# Dimensionality Reduction for General KDE Mode Finding

**Xinyu Luo** [*1] **Christopher Musco** [*2] **Cas Widdershoven** [*3]

## Abstract

Finding the mode of a high dimensional probability distribution $\mathcal{D}$ is a fundamental algorithmic problem in statistics and data analysis. There has been particular interest in efficient methods for solving the problem when $\mathcal{D}$ is represented as a *mixture model* or *kernel density estimate*, although few algorithmic results with worst-case approximation and runtime guarantees are known. In this work, we significantly generalize a result of (Lee et al., 2021) on mode approximation for Gaussian mixture models. We develop randomized dimensionality reduction methods for mixtures involving a broader class of kernels, including the popular logistic, sigmoid, and generalized Gaussian kernels. As in Lee et al.'s work, our dimensionality reduction results yield quasi-polynomial algorithms for mode finding with multiplicative accuracy $(1-\epsilon)$ for any $\epsilon > 0$. Moreover, when combined with gradient descent, they yield efficient practical heuristics for the problem. In addition to our positive results, we prove a hardness result for box kernels, showing that there is no polynomial time algorithm for finding the mode of a kernel density estimate, unless $P = NP$. Obtaining similar hardness results for kernels used in practice (like Gaussian or logistic kernels) is an interesting future direction.

## 1. Introduction

We consider the basic computational problem of finding the mode of a high dimensional probability distribution $\mathcal{D}$ over $\mathbb{R}^d$. Specifically, if $\mathcal{D}$ has probability density function (PDF) $p$, our goal is to find any $x^* \in \mathbb{R}^d$ such that:

$$x^* \in \operatorname{argmax}_{x \in \mathbb{R}^d} p(x)$$

A natural setting for this problem is when $\mathcal{D}$ is specified as a *kernel density estimate* (KDE) or *mixture distribution* (Scott, 2015; Silverman, 2018). In this setting, we are given a set of points $M \in \mathbb{R}^d$ and a non-negative kernel function $\kappa : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^+$ and our PDF equals:

$$p(x) = \mathcal{K}_M(x) = \frac{1}{|M|} \sum_{m \in M} \kappa(x, m).$$

As is typically the case, we will assume that $\kappa$ is *shift-invariant* and thus only depends on the difference between $x$ and $m$, meaning that it can be reparameterized as $\kappa(x-m)$. A classic example of a shift-invariant KDE is any mixture of Gaussians distribution, for which $\kappa(x-m) = C \cdot e^{-\|x-m\|_2^2}$ is taken to be the Gaussian kernel. Here $C = \pi^{-d/2}$ is a normalizing constant. Kernel density estimates are widely used to approximate other distributions in a compact way (Botev et al., 2010; Kim & Scott, 2012), and they have been applied to applications ranging from image annotation (Yavlinsky et al., 2005), to medical data analysis (Sheikhpour et al., 2016), to outlier detection (Kamalov & Leung, 2020). The specific problem of finding the mode of a KDE has found applications in object tracking (Shen et al., 2007), super levelset computation (Phillips et al., 2015), typical object finding (Gasser et al., 1997), and more (Lee et al., 2021).

### 1.1. Prior Work

Despite its many applications, the KDE mode finding problem presents a computational challenge in high-dimensions. For any practically relevant kernel $\kappa$ (e.g., Gaussian) there are no known algorithms with runtime polynomial in both $n$ and $d$ for KDEs on $n = |M|$ base points. This is even the case when we only want to find an *$\epsilon$-approximate* mode for some $\epsilon \in (0, 1)$, i.e. a point $\tilde{x}^*$ satisfying

$$\mathcal{K}_M(\tilde{x}^*) \geq (1-\epsilon) \max_{x \in \mathbb{R}^d} \mathcal{K}_M(x),$$

There has been extensive work on heuristic local search methods like the well-known "mean-shift" algorithm (Carreira-Perpiñán, 2000; 2007), which can be viewed as

[*]Equal contribution [1]Department of Computer Science, Purdue University, Indiana, USA [2]Tandon School of Engineering, New York University, New York, USA [3]State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Sciences, Beijing, China. Correspondence to: Xinyu Luo <luo466@purdue.edu>, Christopher Musco <cmusco@nyu.edu>, Cas Widdershoven <cas.widdershoven@ios.ac.cn>.

*Proceedings of the $40^{th}$ International Conference on Machine Learning*, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

a variant of gradient descent, and often works well in practice. However, these methods do not come with theoretical guarantees and can fail on natural problem instances.

While polynomial time methods are not known, for some kernels it is possible to provably solve the $\epsilon$-approximate mode finding problem in *quasi-polynomial* time. For example, Shenmaier's work on *universal approximate centers* for clustering can be used to reduce the problem to evaluating the quality of a quasi-polynomial number of candidate modes (Shenmaier, 2019). For the Gaussian kernel, the total runtime is $d \cdot 2^{O(\log^2 n)}$ for constant $\epsilon$. Similar runtimes can be obtained by appealing to results on the approximate Carathéodory problem (Blum et al., 2019; Barman, 2015).

More recently, Lee et al. explore *dimensionality reduction* as an approach to obtaining quasi-polynomial time algorithms for KDE mode finding (Lee et al., 2021). Their work shows that, for the Gaussian kernel, any high-dimensional KDE instance can be reduced to a *lower dimensional instance* using randomized dimensionality reduction methods – specifically Johnson-Lindenstrauss projection. An approximate mode for the lower dimensional problem can then be found with a method that depends exponentially on the dimension $d$, and finally, the low-dimensional solution can be "mapped back" to the high-dimensional space[1]. Ultimately, the result in (Lee et al., 2021) allows all dependencies on $d$ to be replaced with terms that are polynomial in $\log(n)$ and $\epsilon$. The conclusion is that the mode of a Gaussian KDE can be approximated to accuracy $\epsilon$ in time $O\left(ndw + 2^w\right)$, where $w = \mathrm{poly}(\log n, 1/\epsilon)$. The leading $ndw$ term is the cost of performing the dimensionality reduction.

In addition to nearly matching prior quasi-polynomial time methods in theory (e.g., Shenmaier's approach), there are a number of benefits to an approach based on dimensionality reduction. For one, sketching directly reduces the space complexity of the mode finding problem, and vectors sketched with JL random projections can be useful in other downstream data analysis tasks. Another benefit is that dimensionality reduction can speed up even heuristic algorithms: instead of using a brute-force approach to solve the low-dimensional KDE instance, a practical alternative is to apply a local search method, like mean-shift, in the low-dimensional space. This approach sacrifices theoretical guarantees, but can lead to faster practical algorithms.

### 1.2. Our Results

The main contribution of our work is to generalize the dimensionality reduction results of (Lee et al., 2021) to a much broader class of kernels, beyond the Gaussian kernel studied in that work. In particular, we introduce a carefully defined class of kernels called "relative-distance smooth kernels". This class includes the Gaussian kernel, as well as the sigmoid, logistic, and any generalized Gaussian kernel of the form $\kappa(x, y) = e^{-\|x-y\|_2^\alpha}$ for $\alpha > 0$. See Definition 3 for more details. Our first result (Lemma 3.4) is that, for any relative-distance smooth kernel, we can approximate the *value* of the mode $\max_x \mathcal{K}_M(x)$ up to multiplicative error $(1-\epsilon)$ by solving a lower dimensional instance obtained by sketching the points in $M$ using a Johnson-Lindenstrauss random projection. The required dimension of the projection is $O(\log^c(n)/\epsilon^2)$, where $c$ is a constant depending on parameters of the kernel $\kappa$. For most commonly used relative-distance smooth kernels, including the Gaussian, logistic, and sigmoid kernels, $c = 3$. This leads to a dimensionality reduction that is completely independent of the original problem dimension $d$ and only depends polylogarithmically on the number of points in the KDE, $n$.

Moreover, in Section 4, we show how to recover an approximate mode $\tilde{x}$ satisfying $K_M(\tilde{x}) \geq (1-\epsilon)\max_x \mathcal{K}_M(x)$ from the solution of the low-dimensional sketched problem. When the kernel satisfies an additional convexity property, recovery can be performed in $O(nd)$ time using a generalization of the mean-shift algorithm used in (Lee et al., 2021). When the kernel does not satisfy the property, we obtain a slightly slower method using a recent result of (Biess et al., 2019) on constructive Lipschitz extensions. One consequence of our general results is the following claim for a number of common kernels:

**Theorem 1.1.** *Let $\mathcal{K}_M = (\kappa, M)$ be a be a KDE on $n = |M|$ points in $d$ dimensions, where $\kappa$ is a Gaussian, logistic, sigmoid, Cauchy[2], or generalized Gaussian kernel with parameter $\alpha \leq 1$. Let $\Pi$ be a random JL matrix with $w = O\left(\frac{\log^2(n/\epsilon)\log(n/\delta)}{\epsilon^2}\right)$ rows and let $\tilde{x}$ be any point such that $\mathcal{K}_{\Pi M}(\tilde{x}) \geq (1-\beta)\max_{x\in\mathbb{R}^w} \mathcal{K}_{\Pi M}(x)$. Given $\tilde{x}$ as input, Algorithm 2 runs in $O(nd)$ time and returns, with probability $1-\delta$, a point $x' \in \mathbb{R}^d$ satisfying:*

$$\mathcal{K}_M(x') \geq (1-\epsilon-\beta)\max_{x\in\mathbb{R}^d} \mathcal{K}_M(x).$$

Above, $\Pi M$ is the point set $\Pi M = \{\Pi m \text{ for } m \in M\}$ and $\mathcal{K}_{\Pi M}$ is the low-dimensional KDE defined by $\Pi M$ and $\kappa$. Theorem 1.1 implies that an approximate high-dimensional mode can be found by (approximately) solving a much lower dimensional problem. The result exactly matches that of (Lee et al., 2021) in the Gaussian case.

When combined with a simple brute-force method for maximizing $\mathcal{K}_{\Pi M}$, Theorem 1.1 immediately yields a quasi-polynomial time algorithm for mode finding. Again we state a natural special case of this result, proven in Section 5.

[1] Methods that run in time exponential in $d$ are straightforward to obtain via discretization/brute force search. See Section 5.

[2] For the Cauchy kernel, we can actually obtain a better bound with dimension $w = O\left(\frac{\log(n/\epsilon)}{\epsilon^2}\right)$. See Corollary 3.2.

