DP-GRAMS: Differentially Private Gradient Ascent-based Mean Shift for Mode Estimation

Anonymous Author
Anonymous Institution

Abstract

We consider the problem of summarizing multivariate data in a manner that preserves individual-level privacy. Traditional statistics such as the mean capture global trends but often overlook local structures. In contrast, modes offer a more localized summary, making them a compelling alternative. We propose Differentially Private Gradient Ascent-based Mean Shift (DP-GRAMS), a novel, nonparametric algorithm for mode estimation under differential privacy constraints. Our approach is based on reformulating the classic mean-shift procedure as gradient ascent on a kernel density estimator, with each update perturbed by Gaussian noise calibrated to ensure privacy. We prove that DP-GRAMS satisfies (ε, δ)-differential privacy and establish both local and global convergence guarantees. Specifically, when the data are drawn from a smooth distribution with finitely many well-separated modes, the algorithm reliably recovers all true modes with high probability. Our theoretical analysis demonstrates that DP-GRAMS achieves optimal convergence rates with respect to sample size and dimensionality in the low-privacy regime, while also quantifying performance degradation under stricter privacy constraints. We present two natural extensions: DP-PMS, a private modal-regression method, and DP-GRAMS-C, a clustering pipeline. Extensive experiments on synthetic and real data demonstrate favorable privacy–utility trade-offs relative to common baselines.

Preliminary work. Under review by AISTATS 2026. Do not distribute.

1 Introduction

The estimation of modes, which are local maxima of a probability density function, is an important summary of a distribution that retains significant local information, while removing the need to estimate the entire density function. Unlike global summaries such as the mean or median, modes reveal heterogeneous subpopulations and localized concentrations of probability mass, making them indispensable in multimodal or complex settings. Applications span a wide range of domains, from clustering and classification (Avidan (2007); Chen et al. (2016b); Li et al. (2007)) to computer vision tasks such as object tracking and image segmentation (Comaniciu and Meer (2002); Comaniciu et al. (2003)), as well as nonlinear statistical modeling paradigms including manifold learning and modal regression (Einbeck and Tutz (2006); Chen et al. (2016a)). Because modes capture fine structural detail, they are often more informative than averages in uncovering hidden groups, sharp transitions, or localized effects.

From a practical standpoint, computing distribution summaries with sensitive data that routinely arise in domains such as healthcare or finance presents significant privacy challenges. A flurry of recent research has shown that statistical summaries can compromise privacy. Differential privacy (Dwork et al. (2006, 2014)) provides rigorous protection by ensuring that the output of an algorithm is nearly indistinguishable with or without any single individual. Considerable research has developed DP methods for means, regression, and clustering (see surveys Oberski and Kreuter (2020); Dankar and El Emam (2013)), but mode estimation has remained almost entirely unaddressed, despite its centrality in nonparametric statistics. This motivates the current work where we pose the problem of mode estimation in terms of a gradient ascent procedure, and develop practically implementable yet theoretically robust techniques for differentially private mode estimation.

Our mode estimation algorithm is based on the classi-

cal mean shift procedure of Fukunaga and Hostetler (1975); Cheng (1995); Comaniciu and Meer (2002), later adapted to specific settings such as density gradient estimation in Arias-Castro et al. (2016) or modal regression Chen et al. (2016a). Given a kernel K that can be used for density estimation, the mean shift consists of sequential updates according to:

$$m(x) = \frac{\sum_{i=1}^n K\left(\frac{X_i - x}{h}\right) X_i}{\sum_{i=1}^n K\left(\frac{X_i - x}{h}\right)}.$$

By iteratively shifting points toward weighted neighborhood averages, it performs gradient ascent on a related kernel density estimate, while requiring no prior specification of the number or shape of clusters. To preserve differential privacy with the above procedure, our algorithm modifies the update at each step to

$$\tilde{m}(x) = m(x) + z$$

for a suitably scaled random noise vector z that allows the required obfuscation for differential privacy. Note that this procedure is very sensitive to initialization, as the mean shift updates are guaranteed to converge only in local basins of attraction for each mode. We resort to multiple random initializations to tackle this issue. However, this poses another challenge due to privacy leakage from several uses of the same dataset. Fortunately, adding noise vectors correlated with respect to the distance of the respective initializations, as done in Hall et al. (2013), helps to resolve this problem. For the correct choice of bandwidth h, we then have the global convergence of DP-GRAMS, and MSE bounds for each mode μ_j:

$$\begin{aligned} &\mathbb{E}\left[\|x_T - \mu_j\|^2 | X_1, \ldots, X_n\right] \\ &\leq C\left(\frac{\log n}{n}\right)^{4/(d+6)} + \left(\frac{\text{polylog}(n,\delta)}{n^2\varepsilon^2}\right)^{2/(d+3)} \end{aligned}$$

for $1 \leq j \leq k$, with probability at least $1 - 1/n$. Here, the expectation is over the random noise added for privacy. As expected, we find a worsening of the mean squared error (MSE) for stricter privacy budgets ε. The nonparametric nature of the problem leads to the familiar curse of dimensionality in this setting, as seen in (Tsybakov (2008); Arias-Castro et al. (2016); Genovese et al. (2014)). Nevertheless, our real data experiments show that the assumption-lean framework allows us to capture subtle data patterns in clustering problems that classical methods such as DP-K-means (Su et al. (2016)) fail to identify. We also extend our privatized algorithm to clustering and modal regression.

To the best of our knowledge, no prior work develops a differentially private algorithm specifically for mode hunting or mean shift. Our work is related to several elements from the differential privacy literature, where several families of techniques are relevant but do not solve the mode estimation problem.

