On the number of modes of Gaussian kernel density estimators

Borjan Geshkovski, Philippe Rigollet, Yihang Sun

On the number of modes of Gaussian kernel density estimators

Borjan Geshkovski
Inria & Sorbonne Université

Philippe Rigollet
MIT

Yihang Sun
Stanford University

December 13, 2024

Abstract

We consider the Gaussian kernel density estimator with bandwidth $\beta^{-\frac{1}{2}}$ of n iid Gaussian samples. Using the Kac-Rice formula and an Edgeworth expansion, we prove that the expected number of modes on the real line scales as $\Theta(\sqrt{\beta \log \beta})$ as $\beta, n \to \infty$ provided $n^c \lesssim \beta \lesssim n^{2-c}$ for some constant $c > 0$. An impetus behind this investigation is to determine the number of clusters to which Transformers are drawn in a metastable state.

Keywords. Kernel density estimator, Kac-Rice formula, Edgeworth expansion, self-attention, mean-shift.

AMS classification. 62G07, 60G60, 60F05, 68T07.

Contents

1 Introduction

1.1 Setup and main result

For $\beta > 0$ and $X_1, \dots, X_n \overset{\text{iid}}{\sim} N(0,1)$, the *Gaussian kernel density estimator (KDE)* with bandwidth $h = \beta^{-1/2}$ is defined as

$$\widehat{P}_n(t) := \frac{1}{n}\sum_{i=1}^{n} \mathsf{K}_h * \delta_{X_i}(t) = \frac{\sqrt{\beta}}{n\sqrt{2\pi}} \sum_{i=1}^{n} e^{-\frac{\beta}{2}(t-X_i)^2}, \qquad t \in \mathbb{R}. \tag{1.1}$$

Here, "Gaussian" refers to the choice of kernel K_h.

In this paper we are interested in determining the expected number of modes (local maxima) of $\widehat{P}_n$ over $\mathbb{R}$. While this is a classical question, addressed in even more general settings than (1.1)—such as non-Gaussian kernels, compactly supported samples, and higher dimensions [MMF92, Mam95, KM97]—a definite answer has not been given in the literature. Indeed, the best-known results fall into one of two settings: either considering samples drawn from a compactly supported density (instead of $N(0,1)$ as done here), or counting the modes within a fixed compact interval. In the special case of the Gaussian KDE (1.1), one has in the latter setting for instance

Theorem 1.1 ([Mam95, Thm. 1]). *Let $\widehat{P}_n$ be the Gaussian KDE defined in (1.1), with bandwidth $h > 0$, of $X_1, \dots, X_n \overset{iid}{\sim} N(0,1)$. Asymptotically as $n \to \infty$, the expected number of modes of $\widehat{P}_n$ in a fixed interval $[a,b]$ is $1+o(1)$ if $1 \ll \beta \ll n^{2/3}$, and $\widetilde{\Theta}(\sqrt{\beta})$ if $n^{2/3} \lesssim \beta \ll n^2/\log^6 n$.*

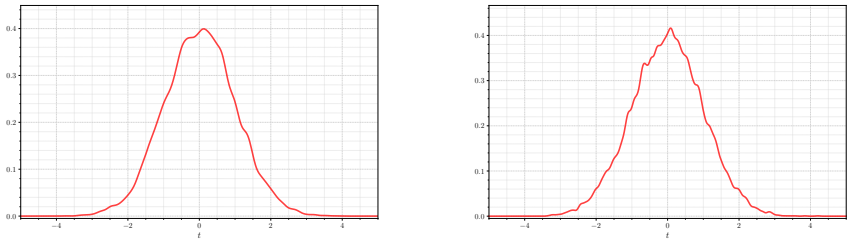


Figure 1: A realization of the random function (1.1) for $n = 10^4$, with $\beta = 100$ (left) and $\beta = 300$ (right).

In [MMF92, Mam95, KM97], the authors additionally conduct more refined casework on the bandwidth to provide more precise estimates, such as pinpointing the leading constants. In fact, [MMF92] *does* count modes in $\mathbb{R}$, but the underlying distribution of the samples X_i is supported on a closed interval (thus excluding $N(0,1)$), so there are no modes outside the interval anyway.

