Metric-Aware Private Approximate Near Neighbors

Martin Aumüller
IT University of Copenhagen
Copenhagen, Denmark
maau@itu.dk

Nikolaj Munk Binder Jensen
IT University of Copenhagen
Copenhagen, Denmark
nimj@itu.dk

Abstract

We study Approximate Near Neighbor search in the local model of Differential Privacy. We prove a lower bound on the error of standard LDP mechanisms and propose LocalTop-1, a new data structure to identify approximate near neighbors in high-dimensional vector spaces. We analyze its privacy guarantees using the notion of Extended Differential Privacy (Fernandes et al., ESORICS'21), provide theoretical bounds on its performance, and present a short experimental evaluation that highlights its applicability.

Keywords

approximate near neighbor search, extended differential privacy, local differential privacy

1 Introduction

Many data analysis tasks in a collection of high-dimensional vectors require efficient subroutines to retrieve close vectors to a given query vector. For example, in density-based clustering [13, 23], clusterings are formed around core points. Identifying these core points requires finding dense areas via a range search.

In many applications, these vectors are derived from sensitive user information, making nearest neighbor search potentially harmful. Recent systems like PACMANN [27] often assume a centralized model where a trusted server holds the data and privacy is only guaranteed for the queries. In this paper, we shift the perspective to the local model, where users consider their own vectors private and are reluctant to share them with a server. In our model, each user i holds a private vector $\mathbf{x}_i$. Together, these form a dataset $S = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$. Building on recent work in the central model [2, 5], we design a data structure constructed from locally privatized outputs that allows to store and search user identifiers. Given a query vector $\mathbf{q}$ and a similarity threshold α, the search returns a set of identifiers of users who are estimated as lying above the similarity threshold. We measure *accuracy* by the fraction of true nearest neighbors, and measure *error* by the number of false positives.

In this paper, we (i) show a lower bound on error for the general (ε, δ)-LDP setting (Section 3), (ii) design a metric-DP mechanism to store and publish a collection of user identifiers, permitting search for approximate near neighbors (Section 4), (iii) analyze the utility-privacy tradeoff of the proposed data structure (Section 5) implemented using the Exponential Mechanism (Section 6), and (iv) experimentally evaluate it for different parameter choices (Section 7). Related work is presented in Section 9.

Compared to previous work on private high-dimensional near neighbor counting [2, 5], the local setting offers the distinct advantage that users never share their raw vectors with any central entity. This allows our data structure to store and retrieve actual user identifiers, enabling downstream applications like nearest neighbor search or clustering that require specific identities rather than just aggregate counts. However, as is typical in the local model, this stronger privacy guarantee comes at the cost of higher error.

2 Preliminaries

2.1 Problem Statement

Let n and d be two integers. User vectors are on the unit sphere $\mathbb{S}^{d-1} := \{\mathbf{x} \in \mathbb{R}^d \mid \|\mathbf{x}\|_2 = 1\}$, and we will use $[n] := \{1, \ldots, n\}$. We will measure similarity by the inner product:

$$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{1 \le i \le d} x_i y_i \in [-1, 1].$$

We work in the local model of differential privacy, where users apply a DP mechanism to their own data before releasing it for analysis. We have n users, each holding a private vector $\mathbf{x}_i \in \mathbb{S}^{d-1}$. Let S be the collection of these vectors. Given these users, a privacy parameter ε, and a similarity threshold $\alpha \in [0, 1]$, the task is to construct a data structure $\mathcal{DS}$ from the output of locally executed DP mechanisms run by each user that supports an operation $\texttt{search}\colon \mathbb{S}^{d-1} \to 2^{[n]}$. $\mathcal{DS}.\texttt{search}(\mathbf{q})$ returns a subset of user identifiers. Let $B_\alpha(\mathbf{q}, S) := \{i \in [n] \mid \langle \mathbf{q}, \mathbf{x}_i \rangle \ge \alpha\}$. We say that a search is *P-accurate* if each identifier $i \in B_\alpha(\mathbf{q}, S)$ is contained in $\mathcal{DS}.\texttt{search}(\mathbf{q})$ with probability at least P. As is standard in approximate near neighbor search, we will enforce a gap between results that we wish to retrieve, and those we deem too far away. Given another threshold $\beta < \alpha$, we let $\bar{B}_\beta(\mathbf{q}, S) = [n] \setminus B_\beta(\mathbf{q}, S)$ denote the users that have inner product less than β. The *error* of the search is $|\mathcal{DS}.\texttt{search}(\mathbf{q}) \cap \bar{B}_\beta(\mathbf{q}, S)|$. We let (α, β, P)-ANN be the problem of building a data structure that supports P-accurate searches, measuring the quality of the data structure by the expected error.

2.2 Differential Privacy

In addition to solving (α, β, P)-ANN, our data structure must also satisfy some notion of privacy. We assume that the perturbed user vectors stored in $\mathcal{DS}$ are public, so an adversary should not be able to infer a user's original vector by analyzing the public dataset. Furthermore, our setting assumes that identifiers are public, so membership in the dataset itself is not considered to be a sensitive property.[1]

In this paper, we use the notion of *differential privacy* (DP). Specifically, we concern ourselves with the *local* (or distributed) model of DP, where a user's raw information is never communicated to anyone else, and where privacy is measured in the distinguishability of outputs for two different user inputs.

Definition 2.1 (Local differential privacy)**.** *A randomized algorithm $\mathcal{M}\colon \mathcal{X} \to \mathcal{Y}$ satisfies* (ε, δ)-local differential privacy *if and only if*

[1] As a real-world example of this, consider a dating application: a user's profile and pictures are public to others, but their precise location should be kept secret.

for every output $Y \subseteq \mathcal{Y}$ and every pair of input values $x, x' \in \mathcal{X}$,

$$\Pr[\mathcal{M}(x) \in Y] \leq e^{\varepsilon} \Pr[\mathcal{M}(x') \in Y] + \delta.$$

When $\delta = 0$ we say that $\mathcal{M}$ satisfies ε-LDP.[2]

As we show in Section 3, the strict privacy guarantees of LDP necessitate high error in the output. In the case of range queries, it is natural to consider weaker notions of privacy, such as *extended differential privacy* [3, 14], where the privacy loss of a mechanism is allowed to vary depending on the similarity of two inputs.

Definition 2.2 (Extended differential privacy [14])**.** *Given two functions $\xi\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$ and $\delta\colon \mathcal{X} \times \mathcal{X} \to [0, 1]$, a randomized algorithm $\mathcal{M}\colon \mathcal{X} \to \mathcal{Y}$ satisfies (ξ, δ)*-extended differential privacy (XDP) *if for all $x, x' \in \mathcal{X}$ and all $Y \in \mathcal{Y}$,*

$$\Pr[\mathcal{M}(x) \in Y] \leq e^{\xi(x,x')} \Pr[\mathcal{M}(x') \in Y] + \delta(x, x').$$

We assume that ξ is a dissimilarity function, hence the privacy guarantee increases with the similarity of x and x'; more similar items are harder to distinguish. When both ξ and δ are constant, we recover the LDP definition; similarly, (ξ, δ)-XDP implies (ε, δ)-LDP for $\varepsilon = \arg\max_{x,x' \in \mathcal{X}} \xi(x, x')$. As such, we mainly focus on XDP in this paper.

DP mechanisms have many properties which make for convenient analysis. In particular, LDP and XDP are preserved under postprocessing and composition.

Lemma 2.1 (Post-processing [12])**.** *Consider an (ε, δ)-DP mechanism $\mathcal{M}\colon \mathcal{X} \to \mathcal{Y}$ and an arbitrary function $f\colon \mathcal{Y} \to \mathcal{Z}$. The composition $f \circ \mathcal{M}$ is also (ε, δ)-DP.*

Lemma 2.2 (Composition of mechanisms [12])**.** *Let $\mathcal{M}_i\colon \mathcal{X} \to \mathcal{Y}_i$ be a mechanism satisfying $(\varepsilon_i, \delta_i)$-DP for $i \in [k]$, and define*

$$\mathcal{M}_{[k]}(x) := \big(\mathcal{M}_1(x), \ldots, \mathcal{M}_k(x)\big).$$

$\mathcal{M}_{[k]}$ is $\big(\sum_{i=1}^{k} \varepsilon_i, \sum_{i=1}^{k} \delta_i\big)$-DP. In particular, the k-fold composition of an (ε, δ)-DP mechanism satisfies $(k\varepsilon, k\delta)$-DP.

Closure under post-processing is a particularly desirable property in the local setting; it means that once we have constructed a private dataset D using a (ξ, δ)-private mechanism, any algorithm used to analyze D is also automatically (ξ, δ)-private. This gives us a great deal of latitude in how we design the `search` procedure for our ANN data structure.

Finally, we will make use of the *exponential mechanism* (EM) which is a natural choice for selection under XDP [9, 16, 26], since its privacy loss depends on the sensitivity of its utility function.

Definition 2.3 (Exponential mechanism (EM) [12, 19])**.** *Consider a utility function $u\colon \mathcal{X} \times \mathcal{R} \to \mathbb{R}$. The exponential mechanism $\mathtt{EM}(x, u, \mathcal{R})$ outputs an element from $\mathcal{R}$ such that for each $r \in \mathcal{R}$ we have:*

$$\Pr[\mathtt{EM}(x, u, \mathcal{R}) = r] = \frac{\exp\left(\frac{\varepsilon}{2} u(x, r)\right)}{\sum_{r' \in \mathcal{R}} \exp\left(\frac{\varepsilon}{2} u(x, r')\right)}.$$

[2]We note that there exists a related privacy definition called (ε, δ)-*probabilistic* DP (pDP), which states that $\mathcal{M}$ must satisfy ε-DP with probability at least $1 - \delta$. It is well-known that (ε, δ)-pDP implies (ε, δ)-DP, and so it is common to prove (ε, δ)-DP via proving (ε, δ)-pDP. We use this strategy for the analysis of the $\mathcal{M}_{\text{IP}}$ mechanism in Section 6.1.