**Theorem 1.2.** *Let $\mathcal{K}_M = (\kappa, M)$ be a be a KDE on $n$ points in $d$ dimensions, where $\kappa$ is a Gaussian, logistic, sigmoid, Cauchy, or generalized Gaussian kernel with $\alpha \leq 1$. There is an algorithm which finds a point $\tilde{x}$ satisfying:*

$$\mathcal{K}_M(\tilde{x}) \geq (1-\epsilon) \max_x \mathcal{K}_M(x)$$

*in $2^{\tilde{O}(\log^3 n/\epsilon^2)} + O(nd\log^3(n)/\epsilon^2)$ time. Here $\tilde{O}(x)$ denotes $O(x \log^c x)$ for constant $c$.*

Interestingly, as in (Lee et al., 2021), the above result falls just short of providing a polynomial time algorithm: doing so would require improving the $\log^3 n$ dependence in the exponent to $\log n$. It is possible to achieve polynomial time by make additional assumptions. For example, if we assume that $\mathcal{K}_M(x^*) \geq \rho$ for some constant $\rho$, then dependencies on $\log(n)$ can be replaced with $\log(1/\rho)$ using existing *coreset methods* (Lee et al., 2021; Phillips & Tai, 2018). However, the question still remains as to whether the general KDE mode finding problem can be solved in polynomial time for any natural kernel. Our final contribution is to take a step towards answering this question in the negative by relating the mode finding problem to the $k$-clique problem, and showing an NP-hardness result for box kernels (defined in the next section). Formally, in Section 6, we prove:

**Lemma 1.3.** *The problem of computing a $\frac{1}{n}$-approximate mode of a box kernel KDE is NP-hard.*

Unfortunately, our lower bound does not extend to commonly used kernels like the Gaussian, logistic, or sigmoid kernels. Proving lower bounds (or finding polynomial time algorithms) for these kernels is a compelling future goal.

**Paper Structure.** Section 2 contains notation and definitions. In Section 3 we provide our main dimensionality results for approximating the objective value for the mode. Then, in Section 4, we show how to recover a high-dimensional mode from a low-dimensional one, providing different approaches for when the kernel is convex and not. Section 5 outlines a brute force method for finding an approximate mode in low dimensions. In Section 6 we show that the approximate mode finding problem is NP-hard for box kernels. Finally, we provide experimental results in Section 7, confirming that dimensionality reduction combined with a heuristic mode finding method yields a practical algorithm for a variety of kernels and data sets.

## 2. Preliminaries

**Notation.** For our purposes, a kernel density estimate (KDE) is defined by a set of $n$ points (a.k.a. centers) $M \subset \mathbb{R}^d$ and a non-negative, shift-invariant kernel function. All of the kernels discussed in this work are also *radial symmetric*. This means that we can actually rewrite the kernel function $\kappa$ to be a scalar function of the squared Euclidean distance $\|x-m\|_2^2$.[3]. Our KDE then has the form:

$$\mathcal{K}_M(x) = \frac{1}{n} \sum_{m \in M} C \cdot \kappa(\|x-m\|_2^2).$$

We further assume that $\kappa : \mathbb{R} \to \mathbb{R}$ is non-increasing, so satisfies $\kappa(t) \geq \kappa(t') \geq 0$ for all $t' \geq t$. In the expression above, $C$ is a normalizing constant that only depends on $\kappa$. It is chosen to ensure that $\int_{t \in \mathbb{R}^d} C \cdot \kappa(t)\, dt = 1$ and thus $\mathcal{K}_M$ is a probability density function. The above function $\mathcal{K}_M(x)$ is invariant to scaling $\kappa$, so to ease notation we further assume that $\kappa(0) = 1$. Note that since $\kappa$ is non-increasing, we thus always have that $\max_t \kappa(t) = \kappa(0) = 1$. We write $\kappa'$ to denote the first-order derivative of $\kappa$ (whenever it exists).

Many common kernels are radial symmetric and non-increasing, so fit the form described above (Silverman, 2018; Altman, 1992; Cleveland & Devlin, 1988). We list a few:

Gaussian: $\kappa(t) = e^{-t}$

Logistic: $\kappa(t) = \dfrac{4}{e^{\sqrt{t}} + 2 + e^{-\sqrt{t}}}$

Sigmoid: $\kappa(t) = \dfrac{2}{e^{\sqrt{t}} + e^{-\sqrt{t}}}$

Cauchy: $\kappa(t) = \dfrac{1}{1+t}$

Generalized Gaussian: $\kappa(t) = e^{-t^\alpha}$

Box: $\kappa(t) = 1$ for $|t| \leq 1$, $\kappa(t) = 0$ otherwise.

Epanechnikov: $\kappa(t) = \max(0, 1-t)$

We are interested in finding a value for $x$ which maximizes or approximately maximizes the kernel density estimate $\mathcal{K}_M(x)$. Again since the problem is invariant to positive scaling, we will consider the problem of maximizing the unnormalized KDE, which we denote by $\bar{\mathcal{K}}_M(x)$:

$$\bar{\mathcal{K}}_M(x) = \sum_{m \in M} \kappa(\|x-m\|_2^2) = \frac{n}{C} \cdot \mathcal{K}_M(x)$$

Our general dimensionality reduction result depends on a parameter of $\kappa$ that we call the "critical radius". For common kernels we later show how to bound this parameter to obtain specific dimensionality reduction results.

**Definition 2.1** ($\alpha$-critical radius, $\xi_\kappa(\alpha)$)**.** For any non-increasing kernel function $\kappa : \mathbb{R} \to \mathbb{R}$, the $\alpha$-critical radius $\xi_\kappa(\alpha)$ is the smallest value of $t$ such that $\kappa(t) \leq \alpha$.

Note that for any $t \geq \xi_\kappa(\alpha)$, we have that $\kappa(t) \leq \alpha$. The value of $\xi_\kappa(\epsilon/2n)$ and $\xi_\kappa(1/n)$ will be especially important in our proofs. Specifically, since $\kappa$ is assumed to have

[3] We let $\|\cdot\|_2^2$ denotes the squared Euclidean norm: $\|a\|_2^2 = \sum_{i=1}^d a_i^2$, where $a_i$ is the $i^{\text{th}}$ entry in the length $d$ vector $a$.

$\kappa(0) = 1$, it is easy to check that any mode for $\mathcal{K}$ must lie within squared distance $\xi_\kappa(1/n)$ from at least one point in $M$, a region which we will call the *critical area*. We will use this fact.

**Johnson-Lindenstrauss Lemma.** Our results leverage the Johnson-Lindenstrauss (JL) lemma, which shows that a set of high dimensional points can be mapped into a space of much lower dimension in such a way that distances between the points are nearly preserved. We use the standard variant of the lemma where the mapping is an easy to compute random linear transformation (Achlioptas, 2001; Dasgupta & Gupta, 2003). Specifically, we are interested in random transformations satisfying the following guarantee:

**Definition 2.2** ($(\gamma, n, \delta)$-Johnson-Lindenstrauss Guarantee)**.** A randomly selected matrix $\Pi \in \mathbb{R}^{w\times d}$ satisfies the $(\gamma, n, \delta)$-JL guarantee for positive error parameter $\gamma$, if for any $n$ data points $v_1, ..., v_n \in \mathbb{R}^d$, with probability $1 - \delta$,

$$\|v_i - v_j\|_2^2 \leq \|\Pi v_i - \Pi v_j\|_2^2 \leq (1+\gamma)\|v_i - v_j\|_2^2 \quad (1)$$

for all pairs $i, j \in \{1, ..., n\}$ simultaneously.

Note that we require one-sided error: most statements of the JL guarantee have a $(1-\gamma)$ factor on the left side of the inequality. This is easily removed by scaling $\Pi$ by $\frac{1}{1-\gamma}$. It is well known that Definition 2.2 is satisfied by a properly i.i.d. random Gaussian or random $\pm 1$ matrix with

$$w = O\left(\frac{\log(n/\delta)}{\min(1, \gamma^2)}\right)$$

rows, and this is tight (Larsen & Nelson, 2017). General sub-Gaussian random matrices also work, as well as constructions that admit faster computation of $\Pi v_i$ (Kane & Nelson, 2014; Ailon & Chazelle, 2009).

**Kirszbraun Extension Theorem.** We also rely on a classic result of (Kirszbraun, 1934). Let $H_1$ and $H_2$ be Hilbert spaces. Kirszbraun's theorem states that if $S$ is a subset of $H_1$, and $f : S \to H_2$ is a Lipschitz-continuous map, then there is a Lipschitz-continuous map $g : H_1 \to H_2$ that extends[4] $f$ and has the same Lipschitz constant. Formally, when applied to Euclidean spaces $\mathbb{R}^w$ and $\mathbb{R}^d$ we have:

*Fact* 2.3. (Kirszbraun Extension Theorem). For any $\mathcal{S} \subset \mathbb{R}^w$, let $f : S \to \mathbb{R}^d$ be an L-Lipschitz function. That is $\forall x, y \in \mathcal{S}$, $\|f(x) - f(y)\|_2 \leq L\|x - y\|_2$. Then, there always exists some function $g : \mathbb{R}^w \to \mathbb{R}^d$ such that:

1. $g(x) = f(x)$ for all $x \in \mathcal{S}$,
2. $g$ is also L-Lipschitz. That is for all $x, y \in \mathbb{R}^w$, $\|g(x) - g(y)\|_2 \leq L\|x - y\|_2$.

[4] I.e. $g(s) = f(s)$ for all $s \in S$.

## 3. Dimensionality Reduction for Approximating the Mode Value

In this section, we show that using a JL random projection, we can reduce the problem of approximating the *value* of the mode of a KDE in $d$ dimensions – i.e., $\max_x \bar{\mathcal{K}}_M(x)$ – to the problem of approximating the value of the mode for a KDE in $d'$ dimensions, where $d'$ depends only on $n$, $\kappa$, and the desired approximation quality. This problem of recovering the mode value is a prerequisite for the harder problem of recovering the *location* of an approximate mode (i.e., a point $x^* \in \mathbb{R}^d$ such that $\mathcal{K}(x^*) \geq (1-\epsilon)\max_{x\in\mathbb{R}^d}\mathcal{K}(x)$), which is addressed in Section 4.

We begin with an analysis for JL projections that bounds $d'$ based on generic properties of $\kappa$. Then, in Section 3.1 we analyze these properties for specific kernels of interest, and prove that $d'$ is in fact small for these kernels – specifically, it depends just polylogarithmically on $n$ and polynomially on the approximation factor $\epsilon$. Our general result follows:

**Theorem 3.1.** *Let $\mathcal{K}_M = (\kappa, M)$ be a $d$-dimensional KDE on a differentiable kernel as defined in Section 2 and let $0 < \epsilon \leq 1$ be an approximation factor. Let $\xi \geq \xi_\kappa(\frac{\epsilon}{2n})$ and let $\kappa'_{\min} \leq \min_{0\leq t\leq 2\xi}\frac{\kappa'(t)t}{\kappa(t)}$. Note that $\kappa'_{\min} \leq 0$ since $\kappa$ is assumed to be non-increasing. We can assume that $\kappa'_{\min} \neq 0$. Let $\gamma = -\frac{\epsilon}{2\kappa'_{\min}} > 0$. Then with probability $(1-\delta)$, for any $\Pi \in \mathbb{R}^{w\times d}$ satisfying the $(\gamma, n+1, \delta)$-JL guarantee, we have:*

$$(1-\epsilon)\max_{x\in\mathbb{R}^d}\mathcal{K}_M(x) \leq \max_{x\in\mathbb{R}^w}\mathcal{K}_{\Pi M}(x) \leq \max_{x\in\mathbb{R}^d}\mathcal{K}_M(x). \quad (2)$$

*Recall that a random $\Pi$ with $w = O\left(\frac{\log((n+1)/\delta)}{\min(1,\gamma^2)}\right)$ rows will satisfy the required $(\gamma, n+1, \delta)$-JL guarantee.*

Note that in the theorem statement above, $\Pi M = \{\Pi m : m \in M\}$ denotes the point set $M$ with dimension reduced by multiplying each point in the set by $\Pi$. Our proof of Theorem 3.1 is included in Appendix A. It leverages Kirszbraun's Exention theorem, and follows along the same lines in (Lee et al., 2021). However, we need to more carefully track the effect of properties of the kernel function $\kappa$, since we do not assume that it has the simple form of a Gaussian kernel.