- **Private density estimation.** Methods based on histograms, orthogonal series (Wasserman and Zhou (2010)), and KDE variants (Hall et al. (2013); Wagner et al. (2023); Liu et al. (2024)) produce private density approximations, but they do not target mode recovery.
- **Private clustering.** Several works are based on the k-means framework (Balcan et al. (2017); Ghazi et al. (2020); Stemmer (2021); Su et al. (2016)) optimize parametric objectives, but are not suited to nonparametric settings or irregular cluster shapes.
- **Private regression.** Most work focuses on linear or generalized linear models (Arora et al. (2022); Alabi et al. (2020); Wang (2018); Sheffet (2017)), leaving nonparametric modal regression unexplored.

Organization. The remainder of the paper is organized as follows. Section 2 reviews some relevant background. Section 3 introduces the DP-GRAMS algorithm. Section 4 states the main theoretical results. Section 5 contains comprehensive empirical evaluation and implementation details. We conclude with a discussion of future work in Section 6. All proofs are deferred to the supplement.

2 Background

This section provides the technical background for DP-GRAMS: kernel density estimation, the mean-shift algorithm and its gradient-ascent interpretation, as well as differential privacy. We retain sufficient detail to make the paper self-contained.

2.1 Kernel Density Estimation

Given i.i.d. samples $X_1, \ldots, X_n \in \mathbb{R}^d$ from density p, the kernel density estimator (KDE)

$$\hat{p}_h(x) = \frac{1}{nh^d} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right), \qquad x \in \mathbb{R}^d,$$

where $K : \mathbb{R}^d \to \mathbb{R}_+$ is a kernel and $h > 0$ is a bandwidth parameter. See, e.g., Tsybakov (2008) for more details.

2.2 The Mean-Shift Algorithm

The Mean shift (MS, Fukunaga and Hostetler (1975); Cheng (1995); Comaniciu and Meer (2002)) is an itera-

tive, nonparametric method for locating local maxima of a density. For a point $x \in \mathbb{R}^d$, the update is

$$m(x) = \frac{\sum_{i=1}^n K\left(\frac{X_i - x}{h}\right) X_i}{\sum_{i=1}^n K\left(\frac{X_i - x}{h}\right)}.$$

Repeatedly setting $x \leftarrow m(x)$ moves the point toward high-density regions and typically converges to a local mode (Yamasaki and Tanaka (2024)). This rationale motivates our private mode hunting algorithm.

Kernel Choice. While our methods can easily be generalized to other kernel choices, we focus on the Gaussian kernel $K(u) = (2\pi)^{-d/2} \exp\left(-\frac{\|u\|^2}{2}\right)$, for several reasons: i) smooth ascent paths (Comaniciu and Meer, 2002, Theorem 2), ii) interpretation as the EM algorithm for Gaussian mixtures (Carreira-Perpinan (2007)), and iii) making mean shift identical to gradient ascent on the logarithm of the KDE (Cheng (1995)).

2.3 Mean Shift as Gradient Ascent

For Gaussian kernels, the gradient of the KDE is

$$\nabla \hat{p}_h(x) = \frac{1}{nh^{d+2}} \sum_{i=1}^n (X_i - x) K\left(\frac{x - X_i}{h}\right)$$

$$m(x) - x = h^2 \nabla \log \hat{p}_h(x) = \; h^2 \nabla \hat{p}_h(x) / \hat{p}_h(x).$$

Thus, each MS step is equivalent to scaled gradient ascent on $\log \hat{p}_h$. Directly privatizing $m(x)$ is infeasible, as its global ℓ_2-sensitivity grows quickly with dimension and data range, forcing noise at a scale that destroys utility. By viewing MS as gradient ascent, however, the update decomposes into per-sample contributions. This structure permits the use of differentially private optimization tools and forms the basis for our DP-GRAMS algorithm.

2.4 Differential Privacy

Differential privacy (DP, Dwork et al. (2006, 2014)) formalizes protection of individual-level information.

Definition 2.1 (Differential Privacy)**.** *A randomized mechanism $\mathcal{M} : \mathcal{X}^n \to \mathcal{Y}$ is (ε, δ)-differentially private if, for all neighboring datasets $\mathbf{X}, \mathbf{X}'$ differing in one record and all measurable $A \subseteq \mathcal{Y}$,*

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

Several classical DP tools will be used in our analysis. The Gaussian mechanism adds noise proportional to the ℓ_2-sensitivity of a function to ensure privacy. Privacy amplification by subsampling improves guarantees when mechanisms are applied to randomly selected subsets of the data (Balle et al. (2018)). Advanced composition quantifies the cumulative privacy loss for adaptive sequences of mechanisms (Dwork et al. (2010)). Finally, post-processing invariance guarantees that any data-independent transformation of a DP output preserves its privacy. Differentially private gradient descent (see, e.g., Abadi et al. (2016)) combines these principles in practice.

3 DP-GRAMS Algorithm

DP-GRAMS estimates the modes of a dataset under differential privacy by performing noisy gradient ascent on a KDE. The steps are detailed in Algorithm 1.

Algorithm 1: DP-GRAMS: Differentially Private Gradient Ascent-based Mean Shift

Input : Data $S = \{X_i\}_{i=1}^n \subset \mathbb{R}^d$, privacy parameters (ε, δ), minibatch size m, steps $T = \lceil \log n \rceil$, bandwidth h, initialization fraction p_0, stepsize $\eta > 0$, gradient clip C_*