Lemma 2.3 (Adapted from [12])**.** *Given arbitrary inputs $x, y \in \mathcal{X}$ and output domain $\mathcal{R}$, let $D_u(x, y) := \sup_{r \in \mathcal{R}} |u(x, r) - u(y, r)|$ denote the sensitivity of u w.r.t. inputs x and y. If the bound $D_u(x, y)$ holds with probability at least $1 - \delta$, then the exponential mechanism $\mathtt{EM}(\cdot, u, \mathcal{R})$ satisfies $(\varepsilon D_u(x, y), \delta)$-XDP.*

3 An XDP Lower Bound for (α, β, P)-ANN

The following theorem motivates the study of a mechanism's privacy properties in terms of XDP, in which privacy guarantees smoothly degrade with the distance of points, versus the strictly stronger notion of LDP.

Given an (α, β)-ANN query $\mathcal{DS}.\mathtt{search}(\mathbf{q})$, we use the term "close point" to describe a point $\mathbf{x}_\alpha$ such that $\langle \mathbf{q}, \mathbf{x}_\alpha \rangle \geq \alpha$ and the term "far point" to describe a point $\mathbf{x}_\beta$ where $\langle \mathbf{q}, \mathbf{x}_\beta \rangle < \beta$. We let the *false positive rate* (FPR) denote the probability of erroneously including a far point (i.e., the type I error of `search` and let *false negative rate* (FNR) denote the probability of failing to include a close point (type II error). We then define the tradeoff function [11]

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

Lemma 3.1 (Adapted from [11])**.** *Let $\mathcal{M}$ be a mechanism with* FNR $= a$, FPR $= b$. *$\mathcal{M}$ satisfies (ε, δ)-XDP if and only if $f_{\varepsilon,\delta}(a) \leq b$. Equivalently, $\mathcal{M}$ satisfies (ε, δ)-XDP if and only if $f_{\varepsilon,\delta}(b) \leq a$.*

$f_{\varepsilon,\delta}$ then bounds the feasible region of FPR, FNR of an (ε, δ)-DP mechanism. This lets us perform simple accuracy analysis without knowing the precise behavior of a mechanism. For example, if we focus on LDP only, then a data structure for (α, β, P) must incur large error.

Theorem 3.2. *Let $S \subseteq \mathbb{S}^{d-1}$, and let $0 \leq \beta < \alpha \leq 1$. Any data structure constructed from an (ε, δ)-LDP mechanism $\mathcal{M}$ that solves (α, β, P)-ANN on S has expected error at least $|\bar{B}_\beta(\mathbf{q}, S)| f_{\varepsilon,\delta}(1 - P)$ for each $\mathbf{q} \in \mathbb{S}^{d-1}$.*

Proof. Fix the mechanism $\mathcal{M}$ and a query $\mathbf{q}$. Let $\mathcal{DS}$ be a data structure built from the output of running $\mathcal{M}$ for all users. For user i, we let $\mathbf{x}$ be its vector, and we assume that $\langle \mathbf{x}, \mathbf{q} \rangle \geq \alpha$.

Since $\mathcal{DS}$ solves (α, β, P)-ANN, we have FNR $\leq 1 - P$. Then because $\mathcal{M}$ is (ε, δ)-LDP, Lemma 3.1 states that if i had any other vector $\mathbf{y}$, regardless of its inner product with $\mathbf{q}$, the probability of including $\mathbf{y}$ must be at least $f_{\varepsilon,\delta}(1 - P)$. The result holds by linearity of expectation over the set of user vectors with inner product at most β. □

For example, if $\varepsilon \leq 1, \delta = 0$ and $P \approx 1$, then *any* point in S, including any far point, is returned with probability at least $e^{-1} \approx 0.37$. This is in stark contrast to the central setting in which an absolute error can be achieved [5]. On the other hand, input-sensitive XDP allows for a more nuanced accuracy bound.

Lemma 3.3. *Consider a mechanism $\mathcal{M}$ satisfying (ξ, δ)-XDP. Assume some query point $\mathbf{q} \in \mathbb{S}^{d-1}$ and a close point $\mathbf{x}_\alpha \in S$ such that $\langle \mathbf{x}_\alpha, \mathbf{q} \rangle \geq \alpha$. Define the set of far points $B := \big\{\mathbf{y} \in S \mid \langle \mathbf{q}, \mathbf{y} \rangle < \beta\big\}$.*

Let $g(x; \mathbf{y})$ be defined as $f_{\xi(\mathbf{x}_\alpha, \mathbf{y}), \delta(\mathbf{x}_\alpha, \mathbf{y})}$ from (1). Then if $\mathcal{M}$ solves (α, β, P)-ANN on S, it has expected error at least $\sum_{\mathbf{y} \in B} g(1 - P; \mathbf{y})$.

Proof. This follows directly from Lemma 3.1. □

Algorithm 1: LocalTop1 construction procedure

Data: n, number of users; ε, δ, privacy parameters; m, τ, tensoring parameters; $\mathcal{M}$, a mechanism from $\mathbb{S}^{d-1} \times (\mathbb{R}^d)^m$ to $[m]$.

$\mathcal{A}_1, \ldots, \mathcal{A}_\tau \leftarrow (\mathbf{a}_{1,1} \ldots, \mathbf{a}_{1,m}), \ldots, (\mathbf{a}_{\tau,1} \ldots, \mathbf{a}_{\tau,m})$;

▷ `All` $\mathbf{a}_{i,j} \overset{\text{iid}}{\sim} \mathcal{N}(0,1)^d$

$T \leftarrow$ empty hash table with default value {};

for ***each*** *user* $i \in [n]$ **do**

  ▷ `Local computation carried out by user` i

  $\mathbf{x}_i \leftarrow$ user i's private vector;

  $(j_1, \ldots, j_\tau) \leftarrow (\mathcal{M}(\mathbf{x}_i, \mathcal{A}_1), \ldots, \mathcal{M}(\mathbf{x}_i, \mathcal{A}_\tau))$;

  send $(j_1, \ldots, j_\tau)$ to server;

  ▷ `Central aggregation`

  $T[(j_1, \ldots, j_\tau)] \leftarrow T[(j_1, \ldots, j_\tau)] \cup \{i\}$;

return $D = (T, \mathcal{A}_1, \ldots, \mathcal{A}_\tau)$

Algorithm 2: LocalTop1 query procedure

Data: $D = (T, \mathcal{A}_1, \ldots \mathcal{A}_\tau)$, data structure; $\mathbf{q} \in \mathbb{S}^{d-1}$, query point; η, search threshold.

for $i \in [\tau]$ **do**

  $B_i \leftarrow \{j \in [m] \mid \langle \mathbf{q}, \mathbf{a}_{i,j} \rangle \geq \eta\}$;

$\mathcal{I} \leftarrow B_1 \times \ldots \times B_\tau$; ▷ `Cartesian product`

return $\bigcup_{b \in \mathcal{I}} T[b]$

While this guarantee is data dependent and does not permit as clean an error bound as Theorem 3.2, it reflects the reasonable assumption that including a very dissimilar point in a query result is more harmful than a point close to the β threshold. If $\varepsilon \leq 1, \delta = 0$, and $P \approx 1$ as in our example before, but additionally we have two far points $\mathbf{y}_\beta$ and $-\mathbf{q}$ such that, e.g., $\xi(\mathbf{x}_\alpha, \mathbf{y}_\beta) = \varepsilon$ and $\xi(\mathbf{x}_\alpha, -\mathbf{q}) = 2\varepsilon$, then $\mathbf{y}_\beta$ is still included with probability at least $e^{-\varepsilon} \approx 0.37$, but now the lower bound on the probability of including the maximally dissimilar point $-\mathbf{q}$ is instead $e^{-2\varepsilon} \approx 0.14$. We show an example of how we can apply Lemma 3.3 in Section 6.

4 Top-1 Data Structure

We now describe the LocalTop-1 data structure for finding approximate near neighbors in the local model. The data structure is based on the Top-1 data structure of Aumüller et al. [5] for the central model. The main difference is that each user runs a local mechanism $\mathcal{M}$ to choose a signature given the public random vectors. This local choice poses the main challenge in the analysis and has to be taken into account in the design of the search algorithm.

Algorithm 1 describes the construction procedure for LocalTop-1, assuming a suitable selection mechanism $\mathcal{M} \colon \mathbb{S}^{d-1} \times (\mathbb{R}^d)^m \to [m]$. It requires two user parameters $m, \tau \geq 1$, to be discussed later. First, the server generates τ sets of random vectors $\mathcal{A}_1, \ldots, \mathcal{A}_\tau$, where each $\mathcal{A}_i$ is a set of m i.i.d. samples from $\mathcal{N}(0,1)^d$. Each user i applies $\mathcal{M}$ to their private vector $\mathbf{x}_i$ and $\mathcal{A}_1, \ldots, \mathcal{A}_\tau$ to obtain a tuple of indices $(\mathcal{M}(\mathbf{x}_i, \mathcal{A}_1), \ldots, \mathcal{M}(\mathbf{x}_i, \mathcal{A}_\tau))$, which is sent to the server. The server then stores the user identifier using a hash table T.

Algorithm 2 describes the query procedure for LocalTop-1. Given a query point $\mathbf{q} \in \mathbb{S}^{d-1}$ and a search threshold η, the server first computes the sets $B_1, \ldots, B_\tau$ of indices for which the inner product with $\mathbf{q}$ exceeds η. The server then returns the user identifiers stored in T corresponding to the Cartesian product of these sets.