With Theorem 3.1 in place, we can apply it to any non-increasing differentiable kernel to obtain a dimensionality reduction result: we just need to compute a lower bound $\kappa'_{\min} \leq \min_{0\leq t\leq 2\xi}\frac{\kappa'(t)t}{\kappa(t)}$. For some kernels we can do so directly. For example, consider the Cauchy kernel, $\kappa(t) = \frac{1}{1+t}$. It can be shown that we can pick $\kappa'_{\min} = -1$ (since $\kappa'(t)t/\kappa(t) \geq -1$ for all $t$). Plugging into Theorem 3.1 we obtain:

**Corollary 3.2.** *Let $\mathcal{K}_m = (\kappa, M)$ be a KDE and, for any $\delta, \epsilon \in (0,1)$, let $\Pi$ be a random JL matrix with $w = O\left(\frac{\log(n/\delta)}{\epsilon^2}\right)$ rows. If $\kappa$ is a Cauchy kernel, then*

*with probability* $1-\delta$,

$$(1-\epsilon)\max_{x\in\mathbb{R}^d}\mathcal{K}_M(x) \leq \max_{x\in\mathbb{R}^w}\mathcal{K}_{\Pi M}(x) \leq \max_{x\in\mathbb{R}^d}\mathcal{K}_M(x).$$

In the following subsection we will describe a broader class of kernels for which we can also obtain good dimensionality reduction results, but for which bounding $\kappa'_{\min}$ is a bit more challenging.

### 3.1. Relative-Distance Smooth Kernels

Specifically, we consider a broad class of kernels, that included the Gaussian kernel:

**Definition 3.3** (Relative-distance smooth kernel)**.** A non-increasing differentiable kernel $\kappa$ is *relative-distance smooth* if there exist constants $c_1, d_1, q_1, c_2, d_2 > 0$ such that

$$c_1 t^{d_1} - q_1 \leq \frac{-\kappa'(t)t}{\kappa(t)} \leq c_2 t^{d_2} \qquad \text{for all} \qquad t \geq 0.$$

In addition to the Gaussian kernel, this class includes other kernels commonly used in practice, like the logistic, sigmoid, and generalized Gaussian kernels:

$$\text{Gaussian: } t^1 \leq \frac{-\kappa'(t)t}{\kappa(t)} = t \leq t^1$$

$$\text{Logistic: } \frac{t^{1/2}}{2} - \frac{1}{2} \leq \frac{-\kappa'(t)t}{\kappa(t)} = \frac{(e^{\sqrt{t}}-1)\sqrt{t}}{2(e^{\sqrt{t}}+1)} \leq \frac{t^{1/2}}{2}$$

$$\text{Sigmoid: } \frac{t^{1/2}}{2} - \frac{1}{2} \leq \frac{-\kappa'(t)t}{\kappa(t)} = \frac{(e^{2\sqrt{t}}-1)\sqrt{t}}{2(e^{2\sqrt{t}}+1)} \leq \frac{t^{1/2}}{2}$$

$$\text{Generalized Gaussian: } \alpha t^\alpha \leq \frac{-\kappa'(t)t}{\kappa(t)} = \alpha t^\alpha \leq \alpha t^\alpha$$

A few common non-increasing kernels, including the rational quadratic kernel, are *not* relative distance smooth. Our main result is that for *any* relative-distance smooth kernel, we can sketch the KDE to dimension $w$ which depends only polylogarithically on $n = |M|$ and quadratically on $1/\epsilon$:

**Lemma 3.4.** *Let $\mathcal{K}_m$ be a KDE for a relative-distance smooth kernel $\kappa$ with parameters $c_1, d_1, q_1, c_2, d_2$. There is a fixed constant $c'$ such that if $\gamma = \frac{\epsilon}{c'}\log^{-d_2/d_1}\left(\frac{2n}{\epsilon}\right)$, then with probability $(1-\delta)$, for any $\Pi \in \mathbb{R}^{w\times d}$ satisfying the $(\gamma, n+1, \delta)$-JL guarantee, Equation (2) holds. To obtain this JL guarantee, it suffices to take $\Pi$ to be a random JL matrix with $w = O\left(\frac{\log^{2d_2/d_1}(n/\epsilon)\log(n/\delta)}{\epsilon^2}\right)$ rows.*

Lemma 3.4 is proven in Appendix A. It uses an intermediate result that bounds the $\frac{\epsilon}{2n}$-critical radius for any relative-distance smooth kernel, which is required to invoke Theorem 3.1. Interestingly, the polylogarithmic factor in Lemma 3.4 only depends on the ratio of the parameters $d_2$ and $d_1$ of the relative-distance smooth kernel $\kappa$. For all of the example kernels discussed above, this ratio equals 1, so we obtain a dimensionality reduction result exactly matching (Lee et al., 2021) for the Gaussian kernel:

**Corollary 3.5.** *Let $\mathcal{K}_m = (\kappa, M)$ be a KDE and, for any $\delta, \epsilon \in (0,1)$, let $\Pi$ be a random JL matrix with $w = O\left(\frac{\log^2(n/\epsilon)\log(n/\delta)}{\epsilon^2}\right)$ rows. If $\kappa$ is a Gaussian, logistic, sigmoid kernel, or generalized Gaussian kernel, then with probability $1-\delta$,*

$$(1-\epsilon)\max_{x\in\mathbb{R}^d}\mathcal{K}_M(x) \leq \max_{x\in\mathbb{R}^w}\mathcal{K}_{\Pi M}(x) \leq \max_{x\in\mathbb{R}^d}\mathcal{K}_M(x).$$

## 4. Recovering an Approximate Mode in High Dimensions

In Section 3, we discussed how to convert a high dimensional KDE into a lower dimensional KDE whose mode has an approximately equal value. However, in applications, we are typically interested in computing a point in the high-dimensional space whose value is approximately equal to the value of the mode. I.e., using our dimensionality reduced dataset, we want to find some $\tilde{x}$ such that:

$$\mathcal{K}_M(\tilde{x}) \geq (1-\epsilon)\max_x \mathcal{K}_M(x).$$

We present two approaches for doing so. The first is based on Kirszbraun's extension theorem and the widely used mean-shift heuristic. It extends the approach of (Lee et al., 2021) to a wider class of kernels – specifically to any *convex and non-increasing* kernel $\kappa$. This class contains most of the relative-distance smooth kernels discussed in Section 3.1, including the Gaussian, sigmoid, and logistic kernels, and generalized Gaussian kernels when $\alpha \leq 1$. It also includes common kernels like the Cauchy kernel, for which we have shown a strong dimensionality reduction results, and the Epanechnikov, biweight, and triweight kernels. Recall that we define $\kappa(t)$ so that $t$ represents the *squared* Euclidean distance between two points; we specifically need $\kappa$ as defined in this way to be convex.

For non-convex kernels, we briefly discuss a second approach in Appendix B based on recent work on explicit one point extensions of Lipschitz functions (Biess et al., 2019). While less computationally efficient, this approach works for any non-increasing $\kappa$. Common examples of non-convex kernels include the tricube kernel $\kappa(t) = (1-t^{3/2})^3$ (Altman, 1992), $\kappa(t) = 1-t^2$ (Comaniciu, 2000), or any generalized Gaussian kernel with $\alpha > 1$.

### 4.1. Mean-shift for Convex Kernels

Based on ideas proposed by Fukunaga and Hostetler (Fukunaga & Hostetler, 1975), the mean-shift method is a commonly used heuristic for finding an approximate mode (Cheng, 1995). The idea behind the algorithm is to iteratively refine a guess for the mode. At each update, a new

**Algorithm 1** Mean-Shift Algorithm

**Require:** Set of $n$ points $M \subset \mathbb{R}^d$, number of iterations $\tau$, differentiable kernel function $\kappa$.

1: Select initial point $x^{(0)} \in \mathbb{R}^d$

2: For $i = 0, ..., \tau - 1$:

$$x^{(i+1)} = \sum_{m \in M} m \cdot \frac{\kappa'\left(\left\|x^{(i)} - m\right\|_2^2\right)}{\sum_{j \in M} \kappa'\left(\left\|x^{(i)} - j\right\|_2^2\right)}$$

3: return $x^{(\tau)}$

guess $x^{(i+1)}$, is obtained by computing a weighted average of all points in $M$ that define the KDE. Points that are closer to the previous guess $x^{(i)}$ are included with higher weight than points that are further. The exact choice of weights depends on the first derivative $\kappa'(t)$, where $t$ is the distance from the current mode to a point in $M$. For any non-increasing, convex kernel, $\kappa'(t)$ is non-positive and decreasing in magnitude – i.e., $|\kappa'(t)|$ is largest for $t$ close to 0, which ensures that points closest to the current guess for the mode are weighted highest when computing the new guess[5]. We include pseudocode for mean-shift as Algorithm 1. The method can be alternatively viewed as an instantiation of gradient ascent for the KDE mode objective with a specifically chosen step size – we do not discuss details here.

A powerful property of the mean-shift algorithm is that it always converges for kernels that are non-increasing and convex. In fact, it is known to provide a monotonically improving solution. Specifically:

*Fact* 4.1 (Comaniciu & Meer (2002)). Let $x^{(0)} \in \mathbb{R}^d$ be an arbitrary starting point and let $x^{(1)}, \ldots, x^{(\tau)}$ be the resulting iterates of Algorithm 1 run on point set $M$ with kernel $\kappa$. If $\kappa$ is convex and non-increasing, then for any $i \in 1, \ldots, \tau$:

$$\mathcal{K}_M(x^{(i)}) \geq \mathcal{K}_M(x^{(i-1)}).$$

In Appendix A, we use this fact to prove that with a (modified) mean-shift method, run for only a single iteration, we can translate any approximate solution for a dimensionality reduced KDE problem to a solution for the original high dimensional problem. Formally, we prove the following:

**Theorem 4.2.** *Let $M$ be a set of points in $\mathbb{R}^d$ and let $\mathcal{K}_M = (\kappa, M)$ be a KDE defined by a shift-invariant, non-increasing, and convex kernel function $\kappa$. Let $x^* \in \operatorname{argmax}_x \mathcal{K}_M(x)$. Let $\Pi \in \mathbb{R}^{w \times d}$ be a JL matrix as in Definition 2.2 and assume that $w$ is chosen large enough so that for all $a, b$ in the set $\{x^*\} \cup M$,*

$$\|a - b\|_2^2 \leq \|\Pi a - \Pi b\|_2^2 \tag{3}$$

*and $(1-\epsilon) \max_{x \in \mathbb{R}^d} \mathcal{K}_M(x) \leq \max_{x \in \mathbb{R}^w} \mathcal{K}_{\Pi M}(x) \leq \max_{x \in \mathbb{R}^d} \mathcal{K}_M(x)$.*

*Let $\tilde{x} \in \mathbb{R}^w$ be an approximate maximizer for $\mathcal{K}_{\Pi M}$ satisfying $\mathcal{K}_{\Pi M}(\tilde{x}) \geq (1-\alpha) \max_{x \in \mathbb{R}^w} \mathcal{K}_{\Pi M}(x)$. Then if we choose $x' = \sum_{m \in M} m \cdot \frac{\kappa'(\|\tilde{x} - \Pi m\|^2)}{\sum_{m \in M} \kappa'(\|\tilde{x} - \Pi m\|^2)}$, we have:*

$$\mathcal{K}_M(x') \geq (1 - \epsilon - \alpha) \max_{x \in \mathbb{R}^d} \mathcal{K}_M(x).$$

Note that $x'$ above is set using a single-iteration of what looks like mean-shift. However, instead of using weights based on the distances of points in $M$ to a previous guess for a high-dimensional mode, we use distances between the points $\Pi M$ in our low-dimensional space to the approximate low-dimensional mode, $\tilde{x}$. Also note that Theorem 4.2 is independent of exactly how $\tilde{x}$ is computed – it could be computed using brute force search, using an approximation algorithm tailored to low-dimensional problems, as in (Lee et al., 2021), or using a heuristic like mean-shift itself.