Our main result provides the answer to the case of counting modes of (1.1) over $\mathbb{R}$, and reads as follows. In particular, it generalizes Theorem 1.1.

Theorem 1.2. *Let $\widehat{P}_n$ be the Gaussian KDE defined in (1.1), with bandwidth $\beta^{-1/2}$, of $X_1, \dots, X_n \overset{iid}{\sim} N(0,1)$. Suppose $n^c \lesssim \beta \lesssim n^{2-c}$ for arbitrarily small $c > 0$, and Assumption 4.2, then asymptotically as $n, \beta \to \infty$,*

1. *In expectation over X_i, the number of modes of $\widehat{P}_n$ is $\Theta(\sqrt{\beta \log \beta})$.*
2. *Almost all modes lie in two intervals of length $\Theta(\sqrt{\log \beta})$—namely, the expected number of modes $t \in \mathbb{R}$, such that $t^2 \notin [2\log n - 3\log\beta, 2\log n - \log\beta]$, is $o(\sqrt{\beta \log \beta})$.*

Assumption 4.2 is related to the decay of the tails of the modulus of continuity of $\widehat{P}_n''(\cdot)$, and is needed to apply the *Kac-Rice formula* (Theorem 2.1), which states that the expected number of modes of $\widehat{P}_n$ is some conditional expectation with respect to the joint law of $(\widehat{P}_n'(t), \widehat{P}_n''(t))$. We postpone a further discussion to Section 4.1.

Remark 1.3. • *To better appreciate the range of values for β in this theorem as well as subsequent ones, we use minimax theory as a benchmark; see, e.g., [Tsy09]. The reparametrization $h = \sqrt{\beta}$ is motivated by the connection to the Transformer model described in Section 1.2. Using an optimal bias-variance tradeoff [Tsy09, Chapter 1], we see that the optimal scaling of the bandwidth parameter h depends on the smoothness of the underlying density of interest: if the underlying density has s bounded (fractional) derivatives, then the optimal choice of h is given by $h \asymp n^{-\frac{1}{2s+1}}$. This gives $\beta \asymp n^{\frac{2}{2s+1}}$. For $s \in (0, \infty)$, we get $\beta \in [n^c, n^{2-c}]$ for some $c > 0$. In particular, the transition of the number of modes from 1 to $\sqrt{\beta}$ in Theorem 1.1 is achieved for $\beta \approx n^{2/3}$, which is the optimal choice for Lipschitz*

densities. The message of our main Theorem 1.2 below is that this scaling in $\sqrt{\beta}$ is the prevailing one for the whole range $\beta \in [n^c, n^{2-c}]$ if one does not restrict counting modes in a bounded interval $[a, b]$.

- *Point 2. in Theorem 1.2 shows that most of the modes are at distance at least $C \log n$ from the origin provided $\beta > n^{\frac{2-C}{3}}$ for $C > 0$ small. This corresponds to a choice of a bandwidth adapted to smoothness $s < 1$. This result is in agreement with and completes the picture drawn by Theorem 1.1.*

Remark 1.4. *Through refined computations, one can determine the modes in the regime $1 \ll \beta \ll n^2$ and also pinpoint the leading constant. For the sake of clarity, we stick to the regime where*

$$2 \log n - \log \beta \asymp \log \beta \asymp \log n,$$

and comment on how to do expand the regime in appropriate places.

Remark 1.5. *We further motivate Point 2. in Theorem 1.2 by considering a qualitative picture of the distribution of the modes displayed in Figure 4.*

- *Near the origin, we find most of the samples X_i and they are densely packed in the shape of a Gaussian. The corresponding Gaussian summands in (1.1) cancel to create one mode, as shown already in Theorem 1.1.*
- *In the two intervals of length $\Theta(\sqrt{\log \beta})$, the samples X_i are separated enough that the corresponding Gaussian summands do not cancel, but rather form $\Omega(\sqrt{\log \beta})$* isolated bumps*, as discussed in more generality in [DG85, Section 9.3].*
- *Further away at the tails, the phenomena of isolated bumps occur, but there are so few samples X_i that the number of modes created is a negligible fraction.*

We revisit this discussion and Figure 4 in Remark 3.3.