Aumüller et al. [5] proved the following theorem (adapted to our notation) regarding the utility of the Top-1 data structure in the non-private setting.[3]

Theorem 4.1 ([5], Theorem 19). *Let $\mathcal{S} = \{x_i\}_{i=1,\ldots,n}$ be a dataset in $\mathbb{S}^{d-1}$ and let $0 \leq \beta < \alpha < 1$. Define $\rho = \frac{(1-\alpha^2)(1-\beta^2)}{(1-\alpha\beta)^2}$ and let $m = \lceil n^{\rho/\tau} \rceil$ with $\tau = \left\lceil \frac{\log^{1/8} n}{1-\alpha^2} \right\rceil$. Let*

$$\eta = \alpha\sqrt{2\log m} - \sqrt{2(1-\alpha^2)\log\log m}.$$

`CloseTop-1` *uses space $O(d \cdot n)$, preprocessing time $d \cdot n^{1+o(1)}$, and expected query time $d \cdot n^{\rho+o(1)}$. It is $1 - o(1)$-accurate and has error at most $n^{\rho+o(1)}$.*

In [5], the data structure is applied to differentially private range counting. The idea is to privatize the data structure by replacing the sets in T with their cardinality, and adding Laplace noise to each count. [5] show that this mechanism satisfies (ε, δ)-DP, and that the *additional* error due to noise is at most $n^{\rho+o(1)} \log(1/\delta)/\varepsilon$.

The basic insight into why this data structure works (we show this formally in Theorem 5.1) is the following: Let $\mathbf{a} \sim \mathcal{N}(0,1)^d$, and let $\mathbf{x}, \mathbf{y}$ be two points such that $\langle \mathbf{x}, \mathbf{y} \rangle = \alpha$. Lemma A.1 (in the appendix) states that if $\langle \mathbf{x}, \mathbf{a} \rangle = X$, then $\langle \mathbf{y}, \mathbf{a} \rangle \sim \mathcal{N}(\alpha X, 1 - \alpha^2)$. With high probability, the vector $\mathbf{a} = \mathcal{M}(\mathbf{x}, \mathcal{A}_j)$ will have inner product around $\sqrt{2\log m}$ with $\mathbf{x}$. This means that $\langle \mathbf{y}, \mathbf{a} \rangle$ is normally distributed with a mean around $\alpha\sqrt{2\log m}$, and variance $1 - \alpha^2$. For $\eta = \alpha\sqrt{2\log m} - \sqrt{2(1-\alpha^2)\log\log m}$, $D.\mathtt{search}(\mathbf{y}, \eta)$ will inspect the points that are associated with $\mathbf{a}$ with probability $1 - o(1)$, and thus find $\mathbf{x}$. The error analysis follows along the same line using that the expected value for a filter storing a far away point is at most $\beta\sqrt{2\log m}$ with variance $1 - \beta^2$. This makes it unlikely to exceed the threshold η.

We note that while previous analyses often rely on asymptotic bounds, we will use the exact CDF of the standard normal distribution, denoted by $\Phi(t)$. Similarly, we use Φ^{-1} to denote the inverse CDF of the standard normal distribution.

5 The Framework and its Analysis

We next present a framework for analyzing the privacy and utility guarantees of a general class of mechanisms that can be used together with LocalTop-1. The framework is based on the notion of a "sensitive" mechanism, which is defined as follows.

Definition 5.1. *Given an integer $m \geq 1$, let $\mathcal{A}$ be some collection of standard Gaussian random vectors $(\mathbf{a}_1, \ldots, \mathbf{a}_m)$ with $\mathbf{a}_i \sim \mathcal{N}(0,1)^d$. Consider a selection mechanism $\mathcal{M} \colon \mathbb{S}^{d-1} \times (\mathbb{R}^d)^m \to [m]$. We say for $0 \leq \beta < \alpha \leq 1$ that $\mathcal{M}$ is $(\eta, p_\alpha, p_\beta)$-*sensitive *if for all $\mathbf{x}, \mathbf{y} \in \mathbb{S}^{d-1}$, the following holds for $X = \langle \mathbf{x}, \mathbf{a}_I \rangle$ with $I = \mathcal{M}(\mathbf{x}, \mathcal{A})$:*

- *If $\langle \mathbf{x}, \mathbf{y} \rangle \geq \alpha$, then $\Pr[X \geq \eta] \geq p_\alpha$.*
- *If $\langle \mathbf{x}, \mathbf{y} \rangle < \beta$, then $\Pr[X \geq \eta] \leq p_\beta$.*

[3] This setting would be equivalent to defining $\mathcal{M}(\mathbf{x}_i, \mathcal{A}_j)$ as picking a random vector in $\{\mathbf{a} \in \mathcal{A}_j \mid \sqrt{2\log m} - o(1) \leq \langle \mathbf{a}, \mathbf{x}_i \rangle \leq \sqrt{2\log m}\}$.

We often abuse notation and let $\mathcal{M}(\mathbf{x}, \mathcal{A})$ refer to the *vector* selected by $\mathcal{M}$ instead of its index. Using this definition, we can summarize the quality of LocalTop-1 as follows:

Theorem 5.1. *Let* LT1 *be an instance of LocalTop-1 using mechanism $\mathcal{M}$ as its selection mechanism and with parameters m, τ being positive integers. If $\mathcal{M}$ is $(\eta, p_\alpha, p_\beta)$-sensitive, then* LT1 *is p_α^τ-accurate and has expected error at most $n \cdot p_\beta^\tau$.*

Proof. By construction of the query procedure in Algorithm 2, identifier i is only included in the query result if $\langle \mathcal{M}(\mathbf{x}_i, \mathcal{A}_j), \mathbf{q} \rangle \geq \eta$ for all $j \in [m]$. Since all $\mathcal{A}_j$ are i.i.d., this happens with probability $\prod_{j=1}^{\tau} \Pr[\langle \mathcal{M}(\mathbf{x}_i, \mathcal{A}_j), \mathbf{q} \rangle \geq \eta]$ which by definition of $(\eta, p_\alpha, p_\beta)$-sensitivity is at least p_α^τ when $\langle \mathbf{x}_i, \mathbf{q} \rangle \geq \alpha$.

The argument for the worst-case expected error of LT1 is similar: If $\langle \mathbf{x}_i, \mathbf{q} \rangle < \beta$, then i is included in the result with probability at most p_β^τ. In the worst case this is the case for all n data points, hence expected error is at most np_β^τ. □

It follows from Theorem 5.1 that we can obtain P-accuracy for LocalTop-1 by using a $(\eta, P^{1/\tau}, p_\beta)$-sensitive selection mechanism which must then also provide our desired level of privacy.

While this framework naturally permits quite detailed analysis of a mechanism's output distribution, it turns out that we can achieve useful bounds on its accuracy as long as it satisfies basic probabilistic guarantees.

Lemma 5.2. *Consider a selection mechanism $\mathcal{M}$. For $q,\ p_t \in (0,1)$ and for some threshold $t \in \mathbb{R}$, let $\eta = \alpha t - \sqrt{1-\alpha^2}\Phi^{-1}(q)$. If we have $\Pr[\langle \mathcal{M}(\mathbf{x}, \mathcal{A}), \mathbf{x} \rangle \geq t] \geq p_t$, then for $\delta \in (0,1)$ we have with probability at least $1-\delta$ that $\mathcal{M}$ is $(\eta, p_\alpha, p_\beta)$-sensitive for*

$$p_\alpha = q \cdot p_t, \quad p_\beta = Q_1 + Q_2(1 - p_t),$$

where

$$Q_1 = \Pr[Z_1 \geq \eta], \quad Z_1 \sim \mathcal{N}(\beta\sqrt{2\log(m/\delta)}, 1-\beta^2),$$
$$Q_2 = \Pr[Z_2 \geq \eta], \quad Z_2 \sim \mathcal{N}(\beta t, 1-\beta^2).$$

Proof. We will use $\mathbf{a}$ to denote the vector $\mathcal{M}(\mathbf{x}, \mathcal{A})$. Let $\mathcal{E}$ be the event that $\langle \mathbf{a}, \mathbf{y} \rangle \geq \eta$ and let $\mathcal{T}$ be the event that $\langle \mathbf{a}, \mathbf{x} \rangle \geq t$.

We first show that $\Pr[\mathcal{E} \mid \langle \mathbf{x}, \mathbf{y} \rangle \geq \alpha] \geq p_\alpha = q \cdot p_t$. Since $\Pr[\mathcal{E}]$ is increasing in $\langle \mathbf{x}, \mathbf{y} \rangle$, we assume without loss of generality that $\langle \mathbf{x}, \mathbf{y} \rangle = \alpha$. Taking the probability over the coin flips of $\mathcal{M}$, we use a union bound to see that

$$\Pr[\mathcal{E}] \geq \Pr[\mathcal{E} \mid \mathcal{T}] \Pr[\mathcal{T}] \geq \Pr[\mathcal{E} \mid \mathcal{T}] \cdot p_t.$$

Let us assume without loss of generality that $(\langle \mathbf{x}, \mathbf{a} \rangle \mid \mathcal{T}) = t$. Then by Lemma A.1, $(\langle \mathbf{y}, \mathbf{a} \rangle \mid \mathcal{T})$ has distribution $\mathcal{N}(\alpha t, 1-\alpha^2)$. It follows that setting $\eta = \alpha t - \sqrt{1-\alpha^2}\Phi^{-1}(q)$ ensures that

$$\Pr[\mathcal{E} \mid \mathcal{T}, \langle \mathbf{x}, \mathbf{y} \rangle = \alpha] \cdot p_t \geq q \cdot p_t.$$