Output: Private mode set $\mathcal{M}$

Sample $k = \max(1, \lfloor np_0 \rfloor)$ points from S and denote them as $\mathcal{I} = \{x_{01}, \ldots, x_{0k}\}$.;
$\mathcal{M} \leftarrow \emptyset$;
Compute σ^2 based on $C_*, \varepsilon, \delta, T$;
Compute kernel matrix $\mathbf{K}$ with entries $K_{ij} = \exp(-\|x_{0i} - x_{0j}\|^2 / 2h^2)$ for $1 \leq i, j \leq k$;
Simulate dT correlated Gaussian vectors $z_{j,t} \sim \mathcal{N}(0, \sigma^2 \mathbf{K})$ for $1 \leq j \leq d$, $1 \leq t \leq T$;
for $l \in [k]$ **do**
    Set $x_0 = x_{0l}$; **for** $t = 0$ **to** $T - 1$ **do**
        Sample minibatch $\mathcal{B}_t \subseteq [n]$, $|\mathcal{B}_t| = m$;
        Compute per-sample gradients $q_i(x_t)$ using KDE weights;
        Clip the norm of each $q_i(x_t)$ to C_*;
        Compute mean gradient $\bar{q}(x)$;
        Compute $z_t = (z_{1,t,l}, z_{2,t,l}, \ldots, z_{d,t,l}) \in \mathbb{R}^d$;
        Compute $x_{t+1} = x_t + \eta(\bar{q}(x_t) + z_t)$;
    Add x_T to $\mathcal{M}$;
Merge nearby outputs in $\mathcal{M}$;
return $\mathcal{M}$

Initialization. Let $X_1, \ldots, X_n \in \mathbb{R}^d$ denote the dataset. To initialize candidate modes, we select a random subset $\mathcal{I} \subset \{X_1, \ldots, X_n\}$ by sampling an indicator vector $w \in \{0, 1\}^n$ with independent entries $w_i \sim \text{Bernoulli}(p_0)$, $i \in [n]$. The initialization set is then

$$\mathcal{I} := \{X_i \mid w_i = 1\}.$$

While not required at sufficiently large sample sizes, we have found that practical implementations benefit from filtering $\mathcal{I}$ by the following high-density (HD) filter which removes low-density candidates before the mean shift updates:

$$\mathcal{I}_{\text{HD}} := \left\{ x \in \mathcal{I} \;\middle|\; \hat{p}_h(x) \geq \gamma \cdot \max_{y \in \mathcal{I}} \hat{p}_h(y) \right\}$$

for a filter parameter $\gamma \in (0, 1]$ specified by the user.

Bandwidth Selection. After initializing candidate modes, DP-GRAMS selects the KDE bandwidth h in a privacy-aware manner which is then fixed and used consistently throughout all gradient ascent updates and mode merging steps.

Algorithm Overview. DP-GRAMS iteratively updates candidate modes via a mean-shift-style gradient ascent on the log-KDE:

$$m(x) - x = h^2 \nabla \log \hat{p}_h(x) = \frac{1}{n} \sum_{i=1}^{n} h^2 \frac{g_i(x)}{\hat{p}_h(x)},$$

where

$$g_i(x) := \frac{1}{h^{d+2}} (X_i - x) K\Big(\frac{x - X_i}{h}\Big)$$

for $K(u) = (2\pi)^{-d/2} e^{-\|u\|^2/2}$. The per-sample contribution is defined as

$$q_i(x) := h^2 \frac{g_i(x)}{\hat{p}_h(x)}, \quad i \in [n],$$

and clipped to a fixed norm C_* (see (3)).

Private Gradient Ascent. Each candidate in $\mathcal{I}$ undergoes T iterations of mini-batch gradient ascent on the log-KDE. In each iteration, a mini-batch of size m is sampled uniformly without replacement, and the per-sample contributions $q_i(x)$ are computed and clipped to a fixed norm C_* to ensure bounded sensitivity. The mean of the clipped gradients is then perturbed with Gaussian noise $\mathcal{N}(0, \sigma^2 I_d)$, where σ is calibrated based on the privacy parameters (ε, δ), the batch size m, the number of iterations T, and the sensitivity C_* (see (2)). Candidate modes are updated using a fixed step size $\eta > 0$. After completing T iterations, the algorithm outputs a set of private candidate modes $\mathcal{M}$.

Correlated Noise. The use of multiple initializations implies that the same data is used several times throughout the algorithm, leading to privacy leakage at the post-processing agglomeration stage detailed below. To avoid this, at each step of gradient descent we use noise vectors that are independent across iterations and data dimensions, but correlated across initializations. More specifically, we think of the iterates $(x_0, x_1, \ldots, x_T)$ as a function of x_0 and add suitably correlated noise for two initializations that are close to each other. The exact value of correlation at two initializations x_0 and x_0' is taken to be $\exp(-\|x_0 - x_0'\|^2/2h^2)$, based on the Gaussian kernel. This choice is based on the framework introduced for private kernel density estimation in Hall et al. (2013).

Merging Candidate Modes. Noise in gradient updates and randomness in initialization can produce multiple points corresponding to the same mode. To reduce redundancy, DP-GRAMS applies a post-processing merge. When the number of true modes is known, hierarchical clustering merges $\mathcal{M}$ into the specified number of clusters, replacing each cluster with its mean. When the number of modes is unknown, candidate modes within a multiple of the kernel bandwidth are grouped and replaced with their mean.

4 Theory

Before presenting our theoretical results, we present the required assumptions. In the following, for any radius $r_j > 0$, we write the closed Euclidean ball as

$$\overline{B}(\mu_j, r_j) := \{x \in \mathbb{R}^d : \|x - \mu_j\| \le r_j\}.$$

Assumption 1 (Model assumptions)**.** *Let $d \ge 1$ and let $p : \mathbb{R}^d \to [0, \infty)$ be a probability density. Fix $k \in \mathbb{N}$. Assume that $p(\cdot)$ has k modes at $\mu_1, \ldots, \mu_k$.*

For each $j \in [k]$, there exists $r_j > 0$, such that p is strictly positive on the closed ball $\overline{B}(\mu_j, r_j)$, with

$$p_{\min,j} := \inf_{x \in \overline{B}(\mu_j, r_j)} p(x) > 0.$$

Moreover, $p \in C^3\big(\overline{B}(\mu_j, r_j)\big)$ and there exists a finite constant $M_j > 0$ such that

$$\sup_{x \in \overline{B}(\mu_j, r_j)} \max_{|\alpha| \le 3} \big\|D^\alpha p(x)\big\| \;\le\; M_j,$$

for D^α as the partial derivative of multi-index α.