**Algorithm 2** Mode Recovery for Convex Kernels

**Require:** Shift-invariant, non-increasing, and convex kernel function $\kappa$ with derivative $\kappa'$. Set of $n$ points $M \subset \mathbb{R}^d$, dimensionality reduction parameter $\gamma$, accuracy parameter $\alpha$, failure probability $\delta$.

1: Construct a random JL matrix $\Pi$ with $w = O\left(\frac{\log((n+1)/\delta)}{\min(1, \gamma^2)}\right)$ rows.

2: Construct a set of $n$ points $\Pi M \subset \mathbb{R}^w$ that contains $\Pi m$ for each $m \in M$.

3: Compute $\tilde{x}$ such that $\mathcal{K}_{\Pi M}(\tilde{x}) \geq (1 - \alpha) \max_{x \in \mathbb{R}^w} \mathcal{K}_{\Pi M}(x)$.

4: Return $x' = \sum_{m \in M} m \cdot \frac{\kappa'(\|\tilde{x} - \Pi m\|^2)}{\sum_{m \in M} \kappa'(\|\tilde{x} - \Pi m\|^2)}$

For convex kernels, Theorem 4.2 implies a strengthening of Theorem 3.1 that allows for recovering an approximate mode, not just the value of the mode. Formally, the combined dimensionality reduction and recover procedure we propose is included as Algorithm 2 and we have the following result on the its accuracy:

**Corollary 4.3.** *Let $\mathcal{K}_M = (\kappa, M)$ be a $d$-dimensional shift-invariant KDE as defined in Section 2 and let $\epsilon$ and $\gamma$ (which depends on $\kappa$ and $\epsilon$) be as in Theorem 3.1. If $\kappa$ is differentiable, non-increasing, and convex, then Algorithm 2 run with parameters $\gamma$ and $\alpha$ returns $x'$ satisfying:*

$$\mathcal{K}_M(x') \geq (1 - \epsilon - \alpha) \max_{x \in \mathbb{R}^d} \mathcal{K}_M(x).$$

Note that Line 4 in Algorithm 2 can be evaluated in $O(nw + nd)$ time. So our headline result, Theorem 1.1, follows as a direct corollary.

[5]Note that for the Gaussian kernel, $\kappa(t) = e^{-t}$, so $|\kappa'(t)| = \kappa(t)$. So the method presented here is equivalent to the version of mean-shift used in prior work on dimensionality reduction for mode finding (Lee et al., 2021)

## 5. Solving the Low-Dimensional Problem

We next discuss a simple brute-force search method for approximate mode finding for any KDE with a continuous kernel function $\kappa$. The method has an exponential runtime dependency on the dimension, so its use for high-dimensional problems is limited, but combined with the dimensionality reduction techniques from Section 3 and the mode recovery techniques from Section 4, it yields a quasi-polynomial mode finding algorithm for a large class of kernels.

Recall that the mode of a KDE $\mathcal{K} = (\kappa, M)$ with $|M| = n$ must lie within its critical area, i.e. in a ball of squared radius $\xi_\kappa(1/n)$ around one of the points in $M$ (where $\xi_\kappa(1/n)$ denotes the $1/n$-critical radius). For any $\delta > 0$ we define a finite $\delta$-*covering* $\mathcal{N}(\mathcal{K}, \delta)$ to be a finite set of points such that, for every point $p$ in the critical area of $\mathcal{K}$, there exists a $p' \in \mathcal{N}(\mathcal{K}, \delta)$ such that $\|p - p'\|_2^2 \leq \delta$. Formally:

**Lemma 5.1.** *Given a KDE $\mathcal{K} = (\kappa, M)$ in $\mathbb{R}^d$ with $|M| = n$, and parameter $\delta > 0$, let $\xi = \xi_\kappa(1/n)$ and let $\mathcal{N}(\mathcal{K}, \delta)$ be a set that contains all points of the form*

$$m + \sum_{i=1}^{d} \frac{k_i \sqrt{\delta}}{\sqrt{d}} e_i,$$

*where $m \in M$, $k_i \in \mathbb{Z}$, and $-\frac{\sqrt{d}\xi}{\sqrt{\delta}} \leq k_i \leq \frac{\sqrt{d}\xi}{\sqrt{\delta}}$. Above $e_i$ are the canonical base vectors of $\mathbb{R}^d$. Then for any point $p$ in a $\xi$-ball surrounding one of the points in $M$, there exists a point $p' \in \mathcal{N}(\mathcal{K}, \delta)$ such that $\|p - p'\|_2^2 \leq \delta$. Moreover, we have that $|\mathcal{N}(\mathcal{K}, \delta)| = n(2\sqrt{d}\xi/\sqrt{\delta})^d$.*

By checking every point in $\mathcal{N}(\mathcal{K}, \delta)$ and returning one that maximizes $\kappa$, we obtain the following results on finding an approximate mode, which is proven in Appendix A:

**Theorem 5.2.** *Given a KDE $\mathcal{K} = (\kappa, M)$ in $\mathbb{R}^d$ with $|M| = n$ and a precision parameter $\epsilon > 0$, let $\xi = \xi_\kappa(1/n)$ and let $\delta$ be at most the largest number such that $\kappa(c) - \kappa(c + \delta) \leq \epsilon\kappa(c)$ for all $c \leq \xi$. Then we can find an $\epsilon$-approximate mode in $O\left(n(2\sqrt{d}\xi/\sqrt{\delta})^d\right)$. In particular, if $d \leq O(\log^c(n))$, $\xi \leq O(n^c)$, and $\delta \geq O(n^{-c})$ for some constant $c$, then we can find an $\epsilon$-approximate mode in quasi-polynomial time in the number of data points $n$.*

Our headline result, Theorem 1.2, follows by combining the dimensional reduction guarantee of Lemma 3.4 with the observation that for $\bar{\xi} = \max(1, \xi_\kappa(1/n))$, choosing $\delta = \min\left(\left(\frac{d_2}{c_2}\epsilon\right)^{1/d_2}, \frac{\epsilon}{c_2}(2\bar{\xi})^{1-d_2}\right)$ satisfies the requirement of Theorem 5.2 for any relative-distance smooth kernel $\kappa$ with parameters $c_1, d_1, q_1, c_2, d_2$. Moreover, as established in Lemma A.1, $\xi_\kappa(\frac{1}{n}) \leq c \log^{1/d_1} n$, so we have that the runtime in Theorem 5.2 is $O(n(\log^c n)^d)$ for constant $\epsilon$. The claim in Theorem 1.2 for the Cauchy kernel follows by noting that $\xi = n$ and we can take $\delta = \frac{1}{\epsilon}$ in Theorem 5.2.

Finally, note that Theorem 1.2 also includes the polynomial time cost of multiplying the original data set by a random JL matrix.

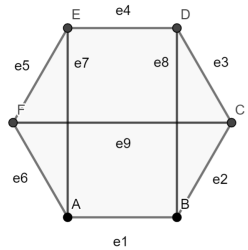
Overall, we conclude that one can compute an approximate mode in quasi-polynomial time for the Cauchy kernel, or any KDE on a relative-distance smooth kernel, and in particular the approximate mode finding problem for KDEs on Gaussian, logistic, sigmoid, or generalized Gaussian kernels can be solved in quasi-polynomial time.

## 6. Hardness Results

The results from the previous sections place the approximate mode finding problem in quasi-polynomial time for a large class of kernels. The question arises whether we can do much better; in this section, we provide some preliminary negative evidence for this possibility. Specifically, we prove NP-hardness of finding an approximate mode of a box kernel KDE, where we recall that this kernel takes the form $\kappa(t) = 1$ for $|t| \leq 1$ and $\kappa(t) = 0$ otherwise. Our hardness result is based on the hardness of the $k$-clique problem:

*Problem* 6.1 ($k$-clique). Given a $\Delta$-regular graph $G$ and an integer $k$, does $G$ have a complete $k$-vertex subgraph?

Problem 6.1 is known to be NP-hard when $k$ is a parameter of the input. We show how to reduce this problem to KDE mode finding using a reduction inspired by work of Shenmaier on the $k$-enclosing ball problem (Shenmaier, 2015). We start by creating a point set given an input $G$ to Problem 6.1. Specifically, we embed $G$ in $\mathbb{R}^{|E|}$ as follows: let $P$ to be the set of rows of the incidence matrix of $G$, i.e. the matrix $B$ such that $B_{v,e} = 1$ if $e$ is an edge adjacent to the node $v$ and $B_{v,e} = 0$ otherwise (Shenmaier, 2015). See Figure 1 for an example.




$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 \end{pmatrix}$$

*Figure 1.* An simple 3-regular graph and its incident matrix $B$.

We will base our hardness result on the following lemma:

**Lemma 6.2** (Shenmaier (2015))**.** *Given a $\Delta$-regular graph $G = (V, E)$ and integer $k$, let $P$ be defined as above. Let $A = (1 - 1/k)(\Delta - 1)$ and let $R$ be the radius of the smallest ball containing $\geq k$ points in $P$. Then $R^2 \leq A$ if there is a $k$-clique in $G$, and $R^2 \geq A + 2/k^2$ otherwise.*

By rescaling $P$ we can show NP-hardness of the KDE mode finding problem for box kernels:

**Theorem 6.3.** *The problem of computing a $\frac{1}{n}$-approximate mode of a box kernel KDE is NP-hard.*

*Proof.* The proof follows almost directly from Lemma 6.2. Note that the value of the mode of a box kernel KDE is given by the largest number of centers in a ball of radius 1. Let $G$ be an instance of Problem 6.1, and let $P$ be the set of rows of the incidence matrix of $G$ as described above. Now let $M = \{p/\sqrt{A} \mid p \in P\}$. From the lemma, we know that there is a ball of radius 1 containing $k$ points if $G$ has a $k$-clique, so $\max_x \bar{\mathcal{K}}_M(x) \geq k$. On the other hand, if $G$ does not have a $k$-clique then every ball of radius 1 contains at most $k-1$ points, i.e., $\max_x \bar{\mathcal{K}}_M \leq k-1$. So, an approximation algorithm with error $\epsilon = 1/k \geq 1/n$ can distinguish between the two cases. Hence, the ($\epsilon$-approximate) mode finding problem for box kernel KDEs is at least as hard as Problem 6.1 when $\epsilon \leq 1/n$. □

While it provides an initial result and hints at why the mode finding problem might be challenging, the above hardness result leaves a number of open questions. First off, it does not rule out a constant factor approximation, or a method whose dependence on the approximation parameter $\epsilon$ is exponential (as in our quasi-polynomial time methods). Moreover, the result does not apply for kernels like the Gaussian kernel – it strongly requires that the value of the box kernel differs significantly between $t = 1$ and $t = 1 + \frac{1}{k^2}$. Proving stronger hardness of approximation for the box kernel, or any hardness for kernels used in practice (like a relative-distance smooth kernel) are promising future directions.