We next show that $\Pr[\mathcal{E} \mid \langle \mathbf{x}, \mathbf{y} \rangle < \beta] \leq p_\beta = Q_1 + Q_2(1 - p_\alpha)$. Without loss of generality we assume $\langle \mathbf{x}, \mathbf{y} \rangle = \beta$. Taking the probability over the coin flips of $\mathcal{M}$, we use the law of total probability to state

$$\begin{aligned}\Pr[\mathcal{E}] &= \Pr[\mathcal{E} \mid \mathcal{T}] \Pr[\mathcal{T}] + \Pr[\mathcal{E} \mid \mathcal{T}^c] \Pr[\mathcal{T}^c] \\ &\leq \Pr[\mathcal{E} \mid \mathcal{T}] + \Pr[\mathcal{E} \mid \mathcal{T}^c](1 - p_t).\end{aligned}$$

By using Lemma A.2 with a union bound over $\mathcal{A}$, we see that

$$\Pr\left[\max_{\mathbf{a} \in \mathcal{A}} \langle \mathbf{x}, \mathbf{a} \rangle \leq \sqrt{2\log(m/\delta)}\right] = \Phi\left(\sqrt{2\log(m/\delta)}\right)^m \geq 1 - \delta.$$

Since $\Pr[\mathcal{E}]$ is increasing in $\langle \mathbf{x}, \mathbf{a} \rangle$, we get that with probability $1-\delta$ or greater,

$$\Pr[\mathcal{E} \mid \mathcal{T}] \leq \Pr[\mathcal{E} \mid \langle \mathbf{x}, \mathbf{a} \rangle = \sqrt{2\log(m/\delta)}],$$
$$\Pr[\mathcal{E} \mid \mathcal{T}^c] \leq \Pr[\mathcal{E} \mid \langle \mathbf{x}, \mathbf{a} \rangle = t].$$

We then use Lemma A.1 to obtain $Q_1 = \Pr[\mathcal{E} \mid \mathcal{T}, \langle \mathbf{x}, \mathbf{y} \rangle = \beta]$ and $Q_2 = \Pr[\mathcal{E} \mid \mathcal{T}^c, \langle \mathbf{x}, \mathbf{y} \rangle = \beta]$. □

5.1 Time, space, and communication complexity

The performance of our algorithm is summarized as follows:

Theorem 5.3. *Assume LocalTop-1 uses the parameter choices defined in Theorem 5.1, and further assume that $\mathcal{M}$ has a running time of $O(m)$. LocalTop-1 has the following guarantees:*

- *Algorithm 1 runs in time $O(nm\tau \cdot d)$ and uses $O(n + m\tau d)$ words of space.*
- *In Algorithm 1, each user sends $O(\tau \log m)$ bits to the server.*
- *Given a search threshold η, Algorithm 2 has an expected query time of $O((Pm)^\tau + OUT)$ where $P = 1 - \Phi(\eta)$ and OUT is the expected number of points returned.*
- *In particular, if the choice of m, τ, η satisfies $1 - \Phi(\eta) = o\left(\frac{n^{1/\tau}}{m}\right)$, it will inspect a sublinear amount of buckets. This always holds for $m^\tau = o(n)$.*

6 Implementation via Exponential Mechanism

In the following, we describe a specific $(\eta, p_\alpha, p_\beta)$-sensitive selection mechanism $\mathcal{M}_{\mathrm{IP}}$ and analyze its properties. It works simply by assigning output likelihood to each random vector in $\mathcal{A}$ according to its inner product with the input $\mathbf{x}$.

Definition 6.1. *We define mechanism $\mathcal{M}_{\mathrm{IP}}$ to be* EM$(\cdot, u_{\mathrm{IP}}, \mathcal{A})$ *with utility function u_{IP} defined as follows:*

$$u_{\mathrm{IP}}(\mathbf{x}, \mathbf{a}) := \frac{\langle \mathbf{x}, \mathbf{a} \rangle}{\sqrt{2\log(2m/\delta)}}.$$

6.1 Mechanism privacy

We first show that the privacy loss of $\mathcal{M}_{\mathrm{IP}}$ depends on the distance between points. We first need the following technical lemma.

Lemma 6.1. *Let $\mathcal{A}$ be a set of m i.i.d. standard Gaussian vectors and consider two arbitrary unit vectors, $\mathbf{x}, \mathbf{y} \in \mathbb{S}^{d-1}$. Taking the results of u_{IP} over $\mathcal{A}$, with probability at least $1-\delta$, $D_{u_{\mathrm{IP}}}(\mathbf{x}, \mathbf{y}) \leq \|\mathbf{x} - \mathbf{y}\|$.*

Proof. Recall that for output domain $\mathcal{A} = \mathbf{a}_1, \ldots, \mathbf{a}_m$ where each $\mathbf{a}_i \sim \mathcal{N}(0,1)^d$, utility function u_{IP} has sensitivity

$$D_{u_{\mathrm{IP}}}(\mathbf{x}, \mathbf{y}) = \sup_{\mathcal{A}} \sup_{i \in [m]} \left| u_{\mathrm{IP}}(\mathbf{x}, \mathbf{a}_i) - u_{\mathrm{IP}}(\mathbf{y}, \mathbf{a}_i) \right|.$$

Let us first bound $|u_{\mathrm{IP}}(\mathbf{x}, \mathbf{a}_i) - u_{\mathrm{IP}}(\mathbf{y}, \mathbf{a}_i)|$ for a single sample $\mathbf{a}_i$. Let $X = \langle \mathbf{x}, \mathbf{a}_i \rangle, Y = \langle \mathbf{y}, \mathbf{a}_i \rangle$. Without loss of generality, we assume $X \geq Y$ since the reverse holds by symmetry. Then we have

$$|u_{\mathrm{IP}}(\mathbf{x}, \mathbf{a}_i) - u_{\mathrm{IP}}(\mathbf{y}, \mathbf{a}_i)| = \frac{X - Y}{\sqrt{2\log(m/\delta)}}.$$

According to Lemma A.4, $X - Y$ is distributed as $\mathcal{N}(0, \|\mathbf{x}-\mathbf{y}\|^2)$, and we use Lemma A.3 to get $|X - Y| \le \|\mathbf{x}-\mathbf{y}\| \sqrt{2\log(2/\delta')}$ with at least probability $1-\delta'$, implying

$$|u_{\mathrm{IP}}(\mathbf{x}, \mathbf{a}_i) - u_{\mathrm{IP}}(\mathbf{y}, \mathbf{a}_i)| \le \|\mathbf{x}-\mathbf{y}\| \sqrt{\log(2/\delta')/\log(2m/\delta)}.$$

We now need this bound to hold for all m samples in $\mathcal{A}$, so applying a union bound over $\mathcal{A}$ we have that with probability at least $1 - m\delta'$,

$$D_{u_{\mathrm{IP}}}(\mathbf{x}, \mathbf{y}) \le \|\mathbf{x}-\mathbf{y}\| \sqrt{\log(2/(\delta' m))/\log(2m/\delta)}.$$

The result follows by setting $\delta' = \delta/m$, since then

$$\sqrt{\log(2/(\delta' m))/\log(2m/\delta)} = \sqrt{\log(2m/\delta)/\log(2m/\delta)} = 1.$$

□

Lemma 6.2. *$\mathcal{M}_{\mathrm{IP}}$ satisfies $(\varepsilon \|\mathbf{x}-\mathbf{y}\|, \delta)$-XDP.*

Proof. This follows immediately from Lemmas 2.3 and 6.1. □

Lemma 6.3. *LocalTop-1 using $\mathcal{M}_{\mathrm{IP}}$ satisfies $(\varepsilon\tau \|\mathbf{x}-\mathbf{y}\|, \tau\delta)$-XDP.*

Proof. This follows from Lemma 2.2.

It follows directly from Lemma 6.3 that we can make LocalTop-1 satisfy $\varepsilon \|\mathbf{x}-\mathbf{y}\|$-XDP simply by instantiating $\mathcal{M}_{\mathrm{IP}}$ with privacy parameters ε/τ and δ/τ.

6.2 Mechanism accuracy

$\mathcal{M}_{\mathrm{IP}}$ can be shown to be $(\eta, p_\alpha, p_\beta)$-sensitive, permitting easy analysis of its statistical utility.

Theorem 6.4. *Define*

$$P(\varrho) = 1 - \Phi(\eta - \gamma\varrho) \quad \text{with} \quad \gamma = \varepsilon/(2\sqrt{2\log(2m/\delta)}).$$

Then $\mathcal{M}_{\mathrm{IP}}$ is $(\eta, p_\alpha, p_\beta)$-sensitive with

$$p_\alpha = P(\alpha) - O(1/\sqrt{m}), \quad p_\beta = P(\beta) + O(1/\sqrt{m}).$$

The proof of Theorem 6.4 has been deferred to Appendix B.

Using this result, we can derive a search threshold to achieve a desired sensitivity. This allows us to use $\mathcal{M}_{\mathrm{IP}}$ with LocalTop-1.

Lemma 6.5. *For $0 \le p_\alpha \le 1$ and*

$$\eta = \gamma\alpha - \Phi^{-1}(p_\alpha) - O(1/\sqrt{m}) \text{ with } \gamma = \varepsilon/(2\sqrt{2\log(2m/\delta)}),$$

mechanism $\mathcal{M}_{\mathrm{IP}}$ is $(\eta, p_\alpha, p_\beta)$-sensitive with

$$p_\beta = 1 - \Phi\big((\alpha-\beta)\gamma - \Phi^{-1}(p_\alpha) - O(1/\sqrt{m})\big) + O(1/\sqrt{m}).$$

Proof. Both results follow simply by substituting the stated value of η into the result from Theorem 6.4. For p_α, we have

$$\begin{aligned} P(\alpha) &\ge 1 - \Phi\big(\gamma\alpha - \Phi^{-1}(p_\alpha) - O(1/\sqrt{m}) - \gamma\alpha\big) - O(1/\sqrt{m}) \\ &= 1 - \Phi\big(-\Phi^{-1}(p_\alpha) - O(1/\sqrt{m})\big) - O(1/\sqrt{m}) \\ &= p_\alpha + O(1/\sqrt{m}), \end{aligned}$$

hence $p_\alpha \le P(\alpha) - O(1/\sqrt{m})$. For p_β, the result comes from substituting η into $p_\beta = P(\beta) + O(1/\sqrt{m})$. □

Since $\mathcal{M}_{\mathrm{IP}}$ is $(\varepsilon \|\mathbf{x}-\mathbf{y}\|, \delta)$-XDP, we can use the fact that close and far points are separated by a gap to bound the probability of finding a far point.