Assumption 2 (Bandwidth condition)**.** *Let $(h_n)_{n \ge 1}$ be the sequence of bandwidths. We assume*

$$h_n \downarrow 0 \qquad \text{and} \qquad n\, h_n^{d+4} \to \infty \qquad \text{as } n \to \infty.$$

We now present our main results on the privacy and accuracy guarantees of the DP-GRAMS procedure.

4.1 Privacy and Utility Guarantees

The first result concerns the sensitivity of our estimator. Since our algorithm uses gradient ascent, a suitable bound for the gradient is essential to calibrate noise level for differential privacy. See, e.g., Balle et al. (2018); Bassily et al. (2014).

Lemma 1 (Per-Sample Sensitivity Bound)**.** *Under Assumptions 1 and 2, for each $i \in [n]$, the per-sample gradient $q_i(x) = h^2 g_i(x)/\hat{p}_h(x)$ satisfies*

$$\|q_i(x)\| \lesssim \frac{1}{p(x) V_d} h^{1-d}, \quad V_d = \frac{\pi^{d/2}}{\Gamma(1 + d/2)} \tag{1}$$

for every $x \in \overline{B}(\mu_j, r_j)$, with high probability.

Calculating the above sensitivity would require estimating $p(x)$ as x proceeds through the mean shift updates $\{x_t : 0 \le t \le T\}$. However, given an initialization in $\cup_j \overline{B}(\mu_j, r_j)$, each iterate x_t would also be in $\cup_j \overline{B}(\mu_j, r_j)$, and by Assumption 1, we avoid updating $p(x)$, instead opting to bound it below by a constant, due to Assumption 1.

At the t-th step of DP-GRAMS we can then add noise generated as $z_t \sim \mathcal{N}(0, \sigma^2 I_d)$, where

$$\sigma = \frac{C_*/m}{\log(1 + n(e^{\varepsilon_{\text{iter}}} - 1)/m)} \sqrt{8 \log\left(\frac{2.5mT}{n\delta}\right)}, \quad (2)$$

where

$$C_* = \frac{1}{p_{\min} V_d} h^{1-d} \quad \text{and} \quad \varepsilon_{\text{iter}} = \frac{\varepsilon}{\sqrt{8T \log(2/\delta)}}. \quad (3)$$

When $\varepsilon \ll 1$, one can replace $e^{\varepsilon_{\text{iter}}} - 1$ by $\varepsilon_{\text{iter}}$ and $\log(1 + n\varepsilon_{\text{iter}}/m)$ by $n\varepsilon_{\text{iter}}/m$, which yields

$$\sigma_g \approx \frac{8\, C_*}{n\, \varepsilon} \sqrt{T \log\left(\frac{2}{\delta}\right) \log\left(\frac{2.5\, m\, T}{n\, \delta}\right)}. \quad (4)$$

The following theorem guarantees that the DP-GRAMS iterates satisfy (ε, δ) privacy constraint.

Theorem 4.1 (Privacy Guarantee of DP-GRAMS)**.** *Suppose Assumptions 1 and 2 hold. Let x_T be the T-th iterate of Algorithm 1 initialized at $x_0 \in \cup_j \overline{B}(\mu_j, r_j)$, with added noise $z_t \sim \mathcal{N}(0, \sigma^2 I_d)$ at each step, for $1 \le t \le T$ and σ as defined in (2). Then x_T satisfies (ε, δ)-differential privacy.*

We now move on to determine the rate of estimation of the modes through DP-GRAMS. Our results rely on the fact that

$$\ell(x) = \log p(x) \quad \text{and} \quad \hat{\ell}_h(x) = \log \hat{p}_h(x)$$

and the respective first two derivatives would be close. More specifically, let $\mathcal{G}_n^{\text{local},j}$ be the event that

$$\sup_{x \in \overline{B}(\mu_j, r_j)} |\hat{\ell}_h^{(k)}(x) - \ell^{(k)}(x)| < C \left[h^{(2-k)\wedge 2} + \sqrt{\frac{\log n}{n h^{d+2k}}} \right] \quad (5)$$

for some constant $C > 0$ and $k = 0, 1, 2$. Here $\hat{\ell}_h^{(k)}$, $\ell^{(k)}$ denote the k-th derivatives of $\hat{\ell}_h$ and ℓ respectively.

The following additional assumptions ensure that DP-GRAMS iterates, initialized within $\overline{B}(\mu_j, r_j)$ for some j, and a sufficiently small r_j would converge to the mode μ_j. The first condition concerns the Hessian matrix of $\log p(x)$ at the modes μ_j.

Assumption 3 (Hessian conditions)**.**

a. For each $j \in [k]$,

$$\alpha_j := -\lambda_{\max}\big(\nabla^2 \log p(\mu_j)\big) > 0.$$

b. For each $j \in [k]$, there is a constant $L_j > 0$ such that

$$\sup_{x,y \in \overline{B}(\mu_j, r_j)} \|\nabla^2 \log p(x) - \nabla^2 \log p(y)\| \le L_j \|x - y\|.$$

The next assumption presumes that Assumption 1 is satisfied for a suitably small r_j.

Assumption 4 (Radius constraint)**.** *For each $j \in [k]$, Assumption 1 is satisfied for some*

$$r_j \le \min\left\{ \frac{\alpha_j}{2L_j},\ \frac{1}{2} \min_{i \ne j} \|\mu_i - \mu_j\| \right\}.$$

The choice of radius above enables us to control local strong concavity and mode exclusion simultaneously. In particular, we show in the supplement that:

1. $\nabla^2 \log p(x) \preceq -\frac{\alpha_j}{2} I$ for all $x \in \overline{B}(\mu_j, r_j)$
2. $B(\mu_j, r_j)$ contains no other mode μ_i for $i \ne j$.