To prove Theorem 1.2, we truncate the real line to the interval

$$T := \left[-\sqrt{2 \log n - \log \beta - \omega(\beta)}, \sqrt{2 \log n - \log \beta - \omega(\beta)} \right] \tag{1.2}$$

where ω is a fixed, slow growing function such that

$$1 \ll \omega(\beta) \ll \log \log \beta,$$

and so T is well-defined for large enough β. Motivated by Theorem 1.2, we also define the interval

$$T' := \left[-\sqrt{2 \log n - 3 \log \beta}, \sqrt{2 \log n - 3 \log \beta} \right] \tag{1.3}$$

if $\beta \leqslant n^{2/3}$ and define $T' = \varnothing$ if $\beta > n^{2/3}$.

We see how Theorem 1.2 implies Theorem 1.1 with $h = \beta^{-1/2}$. From the former, we see that almost all the modes lie in $T \setminus T'$. If $h \gg n^{-1/3}$ so that $\sqrt{2\log n - 3\log\beta} \gg 1$, then $T \setminus T'$ is disjoint from $[a, b]$, so there are few modes in $[a, b]$; if $h \ll n^{-1/3}$ so $\sqrt{2\log n - 3\log\beta} \ll 1$, the fixed interval $[a, b]$ contains a near constant fraction of length of $T \setminus T'$ and thus a near constant fraction of the modes.

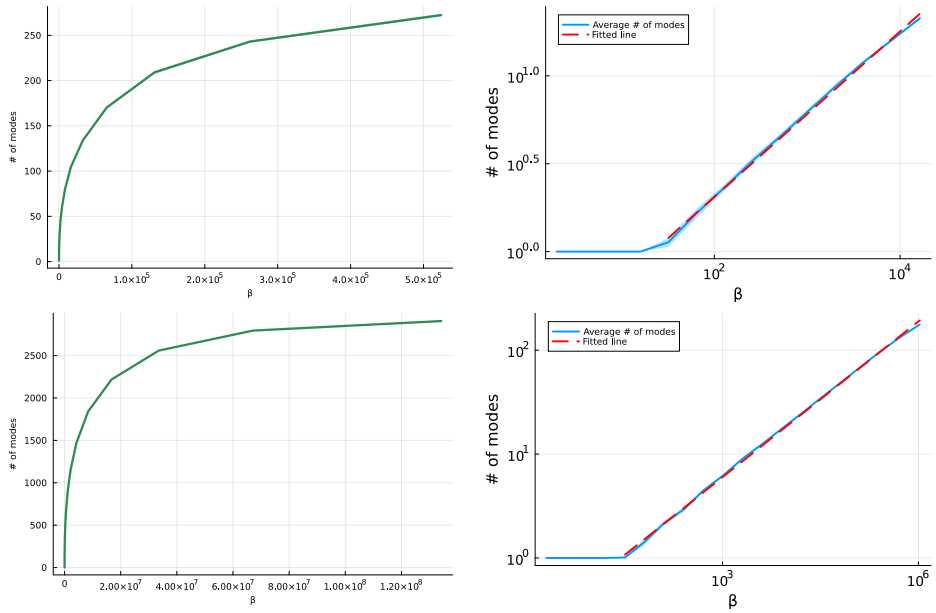


Figure 2: (Left) Plot of the average number of modes as a function of β for $n = 10^3$ (top) and $n = 10^4$ (bottom). (Right) Log-log plot for $n = 10^3$ (top) and $n = 10^4$ (bottom); the predicted linear regression line (red) corroborates a power-law of the form average # of modes $\approx 0.179 \cdot \beta^{0.504}$, in line with Theorem 1.2.