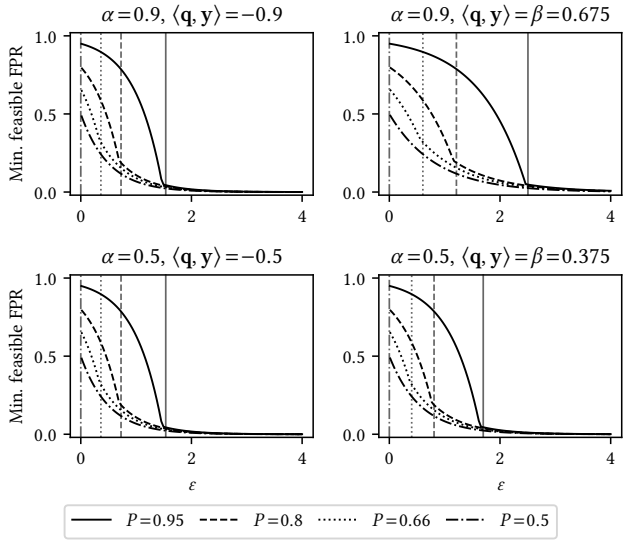

Figure 1: Lower bounds on the minimal required probability of including a far point $\mathbf{y}$ when an $(\varepsilon \|\mathbf{x}-\mathbf{y}\|, \delta)$-XDP mechanism solves (α, β, P)-ANN with fixed $P \in \{0.95, 0.8, 0.66, 0.5\}$. Results are shown for $\alpha \in \{0.9, 0.5\}$ with $\beta = \frac{3}{4}\alpha$. As characterized in Lemma 6.6, the FPR lower bound for each (P, α) lies between the left-column curve ($\langle \mathbf{q}, \mathbf{y}\rangle = \langle \mathbf{q}, -\mathbf{x}\rangle$) and the right-column curve ($\langle \mathbf{q}, \mathbf{y}\rangle = \beta$). Vertical lines indicate the value of ε at which the lower bound equals $1-P$ for each value of P.

Lemma 6.6. *Consider a data structure $\mathcal{DS}$ for (α, β, P)-ANN. Let*

$$D = \sqrt{2 - 2\alpha\beta + 2\sqrt{(1-\alpha^2)(1-\beta^2)}}.$$

Then for query point $\mathbf{q}$, let $\mathbf{y}$ be a far point such that $\langle \mathbf{y}, \mathbf{q}\rangle < \beta$. If $\mathcal{DS}$.search is $(\varepsilon \|\mathbf{x}-\mathbf{y}\|, \delta)$-XDP, then the probability of including $\mathbf{y}$ in the result is at least Q, where for $f_{\cdot,\delta}$ defined as in (1), we have

$$f_{2\varepsilon,\delta}(1-P) \le Q < f_{\varepsilon D,\delta}(1-P).$$

Proof. This follows from Lemma 3.3. By the assumption that $\mathcal{DS}$ is P-accurate, it has FNR at most $1-P$. Let $\mathbf{x}$ be a point such that $\langle \mathbf{x}, \mathbf{q}\rangle = \alpha$. Then the distance $\|\mathbf{x}-\mathbf{y}\|$ depends on $\langle \mathbf{y}, \mathbf{q}\rangle$. Since function $f_{\varepsilon\|\mathbf{x}-\mathbf{y}\|,\delta}(1-P)$ is increasing in $\|\mathbf{x}-\mathbf{y}\|$, it suffices to bound the interval of distances for which $\langle \mathbf{y}, \mathbf{q}\rangle < \beta$.

The lower bound on $\|\mathbf{x}-\mathbf{y}\|$ is found by the fact that $\mathbf{y}$ furthest from $\mathbf{x}$ when $\mathbf{y} = -\mathbf{x}$, and so $\|-\mathbf{x}+\mathbf{x}\| = \|\mathbf{x}+\mathbf{x}\| = 2$.

For the upper bound, we may assume $\langle \mathbf{y}, \mathbf{q}\rangle = \beta$. We must then choose $\mathbf{x}, \mathbf{y}$ such that $\|\mathbf{x}-\mathbf{y}\|$ is maximized. Without loss of generality, let us consider the two-dimensional case with $\mathbf{q} = (1, 0)$. Then the distance between $\mathbf{x}$ and $\mathbf{y}$ is maximal when $\mathbf{x}_2$ and $\mathbf{y}_2$ have different signs, i.e., when $\mathbf{x} = (\alpha, \sqrt{1-\alpha^2})$, $\mathbf{y} = (\beta, -\sqrt{1-\beta^2})$. Then we have

$$\begin{aligned} \|\mathbf{x}-\mathbf{y}\| &= \sqrt{(\alpha-\beta)^2 + (\sqrt{1-\alpha^2} + \sqrt{1-\beta^2})^2} \\ &= \sqrt{2 - 2\alpha\beta + 2\sqrt{1-\alpha^2}\sqrt{1-\beta^2}}. \end{aligned}$$

□

Fig. 1 shows an example of the FPR lower bound of a mechanism for (α, β, P)-ANN which satisfies $(\varepsilon \|\mathbf{x}-\mathbf{y}\|, \delta)$-XDP (this includes LocalTop-1 using $\mathcal{M}_{\text{IP}}$) given different values of P. Very distant far points may be much less likely to be found by a query than points at the β threshold. Additionally, a higher separation between α and β permits a lower FPR for the same FNR.

7 Experiments

We evaluate LocalTop-1 using the $\mathcal{M}_{\text{IP}}$ mechanism for (α, β, P)-ANN in a simple experiment[4] against a baseline using the Gaussian mechanism $\mathcal{M}_G$ [8], calibrated so that both LocalTop-1 and $\mathcal{M}_G$ satisfy $(\varepsilon \|\mathbf{x}-\mathbf{y}\|, \delta)$-XDP. See Appendix C for details about the $\mathcal{M}_G$ baseline.

The purpose of this experiment is to compare the statistical accuracy of outputs produced by each mechanism. While the space and time complexity of LocalTop-1 depends on our parameter choices below, they are not a direct concern in this case and so are not included in the comparison.

7.1 Setup

We performed experiments for the following datasets.

- As a real-world example, we use the SMS Spam Collection [1] corpus of 5 574 SMS messages labelled as spam (13%) or ham (87%). We embed the messages on the unit sphere using `all-MiniLM-L6-v2` [20, 24], a sentence transformer producing 384-dimensional unit-length vectors. We set $\alpha = .50$ and $\beta = .30$ as thresholds. We hold out 20 randomly selected spam messages as queries and build the corpus from the remaining 5,554 messages. At the chosen thresholds, on average 97.1% of items above the α-threshold are genuine spam.
- We generate two instances of Adversarial, a synthetic, adversarially constructed dataset with

  $$n = 100\,000, d = 16, (\alpha, \beta) \in \{(0.9, 0.5), (0.5, 0.3)\}.$$

  For each α, β, we choose a single query point $\mathbf{q}$ and sample 1000 "close" vectors and $n - 1000$ "far" vectors such that

  $$\langle \mathbf{q}, \mathbf{x}_i \rangle \in \begin{cases} [\alpha, \alpha + 0.01], & i \in [1, 1000], \\ [\beta - 0.01, \beta], & i \in [1001, n]. \end{cases}$$

For each dataset, we ran LocalTop-1 using the following sets of parameters:

- Privacy parameters $\varepsilon \in (0, 10], \delta = 1/n$.
- Tensoring parameters: we define a "linear" set of parameters, $(m, \tau) \in \{(\lceil n^{1/\tau} \rceil, \tau) \mid \tau \in \{1, 2, 3\}\}$ and a "sublinear" set $(m, \tau) \in \{(\lceil n^{0.8/\tau} \rceil, \tau) \mid \tau \in \{1, 2, 3\}\}$. According to Theorem 5.3, LocalTop-1 will have expected linear and expected sublinear query times, respectively, for these choices.
- Query threshold η is set according to Lemma 6.5 where we set $p_\alpha = P^{1/\tau}$ for $P = 0.75$. As previously discussed, this ensures that all the LocalTop-1 mechanisms are 0.75-accurate.

Query results were evaluated on their FNR (number of missed close points relative to the actual number of close points) and FPR (number of included far points relative to the actual number of far points). In each dataset, each mechanism was run three times and its results were averaged.

[4] Code for these experiments can be found at https://github.com/nikolajmunk/localtop1.