Theorem 4.2 (Local convergence of DP-GRAMS)**.** *Suppose Assumptions 1, 2, 3, 4 hold, and the initialization satisfies $x_0 \in \overline{B}(\mu_j, r_j)$. Then the output x_T after T steps of* DP-GRAMS *satisfies:*

$$\mathbb{E}\left[\|x_T - \mu_j\|^2 | \mathcal{X}\right] \le C\left(\frac{\log n}{n}\right)^{4/(d+6)} + \left(\frac{\text{polylog}(n, \delta)}{n^2 \varepsilon^2}\right)^{2/(d+3)}$$

for $T = C \log n$ and some numerical constant $C > 0$, provided $\mathcal{X} \in \mathcal{G}_n^{\text{local},j}$, where $\mathcal{G}_n^{\text{local},j}$ is defined in (5).

Our proofs reveal that the theoretically optimal bandwidth happens to be

$$h_{\text{opt}}^{\text{non-DP}} \asymp \left(\frac{\log n}{n}\right)^{1/(d+6)}$$

if the privacy budget ε is sufficiently large, and

$$h_{\text{opt}}^{\text{DP}} \asymp \left(\frac{\text{polylog}(n, \delta)}{n^2 \varepsilon^2}\right)^{1/(2d+6)}$$

otherwise. Note that when ε is large so that the first term in the MSE rate dominates over the noise due to privacy, our MSE rates match the minimax optimal ones derived earlier in Arias-Castro et al. (2016); Genovese et al. (2014).

4.2 Initialization

The error rates derived in the previous section depend crucially on finding an initialization $x_0 \in \overline{B}(\mu_j, r_j)$ for

some $j \in [k]$ and r_j satisfying Assumption 4. We now discuss some strategies for achieving such an initialization and their implications on maintaining differential privacy constraints.

Random initialization: The most natural strategy is to use a subset of points randomly subsampled from the entire dataset. That is, for some pre-fixed $p_0 \in (0,1)$ we sample independent random variables $w_i \sim \text{Bernoulli}(p_0)$ and include the i-th data point in the initialization set $\mathcal{I}$ if and only if $w_i = 1$.

Let us define the in-ball probabilities

$$p_j := \Pr_{X \sim p}\big(X \in \bar{B}(\mu_j, r_j)\big) \quad \text{for} \quad j = 1, \dots, k.$$

The next proposition shows that with high probability, random initialization provides at least one point in $\bar{B}(\mu_j, r_j)$ for every $j \in [k]$.

Proposition 4.3 (Initialization covers all modes)**.** *Under Assumptions 1, 2, 3, 4, with probability at least $1 - n^{-2}$, the initialization pool $\mathcal{I}$ is non-empty and contains at least one point inside each ball $\bar{B}(\mu_j, r_j)$, provided*

$$\frac{n}{\log(kn)} \geq \frac{C}{p_0} \max_{1 \leq j \leq k} \frac{1}{p_j} \tag{6}$$

for some constant $C > 0$.

Now that an initialization is guaranteed we can use local convergence results from Theorem 4.2 to ensure recovery of all modes using DP-GRAMS algorithm. However, using the same data for mean shift updates at each initialization $x_0 \in \mathcal{I}$ leads to privacy loss. To resolve this issue, we use correlated noise vectors to privatize the updates.

Correlated noise: Given two initializations x_0, x_0' the DP-GRAMS algorithm updates use noise vectors $\{z_t(x_0) \in \mathbb{R}^d : 1 \leq t \leq T\}$ and $\{z_t(x_0') \in \mathbb{R}^d : 1 \leq t \leq T\}$ such that the correlation between each component satisfies:

$$\text{Corr}((z_t(x_0))_j, (z_t(x_0'))_j) = \exp(-\|x_0 - x_0'\|^2 / 2h^2)$$

for $1 \leq t \leq T, 1 \leq j \leq d$. The correlation ensures that sufficiently close initializations have very similar noise vectors and thus, protects against privacy leakage due to multiple initializations. Note that we use independent vectors $z_t \sim \mathcal{N}(0, \sigma^2 I_d)$ as $1 \leq t \leq T$, where σ^2 is as defined in (2).

We conclude this section with the following theorem that guarantees privacy of DP-GRAMS as well as its accuracy.

Theorem 4.4 (Global convergence of DP-GRAMS)**.** *Suppose Assumptions 1, 2, 3, 4 hold. Let $\mathcal{I}$ be the initialization pool from Proposition 4.3. Then, with probability at least $1 - n^{-1}$, for every $j \in [k]$ there exists at least one $x_0 \in \mathcal{I}$ such that x_T, the output of Algorithm 1 initialized at x_0 satisfies:*

$$\begin{aligned} &\mathbb{E}\left[\|x_T - \mu_j\|^2 | \mathcal{X}\right] \\ &\leq C\Big(\frac{\log n}{n}\Big)^{4/(d+6)} + \Big(\frac{\text{polylog}(n, \delta)}{n^2 \varepsilon^2}\Big)^{2/(d+3)} \end{aligned}$$

for $T = C \log n$ and some numerical constant $C > 0$.

5 Numerical Experiments

We now illustrate the efficacy of our algorithm on several simulated and real data experiments. We show the applicability of DP-GRAMS in mode hunting and provide its extensions to Private Modal Regression, and Private Clustering.

All experiments use Python libraries with an ARM CPU (8 cores, 8 logical processors) and 8.6 GB RAM, running macOS 15.6.1. All library versions are provided in the supplement. In our clustering experiments, we use DP k-means of Su et al. (2016) via the Holohan et al. (2019) library.