1.2 Motivation

The question of estimating the number of modes as a function of the bandwidth has a plethora of applications in statistical inference and multimodality tests—see [MMF92, Mam95, KM97] and the references therein. Another application which has stimulated some of the recent progress on the topic is data clustering. The latter can be achieved nonparametrically using a KDE, whose modes, and hence clusters, can be detected using the *mean-shift algorithm* [FH75, Che95, CM02, CP00, CPW03, CP07, RL14, CP15], which can essentially be seen as iterative local averaging. The main idea in mean-shift clustering is to perform a mean-shift iteration starting from each data point and then define each mode as a cluster, with all points converging to the same mode grouped into the same cluster. The analysis of this algorithm has led to upper bounds on the number of modes of (1.1) [CPW03].

We were instead brought to this problem from another perspective, motivated

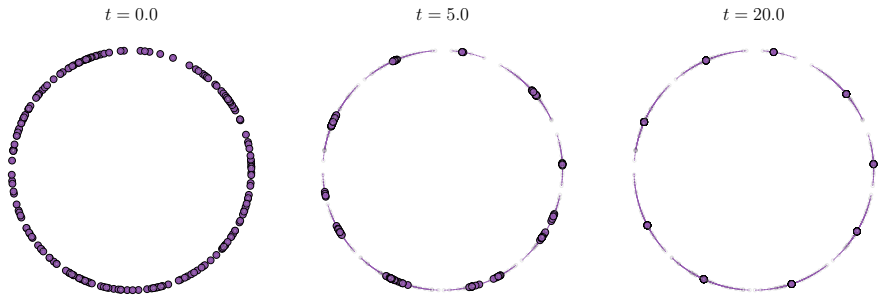
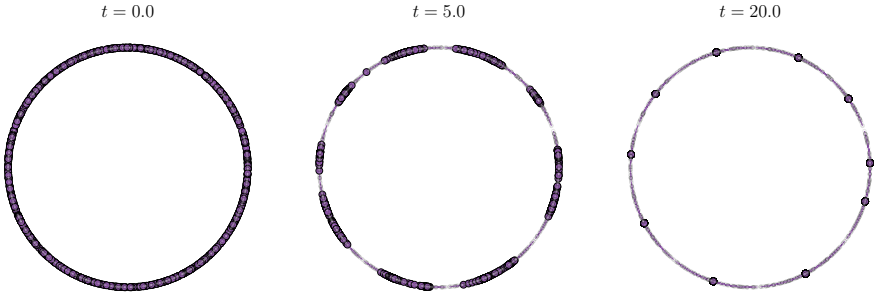


Figure 3: Metastability of self-attention dynamics at temperature $\beta = 81$ initialized with n iid uniform points on the circle, with $n = 200$ (top) and $n = 1000$ (bottom). The number of clusters appears of the correct order $\sim \sqrt{\beta}$. (Code available at github.com/borjanG/2023-transformers-rotf.)



by the study of *self-attention dynamics* [SABP22,GLPR23,GLPR24,GRRB24]—a toy model for *Transformers*, the deep neural network architecture that has driven the success of large language models [VSP+17]. These dynamics form a mean-field interacting particle system

$$\frac{\mathrm{d}}{\mathrm{d}\tau} x_i(\tau) = \sum_{j=1}^{n} \frac{e^{\beta\langle x_i(\tau), x_j(\tau)\rangle}}{\sum_{k=1}^{n} e^{\beta\langle x_i(\tau), x_j(\tau)\rangle}} \mathsf{P}^{\perp}_{x_i(\tau)}(x_j(\tau)),$$

evolving on the unit sphere $\mathbb{S}^{d-1}$ because of $\mathsf{P}^{\perp}_{x} := I_d - xx^{\top}$. Here, $\tau \geqslant 0$ plays the role of layers, the n particles $x_i(\tau)$ represent tokens evolving through a dynamical system. This system is characterized by a temperature parameter $\beta \geqslant 0$ that governs the space localization of particle interations. One sees that all particles move in time by following the field $\nabla \log(\mathsf{K}_{\beta^{-1/2}} * \mu_\tau)$; here, μ_τ is the empirical measure of the particles $x_1(\tau), \ldots, x_n(\tau)$ at time τ.