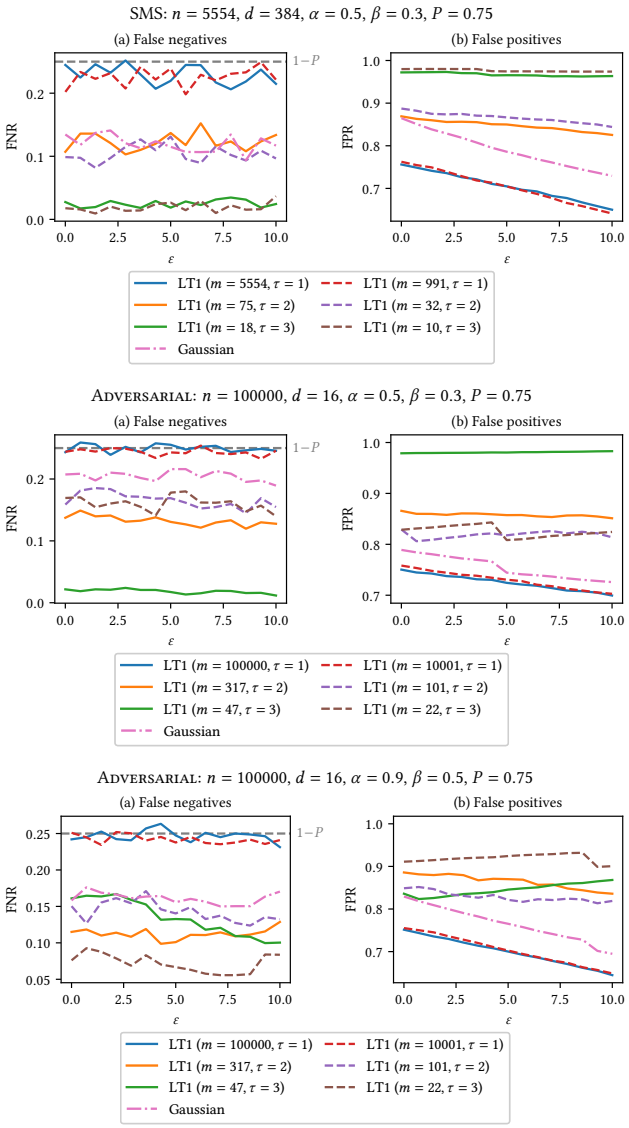

Figure 2: False negative rate (FNR, left) and false positive rate (FPR, right) as functions of ε for the LocalTop-1 (LT1) mechanism under different parameter settings (m, τ), compared against a Gaussian mechanism baseline. For LT1, solid curves correspond to parameter choices with $m^\tau = \Theta(n)$, while dashed curves correspond to $m^\tau = o(n)$. All mechanisms are P-accurate for $P = 0.75$ and satisfy $(\varepsilon \|\mathbf{x}-\mathbf{y}\|, \delta)$-XDP.

7.2 Results

Figure 2 shows that all evaluated mechanisms successfully maintain the target P-accuracy requirement across all three datasets, while exhibiting substantially different FPR behavior.

First, mechanisms sharing the same tensoring parameter τ exhibit similar behavior even when the number m of random vectors differs substantially (with the notable exception of $\tau = 3$). These results suggest that the tensoring level affects empirical accuracy much more than the choice of m, at least for our fairly naive choices of m. This clear stratification by τ is particularly visible for the SMS Spam Collection dataset, where both FNR and FPR results are neatly grouped by τ with accuracy decreasing with the size of τ. This indicates that sublinear query time can be achieved without significantly degrading empirical accuracy.

The Gaussian mechanism baseline generally performs between the $\tau = 1$ and $\tau = 2$ variants across all datasets. In particular, the $\tau = 1$ mechanisms consistently achieve lower FPR than the $\mathcal{M}_G$ baseline, while higher-τ mechanisms tend to incur larger false positive rates. This suggests that with moderate tensoring, LocalTop-1 performs similarly to the baseline while still achieving substantially lower query time complexity.

The strong dependence on τ can likely be understood through the multiplicative structure of the query mechanism. Since a point is found only if all τ sub-queries succeed, maintaining an overall success probability of P requires each sub-query to be $(\eta, P^{1/\tau}, p_\beta)$-sensitive. As discussed in Sections 3 and 6, this necessarily increases the probability of false positives.

Indeed, the experiments reveal a strong inverse relationship between FNR and FPR, which is consistent with the analysis in this paper. With P fixed by construction, increasing ε primarily results in a lower FPR, though this decrease is fairly slow.

The Adversarial datasets provide additional insights into how much geometry affects LocalTop-1's accuracy. Since each of these datasets consists of two "rings" of points at inner product $\approx \alpha$ and $\approx \beta$ w.r.t $\mathbf{q}$, we expect both FNR and FPR to be as high as possible. Indeed most mechanisms seem to exhibit a worse FNR and FPR than for the SMS Spam Collection dataset. Interestingly, there does not seem to be a conclusive difference in mechanism performance for the two different values of α, β, indicating that a more systematic analysis of threshold choice is needed.

Finally, although the empirical results broadly match our theoretical predictions, all evaluated mechanisms (including the Gaussian baseline) perform noticeably worse than the theoretically optimal FNR/FPR frontier derived earlier. This gap suggests that both $\mathcal{M}_{\text{IP}}$ and $\mathcal{M}_G$ are non-optimal mechanisms for private ANN, and that there remains room for significant improvements in the practical design of such mechanisms.

8 Conclusion

We have studied the problem of private (α, β)-approximate near neighbor search through the lens of LocalTop-1, a flexible, random projection based data structure for (α, β)-ANN. We presented a general framework for analyzing the statistical utility of LocalTop-1 based on its selection mechanism, and we showed how to use the XDP guarantees of a mechanism to derive nontrivial lower bounds on its error. As part of this work, we presented a selection mechanism, $\mathcal{M}_{\text{IP}}$, for use with LocalTop-1.

We evaluated several variants of LocalTop-1 against a baseline against both real-world and adversarial datasets. These experiments suggest that although LocalTop-1 is able to match the baseline performance, possibly even with sublinear query times, the error incurred by both privacy-preserving noise and the randomness of the data structure itself means that it is currently not practical for actual use.

Future work will need to address this impracticality, possibly by developing stronger bounds on the necessary error of such mechanisms. One possible avenue of further research is to find more sophisticated heuristics for choosing m and τ. For example, it seems likely that calibrating m to the scale of ε may lead to improvements in space/time requirements and/or statistical utility.

9 Related Work

Fernandes et al. [14] propose distance estimators using Locality-Sensitive Hashing [15] in the extended differential privacy model. Stausholm [25] propose a differentially-private estimator based on the Johnson-Lindenstrauss transform. For set similarity, [6, 18] propose local differentially private estimators for Jaccard similarity. The use of extreme value theory in algorithm design for similarity search has also been applied by [7, 21, 22]. Solutions to the problem of using the exponential mechanism when its selection space has high-density areas have been proposed using weighted or truncated utility functions [9, 16] and preprocessing to select from a suitable subset of the original space [10]. The concept of privacy guarantees degrading with distance is also central to Geo-Indistinguishability [3], which protects location data. Our experimental evaluation uses a similar approach by calibrating ε to provide specific privacy guarantees at a fixed distance.

PACMANN [27] and other approaches [4, 17] focus on a different setting in which only queries have to be answered in a differentially private way or using cryptographic primitives; user vectors in the database are public information. Despite good empirical results, these systems do not provide strong theoretical guarantees on the result quality since they make use of heuristic approaches to nearest neighbor search.

References

[1] Tiago A. Almeida, José María Gómez Hidalgo, and Akebo Yamakami. 2011. SMS Spam Collection Dataset. UCI Machine Learning Repository. Available at https://archive.ics.uci.edu/dataset/228/sms+spam+collection.

[2] Alexandr Andoni, Piotr Indyk, Sepideh Mahabadi, and Shyam Narayanan. 2023. Differentially Private Approximate Near Neighbor Counting in High Dimensions. In *NeurIPS*.

[3] Miguel E. Andrés, Nicolás Emilio Bordenabe, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. 2013. Geo-indistinguishability: differential privacy for location-based systems. In *CCS*. ACM, 901–914.

[4] Hilal Asi, Fabian Boemer, Nicholas Genise, Muhammad Haris Mughees, Tabitha Ogilvie, Rehan Rishi, Guy N. Rothblum, Kunal Talwar, Karl Tarbe, Ruiyu Zhu, and Marco Zuliani. 2024. Scalable Private Search with Wally. *CoRR* abs/2406.06761 (2024).

[5] Martin Aumüller, Fabrizio Boninsegna, and Francesco Silvestri. 2025. Differentially Private High-Dimensional Approximate Range Counting, Revisited. In *FORC (LIPIcs, Vol. 329)*. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 15:1–15:24.

[6] Martin Aumüller, Anders Bourgeat, and Jana Schmurr. 2020. Differentially Private Sketches for Jaccard Similarity Estimation. In *SISAP (Lecture Notes in Computer Science, Vol. 12440)*. Springer, 18–32.

[7] Martin Aumüller, Sariel Har-Peled, Sepideh Mahabadi, Rasmus Pagh, and Francesco Silvestri. 2022. Sampling a Near Neighbor in High Dimensions - Who is the Fairest of Them All? *ACM Trans. Database Syst.* 47, 1 (2022), 4:1–4:40.

[8] Borja Balle and Yu-Xiang Wang. 2018. Improving the Gaussian Mechanism for Differential Privacy: Analytical Calibration and Optimal Denoising. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018 (Proceedings of Machine*