5.1 Experiments on Simulated Data

For all experiments on simulated data, the accuracy of mode estimation is evaluated on the MSE metric

$$\min_{\pi:[k]\to[\hat{k}]} \frac{1}{\max\{k, \hat{k}\}} \sum_{j=1}^{\min\{k,\hat{k}\}} \|\mu_j - \hat{\mu}_{\pi(j)}\|^2$$

where π is an injective map from $[k]$ to $[\hat{k}]$ and $\hat{k}$ is the estimated number of modes. The optimal assignment $(i, \pi(i))$ is calculated by the Hungarian algorithm.

5.1.1 Private Mode Estimation on Simulated Data

Our first set of experiments is on two bivariate distributions:

1. 4-component mixture of isotropic Gaussians:

$$(X, Y) \sim \frac{1}{4} \sum_{k=1}^{4} \mathcal{N}\big(\mu_k, I_2\big),$$

where $\mu_1 = (3,3)$, $\mu_2 = (3,-3)$, $\mu_3 = (-3,3)$, $\mu_4 = (-3,-3)$.

2. 5-component mixture of t distributions:

$$(X, Y) \sim \sum_{k=1}^{5} \pi_k \, t_{\nu_k}\big(\mu_k, \sigma_k^2 I_2\big), \quad \pi_k = 0.2,$$

where the modes are $\mu_1 = (0, 0)$, $\mu_2 = (6, 0)$, $\mu_3 = (-6, 0)$, $\mu_4 = (0, 6)$, $\mu_5 = (0, -6)$, the degrees of freedom are $\nu = [15, 6, 10, 8, 20]$, and the scale parameters are $\sigma = [0.1, 0.9, 1.3, 1.0, 0.4]$.

For each mixture, we run $n_{\text{runs}} = 20$ independent trials. Mode estimation is performed using the classical mean-shift (MS) algorithm and the DP-GRAMS algorithm. Mean-shift operates with Silverman's rule-of-thumb for bandwidth selection and a fixed iteration count of $T = \lceil \log n \rceil$. DP-GRAMS injects Gaussian noise calibrated to satisfy (ε, δ)-differential privacy, processes data in mini-batches of size $m = \lceil n / \log n \rceil$, and adapts the number of iterations $T = \lceil \log n \rceil$ based on the dataset size and privacy parameters. Raw mode estimates from both methods are post-processed via a merging procedure to consolidate nearby modes. Results are plotted in Figure 1.

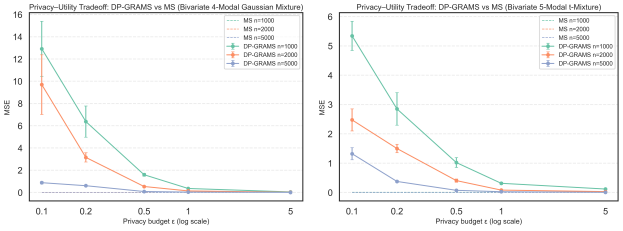

Figure 1: MSE levels at different privacy budgets and sample sizes for bivariate Gaussian (left) and t (right) distributions.

In both Gaussian and t mixture experiments, MSE decreases as the privacy budget ε increases. At the same ε, the MSE of the private estimator decreases with increasing sample size. To visualize mode recovery, we also produce single-run 3D kernel density surfaces with overlaid 2D contour projections in Figure 2, highlighting the close agreement between the private DP-GRAMS estimates and true modes.

5.1.2 Private Modal Regression on Simulated Data

Our second set of experiments builds on DP-GRAMS to design a differentially private modal regression algorithm, which is based on the partial mean shift algorithm (see Chen et al. (2016a)). The algorithm called DP-PMS, for Differentially Private Partial Mean Shift, is summarized in the supplement.

We generate a synthetic regression dataset with three well-separated conditional modes. For a given number of samples per cluster n, the predictor values are drawn uniformly from disjoint intervals: $X_1 \sim \text{Uniform}(0, 0.5)$, $X_2 \sim \text{Uniform}(0.4, 0.7)$, and $X_3 \sim \text{Uniform}(0.6, 1.0)$, which are concatenated into $X = [X_1, X_2, X_3]^\top \in \mathbb{R}^{3n}$. Corresponding responses are

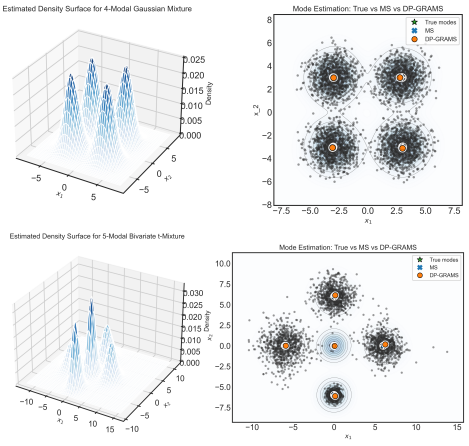

Figure 2: Recovery of Gaussian and t modes for a single run with $n = 4000$ samples. DP-GRAMS recovers all modes even when density values at the modes are different.

sampled from Gaussian distributions centered at distinct modes, $Y_1 \sim \mathcal{N}(\mu_1, \sigma^2)$, $Y_2 \sim \mathcal{N}(\mu_2, \sigma^2)$, $Y_3 \sim \mathcal{N}(\mu_3, \sigma^2)$, with $\mu_1 = 3$, $\mu_2 = 2$, $\mu_3 = 1$, and $\sigma = 0.2$, and concatenated into $Y = [Y_1, Y_2, Y_3]^\top \in \mathbb{R}^{3n}$. Ground-truth modes are computed by the non-private partial mean shift algorithm. For each dataset, we compare three approaches: the vanilla Partial Mean-Shift (PMS) algorithm, the differentially private variant DP-PMS with (ε, δ)-guarantees, and LOWESS as a non-modal reference smoother. DP-PMS is applied on a discrete mesh of predictor values over the set of local samples contributing to the mode estimate.

Mode recovery from a single run are shown in Figure 3.

