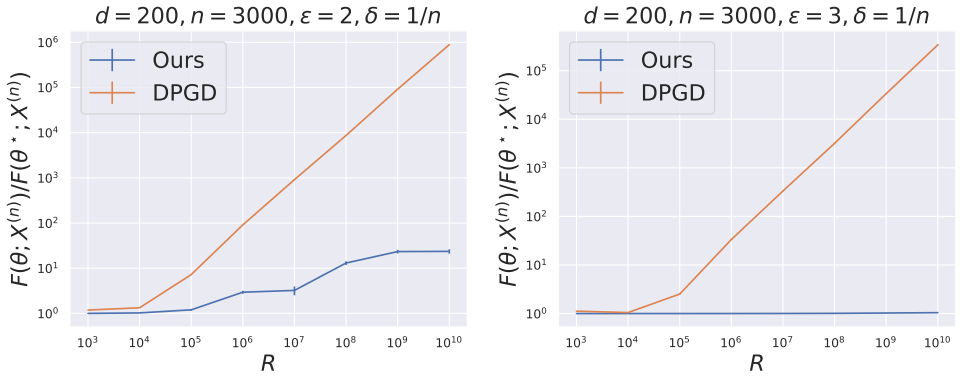


Figure 2: Performance for Different Privacy Budget