Learning Research), Jennifer G. Dy and Andreas Krause (Eds.). PMLR, 403–412. http://proceedings.mlr.press/v80/balle18a.html
[9] Ricardo Silva Carvalho, Theodore Vasiloudis, Oluwaseyi Feyisetan, and Ke Wang. 2023. TEM: High Utility Metric Differential Privacy on Text. In *Proceedings of the 2023 SIAM International Conference on Data Mining, SDM 2023, Minneapolis-St. Paul Twin Cities, MN, USA, April 27-29, 2023*, Shashi Shekhar, Zhi-Hua Zhou, Yao-Yi Chiang, and Gregor Stiglic (Eds.). SIAM, 883–890. https://doi.org/10.1137/1.9781611977653.CH99
[10] Kamalika Chaudhuri, Daniel Hsu, and Shuang Song. 2014. The Large Margin Mechanism for Differentially Private Maximization. In *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger (Eds.), Vol. 27. Curran Associates, Inc.
[11] Jinshuo Dong, Aaron Roth, and Weijie J. Su. 2022. Gaussian Differential Privacy. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 84, 1 (02 2022), 3–37. https://doi.org/10.1111/rssb.12454
[12] Cynthia Dwork and Aaron Roth. 2014. The Algorithmic Foundations of Differential Privacy. *Found. Trends Theor. Comput. Sci.* 9, 3-4 (2014), 211–407. https://doi.org/10.1561/0400000042
[13] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *KDD*. AAAI Press, 226–231.
[14] Natasha Fernandes, Yusuke Kawamoto, and Takao Murakami. 2021. Locality Sensitive Hashing with Extended Differential Privacy. In *ESORICS (2) (Lecture Notes in Computer Science, Vol. 12973)*. Springer, 563–583.
[15] Sariel Har-Peled, Piotr Indyk, and Rajeev Motwani. 2012. Approximate Nearest Neighbor: Towards Removing the Curse of Dimensionality. *Theory Comput.* 8, 1 (2012), 321–350.
[16] Jacob Imola, Shiva Kasiviswanathan, Stephen White, Abhinav Aggarwal, and Nathanael Teissier. 2022. Balancing utility and scalability in metric differential privacy. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence (Proceedings of Machine Learning Research, Vol. 180)*, James Cussens and Kun Zhang (Eds.). PMLR, 885–894. https://proceedings.mlr.press/v180/imola22a.html
[17] Jingyu Li, Zhicong Huang, Min Zhang, Cheng Hong, Jian Liu, Tao Wei, and Wenguang Chen. 2025. Panther: Private Approximate Nearest Neighbor Search in the Single Server Setting. In *CCS*. ACM, 365–379.
[18] Xiaoyun Li and Ping Li. 2023. Differentially Private One Permutation Hashing and Bin-wise Consistent Weighted Sampling. *CoRR* abs/2306.07674 (2023).
[19] Frank McSherry and Kunal Talwar. 2007. Mechanism Design via Differential Privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2007, Providence, RI, USA, October 20-23, 2007, Proceedings*. IEEE Computer Society, 94–103. https://doi.org/10.1109/FOCS.2007.66
[20] Microsoft. 2021. all-MiniLM-L6-v2. https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2.
[21] Ninh Pham. 2021. Simple Yet Efficient Algorithms for Maximum Inner Product Search via Extreme Order Statistics. In *KDD*. ACM, 1339–1347.
[22] Ninh Pham and Tao Liu. 2022. Falconn++: A Locality-sensitive Filtering Approach for Approximate Nearest Neighbor Search. In *NeurIPS*.
[23] Yuan Qiu and Ke Yi. 2025. Approximate DBSCAN under Differential Privacy. *Proc. ACM Manag. Data* 3, 3 (2025), 128:1–128:24. https://doi.org/10.1145/3725265
[24] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, 3980–3990. https://doi.org/10.18653/v1/D19-1410
[25] Nina Mesing Stausholm. 2021. Improved Differentially Private Euclidean Distance Approximation. In *PODS*. ACM, 42–56.
[26] Xinpeng Xie, Chenyang Yu, Yan Huang, Yang Cao, and Chenxi Qiu. 2025. A Decade of Metric Differential Privacy: Advancements and Applications. https://doi.org/10.48550/arXiv.2502.08970 arXiv:2502.08970 [cs].
[27] Mingxun Zhou, Elaine Shi, and Giulia Fanti. 2025. Pacmann: Efficient Private Approximate Nearest Neighbor Search. In *ICLR*. OpenReview.net.

A Useful probabilistic tools

Lemma A.1. *For points $\mathbf{x}, \mathbf{y} \in \mathbb{S}^{d-1}$ such that $\langle \mathbf{x}, \mathbf{y} \rangle = \varrho$ and a random vector $\mathbf{a} \sim \mathcal{N}(0,1)^d$ such that $\langle \mathbf{x}, \mathbf{a} \rangle = X$, we have that $\langle \mathbf{y}, \mathbf{a} \rangle$ is distributed as $\mathcal{N}(\varrho X, 1-\varrho^2)$. Equivalently, $\langle \mathbf{y}, \mathbf{a} \rangle = \varrho X + Z$ where $Z \sim \mathcal{N}(0, 1-\varrho^2)$.*

Lemma A.2 (Gaussian tail bound)**.** *For normal random variable $Z \sim \mathcal{N}(\mu, \sigma^2)$ and $t \geq 0$, we have*

$$\Pr[Z - \mu > t] = \Pr[Z - \mu < -t] \leq e^{-t^2/2\sigma^2}.$$

In particular, if Z is a standard normal variate, then $\Pr[Z \geq t] \leq e^{-t^2/2}$.

Lemma A.3. *For a random variable $Z \sim \mathcal{N}(0, \sigma^2)$ and $0 < \delta < 1$, we have*

$$\Pr[Z > \sigma\sqrt{2\log(1/\delta)}] = \Pr[Z < -\sigma\sqrt{2\log(1/\delta)}] \leq \delta.$$

Proof. This follows by simple algebraic manipulation of the result in Lemma A.2. □

Lemma A.4 (Sensitivity of inner product)**.** *For any $\mathbf{x}, \mathbf{y} \in \mathbb{S}^{d-1}$ and random vector $\mathbf{a} \sim \mathcal{N}(0,1)^d$, we have $\langle \mathbf{x}, \mathbf{a} \rangle - \langle \mathbf{y}, \mathbf{a} \rangle \sim \mathcal{N}(0, \|\mathbf{x} - \mathbf{y}\|_2^2)$.*

Proof. By linearity in first argument, $\langle \mathbf{x}, \mathbf{a} \rangle - \langle \mathbf{y}, \mathbf{a} \rangle = \langle \mathbf{x} - \mathbf{y}, \mathbf{a} \rangle$. This is a linear transformation of a standard Gaussian vector, so $\langle \mathbf{x} - \mathbf{y}, \mathbf{a} \rangle \sim \mathcal{N}(0, \|\mathbf{x} - \mathbf{y}\|_2^2)$. □

B Omitted proofs

B.1 Proof of Theorem 5.3

Proof of Theorem 5.3. We analyze the complexity of the data structure operations as follows:

- **Preprocessing time:** For each insertion of a user vector $\mathbf{x}_i$, $i \in [n]$, the data structure computes $m\tau$ inner products, each taking $O(d)$ time. Thus, the total preprocessing time is $O(nm\tau d)$.
- **Space usage:** The data structure stores the random vectors $\mathcal{A}$ and the hash table T. $\mathcal{A}$ contains $m\tau$ vectors of dimension d, requiring $O(m\tau d)$ space. The hash table T stores one entry per user, requiring $O(n)$ space. The total space is $O(n + m\tau d)$.
- **Communication cost:** Each user i sends the tuple of indices $(\mathcal{M}(\mathbf{x}_i, \mathcal{A}_1), \ldots, \mathcal{M}(\mathbf{x}_i, \mathcal{A}_\tau))$ to the server. Since $\mathcal{M}$ outputs an index in $[m]$, this requires $O(\tau \log m)$ bits per user.
- **Query time:** Given a query $\mathbf{q}$ and threshold η, the algorithm first identifies the candidate sets

  $$B_i = \{j \in [m] \mid \langle \mathbf{q}, \mathbf{a}_{i,j} \rangle \geq \eta\}$$

  for each $i \in [\tau]$. It then iterates over the Cartesian product $B_1 \times \ldots \times B_\tau$. The expected size of each B_i is

  $$m \cdot \Pr_{Z \sim \mathcal{N}(0,1)}[Z \geq \eta] = m(1 - \Phi(\eta)) = mP.$$

  The expected number of buckets to examine is therefore $\prod_{i=1}^{\tau} \mathbb{E}[|B_i|] = (mP)^\tau$. The total query time is proportional to the number of buckets probed plus the number of retrieved items (OUT). If $1 - \Phi(\eta) = o\left(\frac{n^{1/\tau}}{m}\right)$, then $(mP)^\tau = o\left(m\frac{n^{1/\tau}}{m}\right)^\tau = o(n)$, ensuring that the number of inspected buckets is sublinear. This is trivially true when $m^\tau = o(n)$, since then the number of inspected buckets is sublinear regardless of η.

□

B.2 Proof of Theorem 6.4

Proof. Consider arbitrary unit vectors $\mathbf{q}, \mathbf{x}$ such that $\langle \mathbf{q}, \mathbf{x} \rangle = \varrho$. Since the expressions for p_α and p_β are very similar, by definition of $(\eta, p_\alpha, p_\beta)$-sensitivity it suffices to show that

$$P(\varrho) - O(1/\sqrt{m}) \leq \Pr[\langle \mathbf{q}, \mathbf{a}_I \rangle \geq \eta] \leq P(\varrho) + O(1/\sqrt{m}),$$

where $I := \mathcal{M}_{\mathrm{IP}}(\mathbf{x}, \mathcal{A})$ is the index of the vector chosen by $\mathcal{M}_{\mathrm{IP}}$.

Given set of random vectors $\mathcal{A} = \mathbf{a}_1, \ldots, \mathbf{a}_m$, we define

$$X_i := \langle \mathbf{a}_i, \mathbf{x} \rangle, \quad Y_i := \langle \mathbf{a}_i, \mathbf{q} \rangle, \quad W_i := \exp(\gamma X_i).$$

The joint variable (X_i, Y_i) is a bivariate standard normal with covariance ϱ. We note that each W_i is a lognormal random variable with mean $\mu = \mathbb{E}[W_i] = e^{\gamma^2/2}$.

By the definition of $\mathcal{M}_{\mathrm{IP}}$ and the exponential mechanism,

$$\Pr[\mathcal{M}_{\mathrm{IP}}(\mathbf{x}, \mathcal{A}) = i] = \frac{W_i}{\sum_{j=1}^m W_j}.$$

We state our exceedance probability as $\Pr[Y_I \geq \eta]$ and observe that we can express $\Pr[Y_I \geq \eta]$ as

$$\begin{aligned}\Pr[Y_I \geq \eta] &= \sum_{i \in [m] : Y_i \geq \eta} \Pr[\mathcal{M}_{\mathrm{IP}}(\mathbf{x}, \mathcal{A}) = i] \\ &= \frac{\sum_{i \in [m] : Y_i \geq \eta} \exp(\gamma X_i)}{\sum_{j \in [m]} \exp(\gamma X_j)} \\ &= m\mathbb{E}\left[\frac{\exp(\gamma X)}{\sum_{j \in [m]} \exp(\gamma X_j)} \middle| Y \geq \eta \right] \cdot \Pr[Y \geq \eta],\end{aligned}$$

where (X, Y) is bivariate normal, $(X, Y) \sim \mathcal{N}(0, 0, 1, 1, \varrho)$.