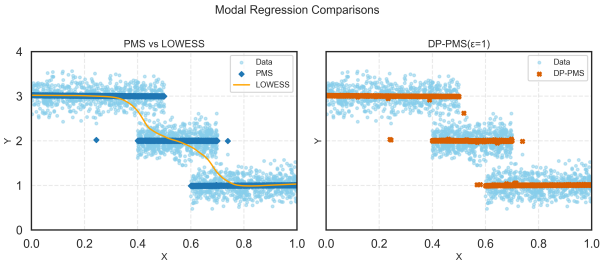

Figure 3: Recovery of modes with non-private partial mean shift and LOWESS (left), and private partial mean-shift via DP-PMS (right) with $n = 2100$ samples divided roughly equally among three clusters. DP-PMS recovers modal structure while exhibiting minor deviations attributable to privacy-preserving noise.

Next, in Figure 4 we see the expected decrease in MSE values for larger privacy budgets ε. At the same ε, the MSE of the private estimator decreases with growing sample size. All experiments are repeated over $n_{\text{runs}} = 10$ independent trials.

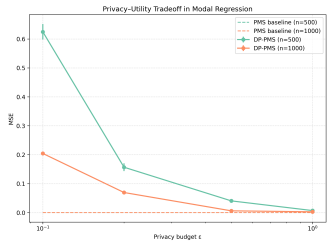

Figure 4: MSE levels at different privacy budgets and sample sizes for modal regression.

5.2 Clustering Experiments on Real Data

In the third set of experiments we study an extension of differentially private mode estimation to clustering problems. More specifically, we use the privately estimated modes to assign cluster labels to all data points, and evaluate the accuracy on standard clustering metrics. MSE values for estimating cluster centers are also calculated. All experiments use well-known benchmark datasets: Iris, Digits, and MNIST.

Note that the Iris dataset has 150 observations on 4 features, with 3 real clusters. The Digits dataset has 1797 observations on 64 features, and MNIST has 70000 observations on 784 features. We reduce the dimension of the MNIST dataset by projecting onto 50 principal components. The other datasets were used as is. Since both Digits and MNIST are classification datasets, we use the corresponding response values $\{0, 1, \ldots, 9\}$ as true cluster labels. We report clustering similarity via Adjusted Rand Index (ARI) here and defer other evaluation metrics to the supplement. MSE values for each algorithm are also reported, along with the corresponding runtimes, in Table (1). We evaluate four methods on the above datasets: non-private Mean-Shift (MS) with agglomerative merging, DP-GRAMS based clustering, standard k-Means, and DP k-Means (from Su et al. (2016)).

The above result is supplemented by a visualization of the different clusters on MNIST data in Figure 5.

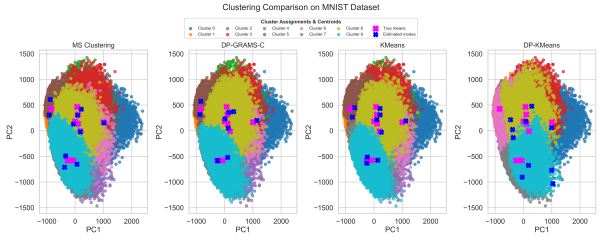

Figure 5: Clustering on MNIST data with two non-private methods and their private analogues. DP-Kmeans does not align with many of the clustering labels, but DP-GRAMS achieves performance similar to the non-private versions.

Data	Method	ARI	MSE	Time (s)
Iris	MS	0.78	0.22	0.01
	DP-GRAMS	0.73	0.07	0.07
	K-means	0.73	0.07	0.04
	DP-Kmeans	0.44	5.14	0.01
Digits	MS	0.71	5.11	2.86
	DP-GRAMS	0.71	5.06	3.67
	K-means	0.67	3.92	0.10
	DP-Kmeans	0.00	2247.70	0.35
MNIST	MS	0.45	37.63	453.20
	DP-GRAMS	0.42	40.50	537.73
	K-means	0.37	39.48	1.02
	DP-Kmeans	0.24	313.22	0.57

Table 1: Clustering similarity (ARI), MSE for cluster centers, and runtimes (in seconds) from non-private and private algorithms. DP-GRAMS at $\varepsilon = 1$ achieves ARI and MSE comparable to non-private methods while ensuring privacy. DP-Kmeans, also at $\varepsilon = 1$, fails to do so.

We also study the performance of the differentially private clustering algorithms across different privacy budgets ε. The curves for MNIST data are given in Figure 6. Visualizations for the other datasets can be found in the supplement.

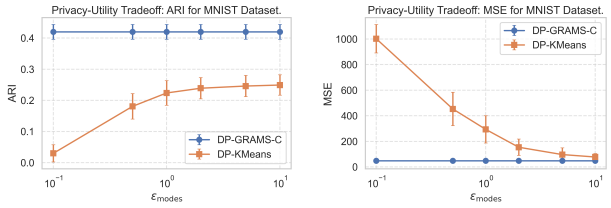

Figure 6: Clustering similarity (ARI) and MSE of cluster centers computed on MNIST data using DP-Kmeans and DP-GRAMS-C for different ε.

6 Conclusion

This paper introduced DP-GRAMS, a differentially private mode-seeking algorithm that leverages the equivalence between mean shift and gradient ascent on the log-density, bringing modern private optimization techniques to nonparametric mode estimation. The key insight is to decompose each mean-shift update into per-sample contributions, clip those contributions to bound sensitivity, and add Gaussian noise calibrated via standard DP accounting. It is worth exploring if our rates can be complemented by minimax lower bounds characterizing the fundamental cost of privacy in this problem. Similarly, exploring kernels with additional structure that adapt to finer properties of densities could improve rates for mode estimation. We intend to explore these in the future.

References

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

Alabi, D., McMillan, A., Sarathy, J., Smith, A., and Vadhan, S. (2020). Differentially private simple linear regression. *arXiv preprint arXiv:2007.05157*.