## 7. Experiments

We support our theoretical results with experiments on two datasets, MNIST (60000 data points, 784 dimensions) and Text8 (71290 data points, 300 dimensions). We use both the Gaussian and Generalized Gaussian kernels with a variety of different bandwidths, $\sigma$. A bandwidth of $\sigma$ means that the kernel function as definied in Section 2 was applied to values $t = \frac{\|m-x\|_2^2}{\sigma^2}$. In general, a large $\sigma$ leads to larger mode value. It also leads to a smoother KDE, which is intuitively easier to maximize. We chose values of $\sigma$ that lead to substantially varying mode values to check the performance of our method across a variety of optimization surfaces.

Since these are high-dimensional datasets, it is not computationally feasible to find an exact mode to compare against. Instead, we obtain a baseline mode value by running the mean-shift heuristic (gradient descent) for 100 iterations, with 60 randomly chosen starting points. To avoid convergence to local optima at KDE centers, these starting points were chosen to be random linear combinations of either all dataset points, or a random pair of points in the data set. The best mode value found was taken as a baseline.

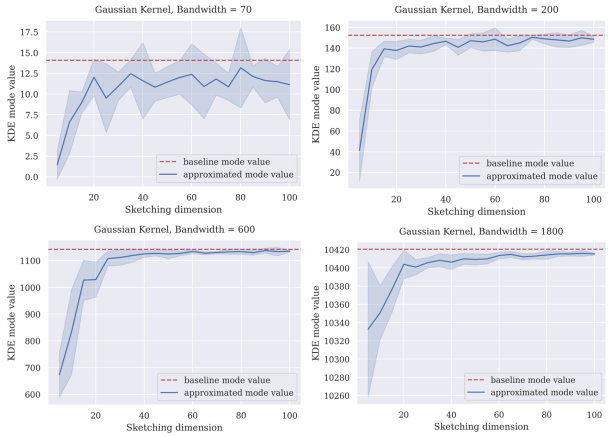



*Figure 2.* MNIST data using a Gaussian kernel with bandwidths 70, 200, 600, and 1800.

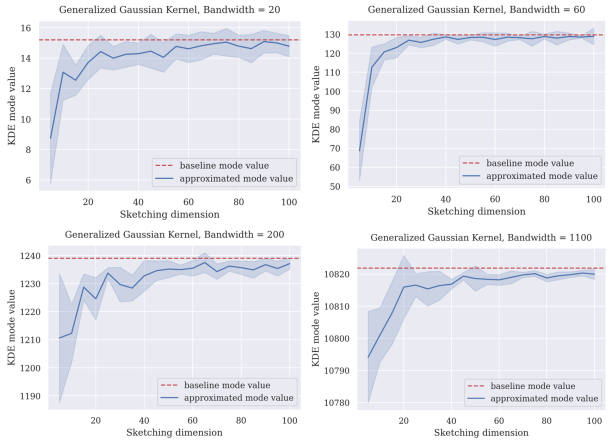



*Figure 3.* MNIST data using a Generalized Gaussian kernel with parameter $\alpha = .5$ and bandwidths 20, 60, 200, and 1100.

Once establishing a baseline, we applied JL dimensionality reduction to each data set and kernel for a variety of sketching dimensions, $w$. Again, for efficiency, we use mean-shift to find an approximate low-dimensional mode, instead of the brute force search method from Section 5. We ran for 10 iterations with 30 random restarts, chosen as described above. To recover a high-dimensional mode from our approximate low-dimensional mode, we use Algorithm 2, since the kernels tested are convex. For each dimension $w$, we ran 10 trials and report the mean and standard deviation of the KDE value of our approximate mode. Results are included in Figures 2-5. Note that, for visual clarity, the y-axis in these figures does not start at zero.

As apparent from the plots, our Johnson-Lindenstrauss dimensionality reduction approach combined with the mean-shift heuristic performs very well overall. In all cases, it was able to recover an approximate mode with value close to the baseline with sketching dimension $w \ll d$. As expected,

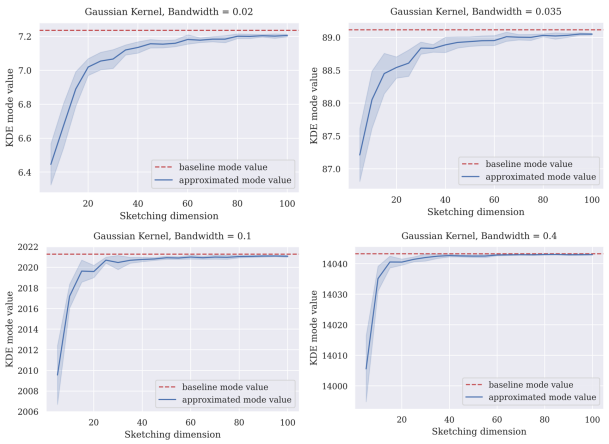



*Figure 4.* Text8 data using a Gaussian kernel with parameter with bandwidths .02, .035, .1, and .4.

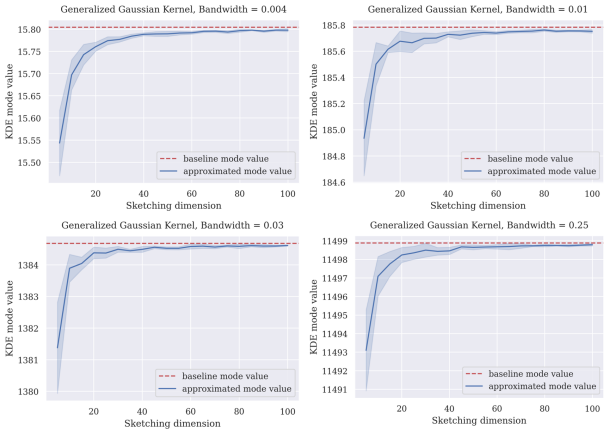



*Figure 5.* Text8 data using a Generalized Gaussian kernel with parameter $\alpha = .5$ and bandwidths .004, .01, .03, and .25.

performance improves with increasing sketching dimension.

## 8. Acknowledgements

This work was partially funded through NSF Award No. 2045590. Cas Widdershoven's work has been partially funded through the CAS Project for Young Scientists in Basic Research, Grant No. YSBR-040.

## References

Achlioptas, D. Database-friendly random projections. In *Proceedings of the Twentieth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pp. 274–281, 2001.

Ailon, N. and Chazelle, B. The fast johnson-lindenstrauss transform and approximate nearest neighbors. *SIAM J. Comput.*, pp. 302–322, 2009.

Altman, N. S. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185, 1992.

Barman, S. Approximating nash equilibria and dense bipartite subgraphs via an approximate version of caratheodory's theorem. In *Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing*, pp. 361–369, 2015.

Biess, A., Kontorovich, A., Makarychev, Y., and Zaichyk, H. Regression via kirszbraun extension with applications to imitation learning. *ArXiv Preprint*, abs/1905.11930, 2019. URL http://arxiv.org/abs/1905.11930.

Blum, A., Har-Peled, S., and Raichel, B. Sparse approximation via generating point sets. *ACM Trans. Algorithms*, 15(3), 6 2019.

Botev, Z. I., Grotowski, J. F., and Kroese, D. P. Kernel density estimation via diffusion. *The annals of Statistics*, 38(5):2916–2957, 2010.

Carreira-Perpiñán, M. A. Mode-finding for mixtures of Gaussian distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11):1318–1323, 2000.

Carreira-Perpiñán, M. A. Gaussian mean-shift is an EM algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29:767–776, 2007.

Cheng, Y. Mean shift, mode seeking, and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(8):790–799, 1995.

Cleveland, W. S. and Devlin, S. J. Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83(403): 596–610, 1988.

Comaniciu, D. and Meer, P. Mean shift: a robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619, 2002.

Comaniciu, D. I. *Nonparametric robust methods for computer vision*. PhD thesis, 2000.

Dasgupta, S. and Gupta, A. An elementary proof of a theorem of johnson and lindenstrauss. *Random Struct. Algorithms*, 22(1):60–65, 1 2003.

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

Gasser, T., Hall, P., and Presnell, B. Nonparametric estimation of the mode of a distribution of random curves. *Journal of the Royal Statistical Society: Series B*, 60: 681–691, 1997.

Kamalov, F. and Leung, H. H. Outlier detection in high dimensional data. *Journal of Information & Knowledge Management*, 19(01), 2020.

Kane, D. M. and Nelson, J. Sparser johnson-lindenstrauss transforms. *Journal of the ACM (JACM)*, 61(1):1–23, 2014.

Kim, J. and Scott, C. D. Robust kernel density estimation. *The Journal of Machine Learning Research*, 13(1):2529–2565, 2012.

Kirszbraun, M. Über die zusammenziehende und lipschitzsche transformationen. *Fundamenta Mathematicae*, 22(1):77–108, 1934.

Larsen, K. G. and Nelson, J. Optimality of the johnson-lindenstrauss lemma. In *58th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 633–638, 2017.

Lee, J. C., Li, J., Musco, C., Phillips, J. M., and Tai, W. M. Finding an approximate mode of a kernel density estimate. In *29th Annual European Symposium on Algorithms (ESA 2021)*, volume 204, pp. 61:1–61:19, 2021.

Phillips, J. M. and Tai, W. M. Near-optimal coresets of kernel density estimates. In *34th International Symposium on Computational Geometry*, 2018.

Phillips, J. M., Wang, B., and Zheng, Y. Geometric inference on kernel density estimates. In *International Symposium on Computational Geometry*, 2015.

Scott, D. W. *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons, 2015.

Sheikhpour, R., Sarram, M. A., and Sheikhpour, R. Particle swarm optimization for bandwidth determination and feature selection of kernel density estimation based classifiers in diagnosis of breast cancer. *Applied Soft Computing*, 40:113–131, 2016.

Shen, C., Brooks, M. J., and van den Hengel, A. Fast global kernel density mode seeking: Applications to localization and tracking. *IEEE Transactions on Image Processing*, 16:1457 – 1469, 2007.