Then for $S_m := \sum_{j \in [m]} W_j$, we have $\mathbb{E}[S_m] = m\mu$.

Since W_j is a lognormal random variable, it has finite third moment. This allows us to use the Berry-Esseen theorem to state that $|S_m - m\mu| = O(\sqrt{m})$. This implies that

$$m\mathbb{E}\left[\frac{e^{\gamma X}}{S_m} \middle| Y \geq \eta \right] = \frac{\mathbb{E}\left[e^{\gamma X} \mid Y \geq \eta \right]}{\mu} \pm O(1/\sqrt{m}).$$

We now compute $\mathbb{E}\left[e^{\gamma X} \mid Y \geq \eta \right]$. We note that since X and Y are jointly Gaussian, $X = \varrho Y + \sqrt{1-\varrho^2} Z$ with $Z \sim \mathcal{N}(0, 1)$ independent of Y. Therefore $\mathbb{E}\left[e^{\gamma X} \mid Y \geq \eta \right] = \mathbb{E}\left[e^{\gamma \sqrt{1-\varrho^2} Z} \right] \mathbb{E}\left[e^{\gamma \varrho Y} \mid Y \geq \eta \right]$. Now Z is a standard normal variate and $(Y \mid Y \geq \eta)$ has a lower-truncated standard normal distribution, so we use their respective MGFs to state

$$\begin{aligned}\mathbb{E}\left[e^{\gamma X} \mid Y \geq \eta \right] &= e^{\gamma^2(1-\varrho^2)/2} e^{\gamma^2 \varrho^2/2} \frac{1 - \Phi(\eta - \gamma\varrho)}{1 - \Phi(\eta)} \\ &= e^{\gamma^2/2} \frac{1 - \Phi(\eta - \gamma\varrho)}{1 - \Phi(\eta)}.\end{aligned}$$

Substituting this back into our expression for $m\mathbb{E}\left[\frac{e^{\gamma X}}{S_m} \middle| Y \geq \eta \right]$, we have

$$\frac{\mathbb{E}\left[e^{\gamma X} \mid Y \geq \eta \right]}{\mu} = \frac{e^{\gamma^2/2} \frac{1 - \Phi(\eta - \gamma\varrho)}{1 - \Phi(\eta)}}{e^{\gamma^2/2}} = \frac{1 - \Phi(\eta - \gamma\varrho)}{1 - \Phi(\eta)}.$$

Then our total expression for $\Pr[Y_I \geq \eta]$ becomes

$$\Pr[Y_I \geq \eta] = \left(\frac{1 - \Phi(\eta - \gamma\varrho)}{1 - \Phi(\eta)} \pm O(1/\sqrt{m}) \right) \cdot \Pr[Y \geq \eta],$$

and noticing that Y is now simply a standard normal random variable, $\Pr[Y \geq \eta] = 1 - \Phi(\eta)$. Hence

$$\Pr[Y_I \geq \eta] = 1 - \Phi(\eta - \gamma\varrho) \pm \frac{O(1/\sqrt{m})}{1 - \Phi(\eta)} = 1 - \Phi(\eta - \gamma\varrho) \pm O(1/\sqrt{m}).$$

□

C Gaussian mechanism baseline

As a point of comparison for LocalTop-1 using $\mathcal{M}_{\mathrm{IP}}$, it is natural to consider the simplest viable mechanism: simply add d-dimensional noise to each input vector and only include its identifier in the query result if the perturbed vector lies within some chosen range of query point $\mathbf{q}$. Specifically, we adapt the Analytic Gaussian mechanism [8] for input perturbation of unit vectors. Given unit vector input $\mathbf{x} \in \mathbb{S}^{d-1}$, we define $\mathcal{M}_G(\mathbf{x}) := \mathbf{x} + Z$ where $Z \sim \mathcal{N}(0, \sigma^2)^d$.

To perform an apples-to-apples-comparison between $\mathcal{M}_G$ and $\mathcal{M}_{\mathrm{IP}}$, we must ensure that they provide the same differential privacy guarantees.

Lemma C.1. *For σ such that*

$$\Phi\left(\frac{1}{\sigma} - \varepsilon\sigma \right) - e^{2\varepsilon} \Phi\left(-\frac{1}{\sigma} - \varepsilon\sigma \right) \leq \delta,$$

$\mathcal{M}_G$ satisfies $(\varepsilon \|\mathbf{x} - \mathbf{y}\|, \delta)$-XDP.

Proof. We first note that by Lemma 3 in [8], the privacy loss of $\mathcal{M}_G$ follows distribution $\mathcal{N}(\gamma, 2\gamma)$ with $\gamma = \frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2}$. If we let $\xi = \varepsilon \|\mathbf{x} - \mathbf{y}\|$, using Lemma 6 in [8], we define

$$\begin{aligned}h(\xi) &= \Phi\left(\frac{\|\mathbf{x}-\mathbf{y}\|}{2\sigma} - \frac{\xi\sigma}{\|\mathbf{x}-\mathbf{y}\|} \right) - e^{\xi} \Phi\left(-\frac{\|\mathbf{x}-\mathbf{y}\|}{2\sigma} - \frac{\xi\sigma}{\|\mathbf{x}-\mathbf{y}\|} \right) \\ &= \Phi\left(\frac{\|\mathbf{x}-\mathbf{y}\|}{2\sigma} - \varepsilon\sigma \right) - e^{\varepsilon\|\mathbf{x}-\mathbf{y}\|} \Phi\left(-\frac{\|\mathbf{x}-\mathbf{y}\|}{2\sigma} - \varepsilon\sigma \right)\end{aligned}$$

and obtain (ξ, δ)-DP if and only if $h(\xi) \leq \delta$ for all possible $\mathbf{x}, \mathbf{y} \in \mathbb{S}^{d-1}$. For fixed ε, function h is monotonically increasing in $\|\mathbf{x} - \mathbf{y}\|$ [8], so to determine a valid σ, we assume the worst case of $\|\mathbf{x} - \mathbf{y}\| = 2$ and solve for σ such that $h(2\varepsilon) \leq \delta$. □

Letting $S' := \{ \mathcal{M}_G(\mathbf{x}_i) \mid i \in [n] \}$ be the set of perturbed user vectors, we can construct a data structure for (α, β, P)-ANN simply by defining the operation $S'.\mathtt{search}(\mathbf{q}) := \{ i \in [n] \mid \langle \mathbf{q}, \mathcal{M}_G(\mathbf{x}_i) \rangle \geq \gamma \}$.

Lemma C.2. *Let $\mathbf{x}, \mathbf{q}$ be unit vectors, $\langle \mathbf{x}, \mathbf{q} \rangle = \varrho$ and let $\mathbf{v} = \mathcal{M}_G(\mathbf{x})$. Then $\Pr[\langle \mathbf{v}, \mathbf{q} \geq \gamma \rangle] = 1 - \Phi((\gamma - \varrho)/\sigma)$.*

Proof. Recall that $\mathbf{v} = \mathbf{x} + Z$, $Z \sim \mathcal{N}(0, \sigma^2)^d$ and therefore $\langle \mathbf{v}, \mathbf{q} \rangle = \langle \mathbf{x}, \mathbf{q} \rangle + \langle Z, \mathbf{q} \rangle$. Now $\langle Z, \mathbf{q} \rangle \sim \mathcal{N}(0, \sigma^2)$, and so $\langle \mathbf{v}, \mathbf{q} \rangle$ is distributed as $\mathcal{N}(\varrho, \sigma^2)$. The result then follows by the fact that $(\langle \mathbf{v}, \mathbf{q} \rangle - \varrho)/\sigma$ is a standard normal variate. □

Lemma C.2 directly lets us choose a threshold η which satisfies our desired P.

Lemma C.3. *Let $\mathbf{x}, \mathbf{q}$ be unit vectors with $\langle \mathbf{x}, \mathbf{q} \rangle = \varrho$ and let $\mathbf{v} = \mathcal{M}_G(\mathbf{x})$. For $0 \leq \beta < \alpha \leq 1$ and $\gamma = \alpha - \sigma \cdot \Phi^{-1}(P)$, the following holds:*

- *If $\varrho \geq \alpha$, then $\Pr[\langle \mathbf{v}, \mathbf{q} \rangle \geq \gamma] \geq P$.*
- *If $\varrho \leq \beta$, then $\Pr[\langle \mathbf{v}, \mathbf{q} \rangle \geq \gamma] \leq 1 - \Phi(\alpha - \beta - \Phi^{-1}(P))$.*

Proof. The results follow by substituting γ into $(\gamma - \varrho)/\sigma$. In the first case,

$$\frac{\gamma - \varrho}{\sigma} = \frac{\alpha - \sigma \cdot \Phi^{-1}(P) - \varrho}{\sigma} \geq \Phi^{-1}(P)$$

and so $1 - \Phi((\gamma - \varrho)/\sigma) \geq P$.

Similarly, in the second case,

$$\frac{\gamma - \varrho}{\sigma} = \frac{\alpha - \sigma \cdot \Phi^{-1}(P) - \varrho}{\sigma} \leq \alpha - \beta - \Phi^{-1}(P),$$

and so $1 - \Phi((\gamma - \varrho)/\sigma) \leq 1 - \Phi(\alpha - \beta - \Phi^{-1}(P))$. □