It is shown that for almost every initial configuration $x_1(0), \ldots, x_n(0)$, and for $\beta \geqslant 0$ in dimension $\geqslant 3$, or $\beta \leqslant 1 \vee \beta \gtrsim n^2$ in dimension 2, all particles converge to a single cluster in infinite time [GLPR23]. Rather than converging quickly, the authors in [GKPR24] prove that the dynamics instead manifest *metastability*: particles quickly approach a few clusters, remain in the vicinity

of these clusters for a very long period, and eventually coalesce into a single cluster in infinite time. Concurrently, and using different methods, the authors in [BPA24] show a similar result: starting from a perturbation of the uniform distribution, beyond a certain time, the empirical measure of the n particles approaches an empirical measure of $O(\sqrt{\beta})$-equidistributed points on the circle in the mean-field limit, and stays near it for long time. This is done by a study of the linearized system and leveraging nonlinear stability results from [Gre00].

Our interest lies in counting the number of clusters during the first metastable phase in dimension $d = 2$. At time $\tau = 0$, particles are initialized at n iid points from the uniform distribution on the circle. In turn, the stationary points of $\mathsf{K}_{\beta^{-1/2}} * \mu_0$ partition the circle into intervals, with points clustering within their respective interval. This highlights the importance of counting the number of stationary points.

Here, we focus on a simplified setting by working on the real line instead of the circle (or higher-dimensional spheres), but we believe the analysis could be extended to these cases pending technical adaptations. Notwithstanding, Theorem 1.2 reflects what is seen in simulations (Figure 3).

1.3 Sketch of the proof

The spirit of the proof of results such as Theorem 1.1 and others presented in [MMF92,Mam95,KM97] is similar to ours—one applies the Kac-Rice formula (Theorem 2.1) to a Gaussian approximation of $(\widehat{P}_n'(t), \widehat{P}_n''(t))$ and argues its validity. However, the main limitation of these works is that modes are counted in a fixed and finite interval $[a, b]$ (and $[0, 1]^d$ in the higher dimensional cases). Extending these techniques to the whole real line demands for different, significantly stronger, approximation results using Edgeworth expansions.

We sketch the key ideas that allow us to count modes over $\mathbb{R}$. We use the Kac-Rice formula to compute the expected number of modes of $\widehat{P}_n$ in the symmetric interval T and T' defined in (1.2) and (1.3). All asymptotics are as $n, \beta \to \infty$.

Proposition 1.6. *If $n^c \lesssim \beta \lesssim n^{2-c}$ for arbitrarily small $c > 0$, then under Assumption 4.2,*

1. *In expectation over X_i, the number of modes of $\widehat{P}_n$ in T is $\Theta\big(\sqrt{\beta \log \beta}\big)$.*
2. *In expectation over X_i, the number of modes of $\widehat{P}_n$ in T' is $O\left(e^{-\frac{\omega(\beta)}{4}} \sqrt{\beta \log \beta}\right)$.*

The Kac-Rice computation appears tractable only when the joint distribution of $(\widehat{P}_n'(t), \widehat{P}_n''(t))$ is Gaussian, which it is not. To overcome this obstacle, we apply the Kac-Rice formula over a Gaussian approximation of the joint distribution in Section 2. For the specific underlying density and KDE in (1.1), we are able to justify in Section 3 the approximation for all t in the growing interval T instead of a fixed interval. This is why Theorem 1.1 only counts modes in a fixed interval.

To show the validity of the Gaussian approximation, we use the *Edgeworth expansion* of the joint distribution of $(\widehat{P}_n'(t), \widehat{P}_n''(t))$ around the Gaussian distribution with matching first two moments. We bound the error due to the third order term of the expansion directly, and deal with the higher order terms by appealing to the error bounds of densities in the Edgeworth approximation similar to [BR10, Theorem 19.2]. We note that [KM97] employ the same theorem to justify the Gaussian process approximation over a fixed interval.

In doing so, we will see that the normal approximation is invalid outside of T (see Remark 3.3), but crucially T is sufficiently large to cover almost all modes, as observed empirically in Remark 1.5 and Figure 4 and given below.