Shenmaier, V. Complexity and approximation of the smallest k-enclosing ball problem. *European Journal of Combinatorics*, 48:81–87, 2015.

Shenmaier, V. A structural theorem for center-based clustering in high-dimensional euclidean space. In *Machine Learning, Optimization, and Data Science*, pp. 284–295. Springer International Publishing, 2019.

Silverman, B. W. *Density estimation for statistics and data analysis*. Routledge, 2018.

Yavlinsky, A., Schofield, E., and Rüger, S. Automated image annotation using global features and robust nonparametric density estimation. In *International Conference on Image and Video Retrieval*, pp. 507–517, 2005.

# A. Additional Proofs

## A.1. Proofs for Section 3

**Theorem 3.1.** *Let $\mathcal{K}_M = (\kappa, M)$ be a $d$-dimensional KDE on a differentiable kernel as defined in Section 2 and let $0 < \epsilon \leq 1$ be an approximation factor. Let $\xi \geq \xi_\kappa(\frac{\epsilon}{2n})$ and let $\kappa'_{\min} \leq \min_{0 \leq t \leq 2\xi} \frac{\kappa'(t)t}{\kappa(t)}$. Note that $\kappa'_{\min} \leq 0$ since $\kappa$ is assumed to be non-increasing. We can assume that $\kappa'_{\min} \neq 0$. Let $\gamma = -\frac{\epsilon}{2\kappa'_{\min}} > 0$. Then with probability $(1-\delta)$, for any $\Pi \in \mathbb{R}^{w \times d}$ satisfying the $(\gamma, n+1, \delta)$-JL guarantee, we have:*

$$(1-\epsilon) \max_{x \in \mathbb{R}^d} \mathcal{K}_M(x) \leq \max_{x \in \mathbb{R}^w} \mathcal{K}_{\Pi M}(x) \leq \max_{x \in \mathbb{R}^d} \mathcal{K}_M(x). \tag{2}$$

*Recall that a random $\Pi$ with $w = O\left(\frac{\log((n+1)/\delta)}{\min(1,\gamma^2)}\right)$ rows will satisfy the required $(\gamma, n+1, \delta)$-JL guarantee.*

*Proof.* First recall the definitions of $\bar{\mathcal{K}}_M(x)$ and $\bar{\mathcal{K}}_{\Pi M}(x)$, which are just fixed positive scalings of $\mathcal{K}_M(x)$ and $\mathcal{K}_{\Pi M}(x)$. It suffices to prove that:

$$(1-\epsilon) \max_{x \in \mathbb{R}^d} \bar{\mathcal{K}}_M(x) \leq \max_{x \in \mathbb{R}^w} \bar{\mathcal{K}}_{\Pi M}(x) \leq \max_{x \in \mathbb{R}^d} \bar{\mathcal{K}}_M(x), \tag{4}$$

To prove (4) we will apply the guarantee of Definition 2.2 to the $n+1$ points in $\{x^*\} \cup M$, where $x^* \in \operatorname{argmax}_{x \in \mathbb{R}^d} \bar{\mathcal{K}}_M(x)$. This guarantee ensures that with probability $(1-\delta)$, $\|a-b\|_2^2 \leq \|\Pi a - \Pi b\|_2^2 \leq (1+\gamma)\|a-b\|_2^2$ for all $a, b$ in this set, where $\Pi \in \mathbb{R}^{w \times d}$ is the JL matrix from the theorem.

We first prove the right hand side of (4) using an argument identical to the proof of Lemma 10 from (Lee et al., 2021). Consider the set of $n$ points $\Pi M$ that contains $\Pi m$ for all $m \in M$. Let $g$ be a map from each point in this set to the corresponding point in $M$. Since $\|\Pi m_1 - \Pi m_2\|_2^2 \geq \|m_1 - m_2\|_2^2$ for all $m_1, m_2 \in M$ as guaranteed above, we have that $g$ is 1-Lipschitz. From Kirszbraun's theorem (Fact 2.3) it follows that there is a function $\tilde{g} : \mathbb{R}^d \to \mathbb{R}^w$ which agrees with $g$ on inputs in $\Pi M$ and satisfies $\|\tilde{g}(s) - \tilde{g}(t)\|_2^2 \leq \|s-t\|_2^2$ for all $s, t \in \mathbb{R}^d$. So for any $x \in \mathbb{R}^w$, there is some $x' = \tilde{g}(x)$ such that, for all $m \in M$,

$$\|x' - m\|_2^2 \leq \|x - \Pi m\|_2^2.$$

The right hand side of (4) then follows: there must be some point $x'$ such that for all $m$, $\|x' - m\|_2^2 \leq \|\tilde{x}^* - \Pi m\|_2^2$ where $\tilde{x}^* \in \operatorname{argmax}_{x \in \mathbb{R}^w} \bar{\mathcal{K}}_{\Pi M}(x)$. Overall we have:

$$\max_{x \in \mathbb{R}^w} \bar{\mathcal{K}}_{\Pi M}(x) = \bar{\mathcal{K}}_{\Pi M}(\tilde{x}^*) = \sum_{m \in M} \kappa(\|\tilde{x}^* - \Pi m\|_2^2) \leq \sum_{m \in M} \kappa(\|x' - m\|_2^2) \leq \max_{x \in \mathbb{R}^d} \bar{\mathcal{K}}_M(x).$$

In the second to last inequality we used that $\kappa$ is non-increasing.

We next prove the left hand side of (4). We first have:

$$\max_{x \in \mathbb{R}^w} \bar{\mathcal{K}}_{\Pi M}(x) = \max_{x \in \mathbb{R}^w} \sum_{m \in M} \kappa(\|x - \Pi m\|_2^2) \geq \sum_{m \in M} \kappa(\|\Pi x^* - \Pi m\|_2^2) \geq \sum_{m \in M, \|x^* - m\|_2^2 \leq \xi} \kappa(\|\Pi x^* - \Pi m\|_2^2),$$

where $x^* \in \operatorname{argmax}_{x \in \mathbb{R}^d} \bar{\mathcal{K}}_M(x)$. Applying the JL guarantee to the $n+1$ points in $\{x^*\} \cup M$, we have that for all $m$, $\|\Pi x^* - \Pi m\|_2^2 \leq (1+\gamma)\|x^* - m\|_2^2$. So plugging into the equation above, we have:

$$\max_{x \in \mathbb{R}^w} \bar{\mathcal{K}}_{\Pi M}(x) \geq \sum_{m \in M, \|x^* - m\|_2^2 \leq \xi} \kappa((1+\gamma)\|x^* - m\|_2^2)$$

We can then bound:

$$
\begin{aligned}
&\sum_{m\in M, \|x^*-m\|_2^2\le\xi} \kappa((1+\gamma)\|x^*-m\|_2^2) \\
&\ge \sum_{m\in M, \|x^*-m\|_2^2\le\xi} \kappa(\|x^*-m\|_2^2) + \gamma\|x^*-m\|_2^2 \min_{z\in[\|x^*-m\|_2^2,(1+\gamma)\|x^*-m\|_2^2]} \kappa'(z) \\
&\ge \sum_{m\in M, \|x^*-m\|_2^2\le\xi} \kappa(\|x^*-m\|_2^2) + \gamma\cdot \min_{z\in[\|x^*-m\|_2^2,(1+\gamma)\|x^*-m\|_2^2]} \kappa'(z)\cdot z \\
&\ge \sum_{m\in M, \|x^*-m\|_2^2\le\xi} \kappa(\|x^*-m\|_2^2) + \gamma\cdot \min_{z\in[\|x^*-m\|_2^2,(1+\gamma)\|x^*-m\|_2^2]} \kappa'(z)\cdot z\cdot \frac{\kappa(\|x^*-m\|_2^2)}{\kappa(z)}.
\end{aligned}
$$

The last inequality follows from the fact that $\kappa$ is non-increasing, so $k'(z)\cdot z$ is negative or zero and $\frac{\kappa(\|x^*-m\|_2^2)}{\kappa(z)} \ge 1$. Invoking our definition of $\kappa'_{\min}$ and choice of $\gamma = -\frac{\epsilon}{2\kappa'_{\min}}$ we can continue:

$$
\begin{aligned}
&\sum_{m\in M, \|x^*-m\|_2^2\le\xi} \kappa(\|x^*-m\|_2^2) + \gamma\cdot \min_{z\in[\|x^*-m\|_2^2,(1+\gamma)\|x^*-m\|_2^2]} \kappa'(z)\cdot z\cdot \frac{\kappa(\|x^*-m\|_2^2)}{\kappa(z)} \\
&\ge \sum_{m\in M, \|x^*-m\|_2^2\le\xi} \kappa(\|x^*-m\|_2^2) + \gamma\cdot\kappa'_{\min}\cdot\kappa(\|x^*-m\|_2^2) \\
&= \sum_{m\in M, \|x^*-m\|_2^2\le\xi} \kappa(\|x^*-m\|_2^2) - \frac{\epsilon}{2}\cdot\kappa(\|x^*-m\|_2^2) \\
&= \left(1-\frac{\epsilon}{2}\right) \sum_{m\in M, \|x^*-m\|_2^2\le\xi} \kappa(\|x^*-m\|_2^2) \\
&= \left(1-\frac{\epsilon}{2}\right)\left(\sum_{m\in M}\kappa(\|x^*-m\|_2^2) - \sum_{m\in M, \|x^*-m\|_2^2>\xi}\kappa(\|x^*-m\|_2^2)\right) \\
&\ge \left(1-\frac{\epsilon}{2}\right)\left(\sum_{m\in M}\kappa(\|x^*-m\|_2^2) - \frac{\epsilon}{2}\right) \\
&= \left(1-\frac{\epsilon}{2}\right)\left(\max_{x\in\mathbb{R}^d}\bar{\mathcal{K}}_M(x) - \frac{\epsilon}{2}\right) \ge \left(1-\frac{\epsilon}{2}\right)^2 \max_{x\in\mathbb{R}^d}\bar{\mathcal{K}}_M(x) \ge (1-\epsilon)\max_{x\in\mathbb{R}^d}\bar{\mathcal{K}}_M(x)
\end{aligned}
$$

Note that in the second to last line we invoked the definition of $\xi \ge \xi_\kappa(\frac{\epsilon}{2n})$. Specifically, we used that, for any $m$ with $\|x^*-m\|_2^2 > \xi$, $\kappa(\|x^*-m\|_2^2) \le \frac{\epsilon}{2n}$. In the last line we use that $\max_{x\in\mathbb{R}^d}\bar{\mathcal{K}}_M(x) \ge 1$. □

**Lemma A.1.** *Let $\mathcal{K}_M = (\kappa, M)$ be a KDE for a point set $M$ with cardinality $n$ and relative-distance smooth kernel $\kappa$ with parameters $c_1, d_1, q_1, c_2, d_2$. Then for any $\epsilon \in (0,1]$, $\xi_\kappa(\frac{\epsilon}{2n}) \le c\log^{1/d_1}\left(\frac{2n}{\epsilon}\right)$ for a fixed constant $c$ that depends on $c_1, d_1, q_1, c_2$, and $d_2$.*