Arias-Castro, E., Mason, D., and Pelletier, B. (2016). On the estimation of the gradient lines of a density and the consistency of the mean-shift algorithm. *The Journal of Machine Learning Research*, 17(1):1487–1514.

Arora, R., Bassily, R., Guzmán, C., Menart, M., and Ullah, E. (2022). Differentially private generalized linear models revisited. *Advances in neural information processing systems*, 35:22505–22517.

Avidan, S. (2007). Ensemble tracking. *IEEE transactions on pattern analysis and machine intelligence*, 29(2):261–271.

Balcan, M.-F., Dick, T., Liang, Y., Mou, W., and Zhang, H. (2017). Differentially private clustering in high-dimensional euclidean spaces. In *International Conference on Machine Learning*, pages 322–331. PMLR.

Balle, B., Barthe, G., and Gaboardi, M. (2018). Privacy amplification by subsampling: Tight analyses via couplings and divergences. *Advances in neural information processing systems*, 31.

Bassily, R., Smith, A., and Thakurta, A. (2014). Differentially private empirical risk minimization: Efficient algorithms and tight error bounds. *arXiv preprint arXiv:1405.7085*.

Carreira-Perpinan, M. A. (2007). Gaussian mean-shift is an em algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(5):767–776.

Chen, Y.-C., Genovese, C. R., Tibshirani, R. J., and Wasserman, L. (2016a). Nonparametric modal regression. *Annals of Statistics*.

Chen, Y.-C., Genovese, C. R., and Wasserman, L. (2016b). A comprehensive approach to mode clustering. *Electronic Journal of Statistics*, 10(1):210–241. Originally submitted as "Enhanced Mode Clustering" (arXiv:1406.1780, June 2014).

Cheng, Y. (1995). Mean shift, mode seeking, and clustering. *IEEE transactions on pattern analysis and machine intelligence*, 17(8):790–799.

Comaniciu, D. and Meer, P. (2002). Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 24(5):603–619.

Comaniciu, D., Ramesh, V., and Meer, P. (2003). Kernel-based object tracking. *IEEE Transactions on pattern analysis and machine intelligence*, 25(5):564–577.

Dankar, F. K. and El Emam, K. (2013). Practicing differential privacy in health care: A review. *Trans. Data Priv.*, 6(1):35–67.

Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I., and Naor, M. (2006). Our data, ourselves: Privacy via distributed noise generation. In *Annual international conference on the theory and applications of cryptographic techniques*, pages 486–503. Springer.

Dwork, C., Roth, A., et al. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407.

Dwork, C., Rothblum, G. N., and Vadhan, S. (2010). Boosting and differential privacy. In *2010 IEEE 51st annual symposium on foundations of computer science*, pages 51–60. IEEE.

Einbeck, J. and Tutz, G. (2006). The fitting of multifunctions: an approach to nonparametric multimodal regression. *A. Rizzi, MV, editor, COMPSTAT 2006, Proceedings in Computational Statistics*, pages 1243–1250.

Fukunaga, K. and Hostetler, L. (1975). The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on information theory*, 21(1):32–40.

Genovese, C. R., Perone-Pacifico, M., Verdinelli, I., and Wasserman, L. (2014). Nonparametric ridge estimation. *Annals of Statistics*, 42(4):1511–1545.

Ghazi, B., Kumar, R., and Manurangsi, P. (2020). Differentially private clustering: Tight approximation ratios. *Advances in Neural Information Processing Systems*, 33:4040–4054.

Hall, R., Rinaldo, A., and Wasserman, L. (2013). Differential privacy for functions and functional data. *The Journal of Machine Learning Research*, 14(1):703–727.

Holohan, N., Braghin, S., Mac Aonghusa, P., and Levacher, K. (2019). Diffprivlib: the ibm differential privacy library. *arXiv preprint arXiv:1907.02444*.

Li, J., Ray, S., and Lindsay, B. G. (2007). A nonparametric statistical approach to clustering via mode identification. *Journal of Machine Learning Research*, 8(Aug):1687–1723.

Liu, E., Hu, J. Y.-C., Reneau, A., Song, Z., and Liu, H. (2024). Differentially private kernel density estimation. *arXiv preprint arXiv:2409.01688*.

Oberski, D. L. and Kreuter, F. (2020). Differential privacy and social science: An urgent puzzle. *Harvard Data Science Review*, 2(1):1–21.

Sheffet, O. (2017). Differentially private ordinary least squares. In *International Conference on Machine Learning*, pages 3105–3114. PMLR.

Stemmer, U. (2021). Locally private k-means clustering. *Journal of Machine Learning Research*, 22(176):1–30.

Su, D., Cao, J., Li, N., Bertino, E., and Jin, H. (2016). Differentially private k-means clustering. In *Proceedings of the sixth ACM conference on data and application security and privacy*, pages 26–37.

Tsybakov, A. B. (2008). Nonparametric estimators. In *Introduction to Nonparametric Estimation*, pages 1–76. Springer.

Wagner, T., Naamad, Y., and Mishra, N. (2023). Fast private kernel density estimation via locality sensitive quantization. In *International Conference on Machine Learning*, pages 35339–35367. PMLR.

Wang, Y.-X. (2018). Revisiting differentially private linear regression: optimal and adaptive prediction & estimation in unbounded domain. *arXiv preprint arXiv:1803.02596*.

Wasserman, L. and Zhou, S. (2010). A statistical framework for differential privacy. *Journal of the American Statistical Association*, 105(489):375–389.

Yamasaki, R. and Tanaka, T. (2024). Convergence analysis of mean shift. *IEEE transactions on pattern analysis and machine intelligence*, 46(10):6688–6698.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes/No/Not Applicable]

   All mathematical details are given in Section 4. All experiment details are given in Section 5 and the supplement.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

   All assumptions are present in Section 4. All proofs are given in the supplement.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

   All experiment details are given in Section 5 and the supplement.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

   Existing libraries and codes used here are given in our experimental details in Section 5.

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]