Proposition 1.7. *If $n^c \lesssim \beta \lesssim n^{2-c}$ for arbitrarily small $c > 0$, then the expectation over X_i of the number of modes of $\widehat{P}_n$ that lie outside of T is $O\left(\sqrt{\beta \exp(\omega(\beta))}\right)$.*

We prove this in Section 4.2 with an argument from scale-space theory: we bound the number of modes outside T by the number of samples X_i outside T, which we then bound naively. This is precisely the argument used by [CPW03, Theorem 2] to show Gaussian mixtures over $\mathbb{R}$ with n components must have at most n modes. This argument crucially relies on the kernel density estimator being Gaussian (see Remark 4.5). Now, Theorem 1.2 follows from Propositions 1.6 and 1.7 upon recalling that $1 \ll \omega(\beta) \ll \log\log\beta$.

1.4 Notation

We adopt standard notation from asymptotic analysis: we write $f(x) \ll g(x)$ or $f(x) = o(g(x))$ if $f(x)/g(x) \to 0$ as $x \to \infty$; $f(x) \lesssim g(x)$ or $f(x) = O(g(x))$ if there exists a finite, positive constant C such that $f(x) \leqslant Cg(x)$; and we write $f(x) \asymp g(x)$ or $f(x) = \Theta(g(x))$ if $f(x) \lesssim g(x)$ and $g(x) \lesssim f(x)$.

2 Kac-Rice for the normal approximation

2.1 The Kac-Rice formula

We say that $\Psi : \mathbb{R} \to \mathbb{R}$ has an *upcrossing of level u* at $t \in \mathbb{R}$ if $\Psi(t) = u$ and $\Psi'(t) > 0$. The Kac-Rice formula allows us to compute the expected number of up-crossings when F is a random field (i.e., a stochastic process).

Theorem 2.1 (Kac-Rice, [AW09, pp. 62], [AT09, Section 11.1])**.** *Consider a random $\Psi : \mathbb{R} \to \mathbb{R}$, some fixed $u \in \mathbb{R}$ and a compact $T \subset \mathbb{R}$. Suppose*

1. *Ψ is a.s. in $\mathcal{C}^1(\mathbb{R})$, and Ψ, Ψ' both have finite variance over T;*
2. *The law of $\Psi(t)$ admits a density $p_t^{[1]}(x)$ which is continuous for $t \in T$ and x in a neighborhood of u;*
3. *The joint law of $(\Psi(t), \Psi'(t))$ admits a density $p_t(x, y)$ which is continuous for $t \in T$, x in a neighborhood of u, and $y \in \mathbb{R}$;*

4. $\mathbb{P}(\omega(\eta) > \varepsilon) = O(\eta)$ *as* $\eta \to 0^+$ *for any* $\varepsilon > 0$, *where* $\omega(\cdot)$ *denotes the modulus of continuity*[1] *of* $\Psi'(\cdot)$.

Define the number of up-crossings in T *of* Ψ *at level* $u \in \mathbb{R}$ *as*

$$U_u(\Psi, T) := \big|\{t \in T : \Psi(t) = u, \Psi'(t) > 0\}\big|.$$

Then, with expectation taken over the randomness of Ψ,

$$\mathbb{E}U_u(\Psi, T) = \int_T \int_0^\infty y p_t(u, y)\, \mathrm{d}y\, \mathrm{d}t. \tag{2.1}$$

The Kac-Rice formula extends to any dimension $d \geqslant 1$, and also on manifolds other than $\mathbb{R}^d$—see [AT09, Section 11.1]. It is the classical tool for computing the expected number of critical points of random fields, with many recent applications including spin glasses [AAČ13, FMM21] and landscapes of loss functions arising in machine learning [MAB20]. While the method applies to general densities, the conditional expectation appears infeasible to compute or estimate beyond the Gaussian case.