*Proof.* Since $\kappa$ is positive, non-increasing, and $\kappa(0) = 1$ we can write $\kappa(t) = e^{-f(t)}$ for some positive, non-decreasing function $f$ with $f(0) = 0$. We have $\frac{-\kappa'(t)t}{\kappa(t)} = f'(t)t \ge c_1 t^{d_1} - q_1$ and thus $f'(t) \ge \max(0, c_1 t^{d_1-1} - \frac{q_1}{t})$. Writing

$\kappa(t) = e^{-\int_0^t f'(x)dx}$, we will upper bound $\kappa(t)$ by lower bounding $\int_0^t f'(x)dx$. Specifically, we have:

$$\begin{aligned}\int_0^t f'(x)dx &\geq \int_0^t \max\left(0, c_1 x^{d_1-1} - \frac{q_1}{x}\right) dx \\ &= \int_0^t c_1 x^{d_1-1} dx - \int_0^t \min(c_1 x^{d_1-1}, \frac{q_1}{x}) dx \\ &= \frac{c_1}{d_1} t^{d_1} - \int_0^{(q_1/c_1)^{1/d_1}} c_1 x^{d_1-1} dx - \int_{(q_1/c_1)^{1/d_1}}^t \frac{q_1}{x} dx \\ &= \frac{c_1}{d_1} t^{d_1} - \frac{q_1}{d_1} - q_1 \log(t) + \frac{q_1}{d_1} \log(q_1/c_1).\end{aligned}$$

It follows that $\kappa(t) \leq e^{-\frac{c_1}{d_1} t^{d_1} + \frac{q_1}{d_1} + q_1 \log(t) - \frac{q_1}{d_1} \log \frac{q_1}{c_1}}$. We want to upper bound the smallest $t$ such that $\kappa(t) \leq \frac{\epsilon}{2n}$. Let $z$ be a sufficiently large constant so that:

$$\frac{c_1}{d_1} z^{d_1} \geq 2\left(\frac{q_1}{d_1} + q_1 \log(z) - \frac{q_1}{d_1} \log \frac{q_1}{c_1}\right)$$

Then it suffices to pick $t \geq \max\left(z, \left(\frac{d_1}{c_1} \log(\frac{2n}{\epsilon})\right)^{1/d_1}\right) = O\left(\log^{1/d_1}(\frac{2n}{\epsilon})\right)$ to ensure that $\kappa(t) \leq \frac{\epsilon}{2n}$. □

**Lemma 3.4.** *Let $\mathcal{K}_m$ be a KDE for a relative-distance smooth kernel $\kappa$ with parameters $c_1, d_1, q_1, c_2, d_2$. There is a fixed constant $c'$ such that if $\gamma = \frac{\epsilon}{c'} \log^{-d_2/d_1}\left(\frac{2n}{\epsilon}\right)$, then with probability $(1-\delta)$, for any $\Pi \in \mathbb{R}^{w \times d}$ satisfying the $(\gamma, n+1, \delta)$-JL guarantee, Equation (2) holds. To obtain this JL guarantee, it suffices to take $\Pi$ to be a random JL matrix with $w = O\left(\frac{\log^{2d_2/d_1}(n/\epsilon)\log(n/\delta)}{\epsilon^2}\right)$ rows.*

*Proof.* With Lemma A.1 in place, our main result for relatively distance smooth kernels follows directly: By Lemma A.1, $\xi = c \log^{1/d_1}\left(\frac{2n}{\epsilon}\right) \geq \xi_\kappa(\frac{\epsilon}{2n})$. And since $\kappa$ is relative-distance smooth, we have that:

$$\min_{0 \leq x \leq 2\xi} \frac{\kappa'(x)x}{\kappa(x)} \geq \min_{0 \leq x \leq 2\xi} -c_2 x^{d_2} = -c_2(2\xi)^{d_2} \geq -c' \log^{d_2/d_1}\left(\frac{2n}{\epsilon}\right),$$

for sufficiently large constant $c'$. Let $\kappa'_{\min} = -c' \log^{d_2/d_1}\left(\frac{2n}{\epsilon}\right)$. Invoking Theorem 3.1, we require that $\gamma = -\frac{\epsilon}{2\kappa'_{\min}} = \frac{\epsilon}{2c'} \log^{-d_2/d_1}\left(\frac{2n}{\epsilon}\right)$. □

### A.2. Proofs for Section 4

**Theorem 4.2.** *Let $M$ be a set of points in $\mathbb{R}^d$ and let $\mathcal{K}_M = (\kappa, M)$ be a KDE defined by a shift-invariant, non-increasing, and convex kernel function $\kappa$. Let $x^* \in \operatorname{argmax}_x \mathcal{K}_M(x)$. Let $\Pi \in \mathbb{R}^{w \times d}$ be a JL matrix as in Definition 2.2 and assume that $w$ is chosen large enough so that for all $a, b$ in the set $\{x^*\} \cup M$,*

$$\|a - b\|_2^2 \leq \|\Pi a - \Pi b\|_2^2 \tag{3}$$

$$\text{and } (1-\epsilon) \max_{x \in \mathbb{R}^d} \mathcal{K}_M(x) \leq \max_{x \in \mathbb{R}^w} \mathcal{K}_{\Pi M}(x) \leq \max_{x \in \mathbb{R}^d} \mathcal{K}_M(x).$$

*Let $\tilde{x} \in \mathbb{R}^w$ be an approximate maximizer for $\mathcal{K}_{\Pi M}$ satisfying $\mathcal{K}_{\Pi M}(\tilde{x}) \geq (1-\alpha) \max_{x \in \mathbb{R}^w} \mathcal{K}_{\Pi M}(x)$. Then if we choose $x' = \sum_{m \in M} m \cdot \frac{\kappa'(\|\tilde{x} - \Pi m\|^2)}{\sum_{m \in M} \kappa'(\|\tilde{x} - \Pi m\|^2)}$, we have:*

$$\mathcal{K}_M(x') \geq (1 - \epsilon - \alpha) \max_{x \in \mathbb{R}^d} \mathcal{K}_M(x).$$

*Proof.* We will show that:

$$\mathcal{K}_M(x') \geq \mathcal{K}_{\Pi M}(\tilde{x}) \tag{5}$$

where $\tilde{x} \in \mathbb{R}^w$ is the approximate maximizer of $\mathcal{K}_{\Pi M}$, as defined in the theorem statement. If we can prove (5) then the theorem follows by the following chain of inequalities:

$$\mathcal{K}_M(x') \geq \mathcal{K}_{\Pi M}(\tilde{x}) \geq (1-\alpha)\max_x \mathcal{K}_{\Pi M}(x) \geq (1-\alpha)(1-\epsilon)\max_x \mathcal{K}_M(x) \geq (1-\alpha-\epsilon)\max_x \mathcal{K}_M(x).$$

To prove (5), we follow a similar approach to (Lee et al., 2021). Since $\Pi$ satisfies (3), the function $f$ mapping $\{\Pi x^*\} \cup \Pi M \to \{x^*\} \cup M$ is 1-Lipschitz. Accordingly, by Kirszbraun's Extension Theorem (Fact 2.3), there is some function $g(x) : \mathbb{R}^w \to \mathbb{R}^d$ that agrees with $f$ on inputs from $\{\Pi x^*\} \cup \Pi M$ and remains 1-Lipschitz. It follows that, if we apply $g$ to $\tilde{x}$ then for all $m \in M$,

$$\|g(\tilde{x}) - m\|_2 \leq \|\tilde{x} - \Pi m\|_2 .$$

In other words, for our approximate low-dimensional mode $\tilde{x}$, there is a high-dimensional point $g(\tilde{x})$ that is closer to all points in $M$ than $\tilde{x}$ is to the points in $\Pi M$. In fact, by extending all points by one extra dimension, we can obtain an exact equality. In particular, let $x'' \in \mathbb{R}^{d+1}$ equal $g(\tilde{x})$ on its first $d$ coordinates, and 0 on its last coordinate. For each $m \in M$ let $m'' \in \mathbb{R}^{d+1}$ equal $m$ on its first $d$ coordinates, and $\sqrt{\|\tilde{x} - \Pi m\|_2^2 - \|g(\tilde{x}) - m\|_2^2}$ on its last coordinate. Let $M'' \subset \mathbb{R}^{d\times 1}$ denote the set of these transformed points. First observe that for all $m'' \in M''$

$$\|x'' - m''\|_2 = \|\tilde{x} - \Pi m\|_2 . \tag{6}$$

Accordingly, $x'$ as defined in the theorem statement is exactly equivalent to the first $d$ entries of the $d+1$ dimensional vector that would be obtained from running one iteration of mean-shift on $x''$ using point set $M''$. Call this $d+1$ dimensional vector $\bar{x}'$. By Fact 4.1, we have that:

$$\mathcal{K}_{M''}(\bar{x}') \geq \mathcal{K}_{M''}(x'') = \mathcal{K}_{\Pi M}(\tilde{x}).$$

The last equality follows from (6). Finally, for any non-increasing kernel we have that:

$$\mathcal{K}_M(x') \geq \mathcal{K}_{M''}(\bar{x}'),$$

because $\|x' - m\|_2^2 \leq \|\bar{x}' - m''\|_2^2$ for all $m$. This is simply because $x'$ and $m$ are equal to $\bar{x}'$ and $m''$, but with their last entry removed, so they can only be closer together. Combining the previous two inequalities proves (5), which establishes the theorem. □

**Corollary 4.3.** *Let $\mathcal{K}_M = (\kappa, M)$ be a $d$-dimensional shift-invariant KDE as defined in Section 2 and let $\epsilon$ and $\gamma$ (which depends on $\kappa$ and $\epsilon$) be as in Theorem 3.1. If $\kappa$ is differentiable, non-increasing, and convex, then Algorithm 2 run with parameters $\gamma$ and $\alpha$ returns $x'$ satisfying:*

$$\mathcal{K}_M(x') \geq (1-\epsilon-\alpha)\max_{x\in\mathbb{R}^d} \mathcal{K}_M(x).$$

*Proof.* Corollary 4.3 immediately follows by combining Theorem 3.1 with Theorem 4.2. In particular, if $\Pi$ is chosen with $w = O\left(\frac{\log((n+1)/\delta)}{\min(1,\gamma^2)}\right)$ rows (as in Algorithm 2) then with probability $1-\delta$, we have that both (2) and (3) hold with probability $1-\delta$, which are the only conditions needed for Theorem 4.2 to hold. □

### A.3. Proofs for Section 5

**Lemma 5.1.** *Given a KDE $\mathcal{K} = (\kappa, M)$ in $\mathbb{R}^d$ with $|M| = n$, and parameter $\delta > 0$, let $\xi = \xi_\kappa(1/n)$ and let $\mathcal{N}(\mathcal{K}, \delta)$ be a set that contains all points of the form*

$$m + \sum_{i=1}^{d} \frac{k_i\sqrt{\delta}}{\sqrt{d}} e_i,$$

*where $m \in M$, $k_i \in \mathbb{Z}$, and $-\frac{\sqrt{d}\xi}{\sqrt{\delta}} \leq k_i \leq \frac{\sqrt{d}\xi}{\sqrt{\delta}}$. Above $e_i$ are the canonical base vectors of $\mathbb{R}^d$. Then for any point $p$ in a $\xi$-ball surrounding one of the points in $M$, there exists a point $p' \in \mathcal{N}(\mathcal{K}, \delta)$ such that $\|p - p'\|_2^2 \leq \delta$. Moreover, we have that $|\mathcal{N}(\mathcal{K}, \delta)| = n(2\sqrt{d}\xi/\sqrt{\delta})^d$.*