For the KDE $\widehat{P}_n$ defined in (1.1), define the random function $F_n : \mathbb{R} \to \mathbb{R}$ by

$$F_n(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^n (t - X_i) e^{-\frac{\beta}{2}(t-X_i)^2} = -\sqrt{\frac{2\pi n}{\beta^3}}\, \widehat{P}_n'(t). \tag{2.2}$$

Then $t \in \mathbb{R}$ is an upcrossing of F_n at level 0 if and only if $F_n(t) = 0$ and $F_n'(t) > 0$. This is equivalent to $\widehat{P}_n'(t) = 0$ and $\widehat{P}_n''(t) < 0$, i.e. t is a mode of $\widehat{P}_n$. Thus, the number of modes of $\widehat{P}_n$ in T is given by $U_0(F_n, T)$. For T, T' defined in (1.2)–(1.3), Propositions 1.6 and 1.7 yield

$$\begin{aligned} \mathbb{E}U_0(F_n, T) &\asymp \sqrt{\beta \log \beta}, \\ \mathbb{E}U_0(F_n, T') &\lesssim e^{-\frac{\omega(\beta)}{4}} \sqrt{\beta \log \beta}, \\ \mathbb{E}U_0(F_n, \mathbb{R} \setminus T) &\lesssim e^{\frac{\omega(\beta)}{2}} \sqrt{\beta}. \end{aligned} \tag{2.3}$$

2.2 Computing the Gaussian approximation

Without loss of generality, fix $t \in T$ with $t \geqslant 0$. We can rewrite $F_n(t)$ from (2.2) and compute its derivative: for independent copies (G_i, G_i') of

$$\begin{bmatrix} G(t) \\ G'(t) \end{bmatrix} = e^{-\frac{\beta}{2}(t-X)^2} \begin{bmatrix} t - X \\ 1 - \beta(t-X)^2 \end{bmatrix}, \tag{2.4}$$

where $X \sim N(0, 1)$, we have

$$\begin{bmatrix} F_n(t) \\ F_n'(t) \end{bmatrix} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \begin{bmatrix} G_i(t) \\ G_i'(t) \end{bmatrix} \sim p_t.$$

We prove that p_t is a well-defined density in Proposition 4.1, and defer the following computation to Appendix A.1.

[1] defined, for $f : \mathbb{R} \to \mathbb{R}$, as $\omega(\eta) = \sup_{t,s:\ |t-s| \leq \eta} |f(t) - f(s)|$.

Fact 2.2. *The mean and covariance matrix of the random vector* $(F_n(t), F_n'(t))$ *are given respectively by*

$$\begin{aligned}\mu_t &:= \sqrt{n}\begin{bmatrix}\mathbb{E}G(t)\\ \mathbb{E}G'(t)\end{bmatrix} \asymp n^{\frac{1}{2}}\beta^{-\frac{3}{2}}e^{-\frac{t^2}{2}}\begin{bmatrix}t\\ -t^2\end{bmatrix}\\ \Sigma_t &:= \begin{bmatrix}\operatorname{Var}G(t) & \operatorname{Cov}(G(t),G'(t))\\ \operatorname{Cov}(G(t),G'(t)) & \operatorname{Var}G'(t)\end{bmatrix} \asymp \beta^{-\frac{3}{2}}e^{-\frac{t^2}{2}}\begin{bmatrix}1 & -t\\ -t & \beta\end{bmatrix}.\end{aligned} \tag{2.5}$$

We proceed to centering and rescaling the density p_t. Let $Y_i(t)$, $i \in [n]$, be independent copies of

$$Y(t) = \Sigma_t^{-\frac{1}{2}}\begin{bmatrix}G(t)-\mathbb{E}G(t)\\ G'(t)-\mathbb{E}G'(t)\end{bmatrix} \tag{2.6}$$

Let q_t denote the density of $\sqrt{n}\sum_{i=1}^n Y_i(t)$. By construction q_t has mean 0 and covariance I_2. Moreover, by the change-of-variables formula, it holds

$$p_t(x,y) = (\det\Sigma_t)^{-\frac{1}{2}}q_t\Big(\Sigma_t^{-\frac{1}{2}}[(x,y)-\mu_t]\Big). \tag{2.7}$$