*Proof.* The second claim on the size of $\mathcal{N}(\mathcal{K}, \delta)$ is immediate. For the first claim, note that $p$ can be written as $p'' + \sum_i \frac{k_i'\sqrt{\delta}}{\sqrt{d}} e_i$ with $p'' \in M$ and $|k_i'| \leq \frac{\sqrt{d}\xi}{\sqrt{\delta}}$. Let $p' = p'' + \sum_i \frac{\lfloor k_i' \rfloor \sqrt{\delta}}{\sqrt{d}} e_i \in \mathcal{N}(\mathcal{K}, \delta)$. Then we have

$$\|p - p'\|_2^2 = \left\| p'' + \sum_i \frac{k_i'\sqrt{\delta}}{\sqrt{d}} e_i - p'' - \sum_i \frac{\lfloor k_i' \rfloor \sqrt{\delta}}{\sqrt{d}} e_i \right\|_2^2 = \left\| \sum_{i=1}^d \frac{(k_i' - \lfloor k_i' \rfloor)\sqrt{\delta}}{\sqrt{d}} e_i \right\|^2$$
$$= \sum_{i=1}^d \left( \frac{(k_i' - \lfloor k_i' \rfloor)\sqrt{\delta}}{\sqrt{d}} \right)^2 \leq \sum_{i=1}^d \left( \frac{\sqrt{\delta}}{\sqrt{d}} \right)^2 = \delta.$$

□

**Theorem 5.2.** *Given a KDE $\mathcal{K} = (\kappa, M)$ in $\mathbb{R}^d$ with $|M| = n$ and a precision parameter $\epsilon > 0$, let $\xi = \xi_\kappa(1/n)$ and let $\delta$ be at most the largest number such that $\kappa(c) - \kappa(c + \delta) \leq \epsilon\kappa(c)$ for all $c \leq \xi$. Then we can find an $\epsilon$-approximate mode in $O\left(n(2\sqrt{d}\xi/\sqrt{\delta})^d\right)$. In particular, if $d \leq O(\log^c(n))$, $\xi \leq O(n^c)$, and $\delta \geq O(n^{-c})$ for some constant $c$, then we can find an $\epsilon$-approximate mode in quasi-polynomial time in the number of data points $n$.*

*Proof.* Since there always exists a mode in the critical area of $\mathcal{K}$, we can use Lemma 5.1 to find a point $p'$ at most $\delta$ away from a mode $p$ of $\mathcal{K}$ in $O\left(n(2\sqrt{d}\xi/\sqrt{\delta})^d\right)$. Then we have

$$\mathcal{K}(p') = \sum_{m \in M} \kappa(\|m - p'\|_2^2) \geq \sum_{m \in M} \kappa(\|m - p\|_2^2 + \|p - p'\|_2^2) \geq \sum_{m \in M} \kappa(\|m - p\|_2^2 + \delta)$$
$$\geq \sum_{m \in M} \kappa(\|m - p\|_2^2)) - \epsilon\kappa(\|m - p\|_2^2)) = (1 - \epsilon) \sum_{m \in M} \kappa(\|m - p\|_2^2)) = (1 - \epsilon)\mathcal{K}(p)$$

□

### A.4. Analysis for Relative-Distance Smooth Kernels

Let $\bar{\xi} = \max(1, \xi_\kappa(1/n))$. We will prove that for any relative distance smooth kernel $\kappa$ with parameters $c_1, d_1, q_1, c_2$, and $d_2$, we have $\kappa(c) - \kappa(c + \delta) \leq \epsilon\kappa(c)$ for all $c \leq \xi = \xi_\kappa(1/n)$ as long as:

$$\delta = \min\left( \left( \frac{d_2}{c_2} \epsilon \right)^{1/d_2}, \frac{\epsilon}{c_2} (2\bar{\xi})^{1-d_2} \right).$$

By the definition of relative distance smooth kernels, we have that $-\kappa'(t) \leq c_2 t^{d_2 - 1}\kappa(t)$. Hence,

$$\kappa(c) - \kappa(c + \delta) \leq \int_c^{c+\delta} c_2 t^{d_2 - 1}\kappa(t)dt \leq \kappa(c) \int_c^{c+\delta} c_2 t^{d_2 - 1} dt.$$

The last step follows from the fact that $\kappa(t)$ is non-increasing in $t$. Since $d_2 > 0$, this simplifies to

$$\kappa(c) - \kappa(c + \delta) \leq \kappa(c) \int_c^{c+\delta} c_2 t^{d_2 - 1} dt = \frac{c_2}{d_2}\kappa(c)((c + \delta)^{d_2} - c^{d_2}).$$

So, we need to show that $\frac{c_2}{d_2}\left((c + \delta)^{d_2} - c^{d_2}\right) \leq \epsilon$. We consider two cases:

**Case 1:** When $d_2$ is $< 1$, consider the function $f(x) = (x + \delta)^{d_2} - x^{d_2}$. This function is non-increasing, indeed, $f'(x) = d_2((x + \delta)^{d_2 - 1} - x^{d_2 - 1}) < 0$. Hence, we have that

$$((c + \delta)^{d_2} - c^{d_2}) \leq \delta^{d_2}.$$

We can pick $\delta = (\frac{d_2}{c_2}\epsilon)^{1/d_2}$.

**Case 2:** On the other hand, when $d_2 \geq 1$, the function $f(x) = x^{d_2}$ is convex, so we have:

$$(c + \delta)^{d_2} - c^{d_2} \leq \delta f'(c + \delta) = \delta d_2 (c + \delta)^{d_2 - 1} \leq \delta d_2 2^{d_2 - 1} \max(\xi^{d_2 - 1}, \delta^{d_2 - 1}) \leq \delta d_2 2^{d_2 - 1} \bar{\xi}^{d_2 - 1}$$

In this case, we can choose $\delta = \frac{\epsilon}{c_2}(2\bar{\xi})^{1-d_2}$.

Hence, picking $\delta = \min\left(\left(\frac{d_2}{c_2}\epsilon\right)^{1/d_2}, \frac{\epsilon}{c_2}(2\bar{\xi})^{1-d_2}\right)$ ensures that $(c+\delta)^{d_2} - c^{d_2} \leq \frac{d_2}{c_2}\epsilon$, and thus that $\frac{c_2}{d_2}\left((c+\delta)^{d_2} - c^{d_2}\right) \leq \epsilon$, as required.

## B. Mode Recovery for Non-convex Kernels

In Section 4.1, we show that the mean-shift method can rapidly recover an approximate mode for any convex, non-increasing kernel from an approximation to the JL reduced problem. In this section, we briefly comment on an alternative method that can also handle non-convex kernels, albeit at the cost of worse runtime. Specifically, it is possible to leverage a recent result from (Biess et al., 2019) on an algorithmic version of the Kirszbraun extension theory. This work provides an algorithm for explicitly extending a function $f$ that is Lipschitz on some fixed set of points to *one additional* point. The main result follows:

**Theorem B.1** ((Biess et al., 2019))**.** *Consider a finite set $(x_i)_{i\in[n]} \subset X = \mathbb{R}^w$, and its image $(y_i)_{i\in[n]} \subset Y = \mathbb{R}^d$ under some $L$-Lipschitz map $f : X \to Y$. There is an algorithm running in $O(nw + nd\log n/\epsilon^2)$ time which returns, for any point $z \in \mathbb{R}^w$, and a precision parameter $\epsilon > 0$, a point $z' \in \mathbb{R}^d$ satisfying for all $i \in [n]$,*

$$\|z' - f(x_i)\|^2 \leq (1+\epsilon)L\|z - x_i\|^2$$

From this result we can obtain a claim comparable to Corollary 4.3:

**Theorem B.2.** *Let $\mathcal{K}_M = (\kappa, M)$ be a $d$-dimensional shift-invariant KDE where $\kappa$ is differentiable and non-increasing but not necessary convex. Let $\epsilon$ and $\gamma$ (which depends on $\kappa$ and $\epsilon$) be as in Theorem 3.1 and let $\Pi$ be a random JL matrix as in Definition 2.2 with $w = O\left(\frac{\log((n+1)/\delta)}{\min(1,\gamma^2)}\right)$ rows. Let $\tilde{x}$ be an approximate maximizer for $\mathcal{K}_{\Pi M}$ satisfying $\mathcal{K}_{\Pi M}(\tilde{x}) \geq (1-\alpha)\max_{x\in\mathbb{R}^w}\mathcal{K}_{\Pi M}(x)$. If we run the algorithm of Theorem B.1 with $X = \Pi M$, $Y = M$, $z = \tilde{x}$, and error parameter $\gamma$, then the method returns $x' \in \mathbb{R}^d$ satisfying:*

$$\mathcal{K}_M(x') \geq (1 - 2\epsilon - \alpha)\max_{x\in\mathbb{R}^w}\mathcal{K}_M(x).$$

*Proof.* For conciseness, we sketch the proof. As discussed, by Definition 2.2, $\Pi M \to M$ is a 1-Lipschitz map. So it follows that the $x'$ returned by the algorithm of (Biess et al., 2019) satisfies for all $m \in M$,

$$\|x' - m\|^2 \leq (1+\gamma)\|\tilde{x} - \Pi m\|^2$$

It follows that:

$$\mathcal{K}_M(x') \geq \sum_{m\in M}\kappa\left((1+\gamma)\|\tilde{x} - \Pi m\|^2\right).$$

By the same argument used in the proof of Theorem 3.1, we have that

$$\sum_{m\in M}\kappa\left((1+\gamma)\|\tilde{x} - \Pi m\|^2\right) \geq \sum_{m\in M}(1-\epsilon)\kappa\left(\|\tilde{x} - \Pi m\|^2\right) = (1-\epsilon)\mathcal{K}_{\Pi M}(\tilde{x}).$$

In turn, since Theorem 3.1 holds under the same conditions as Theorem B.2, we have:

$$\mathcal{K}_{\Pi M}(\tilde{x}) \geq (1-\alpha)\max_x \mathcal{K}_{\Pi M}(x) \geq (1-\alpha)(1-\epsilon)\max_x \mathcal{K}_M(x).$$

The result follows from noting that $(1-\alpha)(1-\epsilon)^2 \geq (1 - 2\epsilon - \alpha)$. By rescaling $\epsilon$ we can obtain equivalent precision to Corollary 4.3. □

Note that for the common relative-distance smooth kernels addressed in Theorem 1.1, we have that $\gamma = O(\log(n/\epsilon)/\epsilon)$. So, the runtime of recovering a high-dimensional model using the method of (Biess et al., 2019) is $O(nd\log^3(n)/\epsilon^2)$. This exceeds the $O(nd)$ runtime of the mean-shift method. However, in contrast to mean-shift, the method can be applied to non-convex kernels like generalized Gaussian kernels with $\alpha > 1$.