Now, let $\varphi : \mathbb{R}^2 \to \mathbb{R}$ be the density of $N(0, I_2)$, i.e.,

$$\varphi(x) := \frac{1}{\sqrt{2\pi}}e^{-\frac{\|x\|^2}{2}}.$$

We aim to approximate the Kac-Rice integral (2.1) as follows:

$$\int_T\int_0^\infty yp_t(0,y)\,\mathrm{d}y \approx \int_T\int_0^\infty y(\det\Sigma_t)^{-\frac{1}{2}}\varphi\Big(\Sigma_t^{-\frac{1}{2}}[(0,y)-\mu_t]\Big)\,\mathrm{d}y\,\mathrm{d}t. \tag{2.8}$$

The validity of this approximation is deferred to Section 3. In the remainder of this section, we solely focus on computing the right hand side integral.

Lemma 2.3. *There exists some* $A_t \asymp \beta^{-\frac{3}{2}}nt^2e^{-\frac{t^2}{2}}$ *such that*

$$\begin{aligned}\varphi\Big(\Sigma_t^{-\frac{1}{2}}[(0,y)-\mu_t]\Big) &\asymp \exp\Big(-A_t-\Theta\Big(\beta^{\frac{1}{2}}e^{\frac{t^2}{2}}\Big)y^2\Big),\\ \int_0^\infty y\varphi\Big(\Sigma_t^{-\frac{1}{2}}[(0,y)-\mu_t]\Big)\,\mathrm{d}y &\asymp \beta^{-\frac{1}{2}}e^{-\frac{t^2}{2}}e^{-A_t}.\end{aligned} \tag{2.9}$$

Proof of Lemma 2.3. Since $\det\Sigma_t \asymp \beta^{-2}e^{-t^2}$, we have

$$\Omega := \Sigma_t^{-1} \asymp \beta^{\frac{1}{2}}e^{\frac{t^2}{2}}\begin{bmatrix}\beta & t\\ t & 1\end{bmatrix}.$$

Now as $t \in T$ and so $t^2 \ll \beta$, we find

$$\Big\|\Sigma_t^{-\frac{1}{2}}[(0,y)-\mu_t]\Big\|^2 = \Big\langle(-\mu_{t,1}, y-\mu_{t,2}), \Sigma_t^{-1}(-\mu_{t,1}, y-\mu_{t,2})\Big\rangle$$

$$
\begin{aligned}
&\asymp \Omega_{11}\mu_{t,1}^2 - 2\Omega_{12}\mu_{t,1}(y-\mu_{t,2}) + \Omega_{22}(y-\mu_{t,2})^2 \\
&\asymp \beta^{\frac{1}{2}} e^{\frac{t^2}{2}} \mu_{t,1}^2 \left[\beta - 2t\left(\frac{y}{\mu_{t,1}} + t\right) + \left(\frac{y}{\mu_{t,1}} + t\right)^2\right] \\
&\asymp \beta^{\frac{1}{2}} e^{\frac{t^2}{2}} \mu_{t,1}^2 (\beta - t^2) + \beta^{\frac{1}{2}} e^{\frac{t^2}{2}} y^2 \\
&\asymp \beta^{-\frac{3}{2}} n t^2 e^{-\frac{t^2}{2}} + \beta^{\frac{1}{2}} e^{\frac{t^2}{2}} y^2.
\end{aligned}
$$

This shows the first statement in (2.9). For the second, we have

$$
\begin{aligned}
\int_0^\infty y\varphi\Big(\Sigma_t^{-\frac{1}{2}}[(0,y)-\mu_t]\Big)\,\mathrm{d}y &\asymp e^{-A_t}\int_0^\infty y e^{-\Theta\left(\beta^{1/2}e^{t^2/2}\right)y^2}\,\mathrm{d}y \\
&\asymp \beta^{-\frac{1}{2}} e^{-\frac{t^2}{2}} e^{-A_t}
\end{aligned}
$$

by Gaussian integral computations (see Fact A.1). □

2.3 The Kac-Rice integral over φ

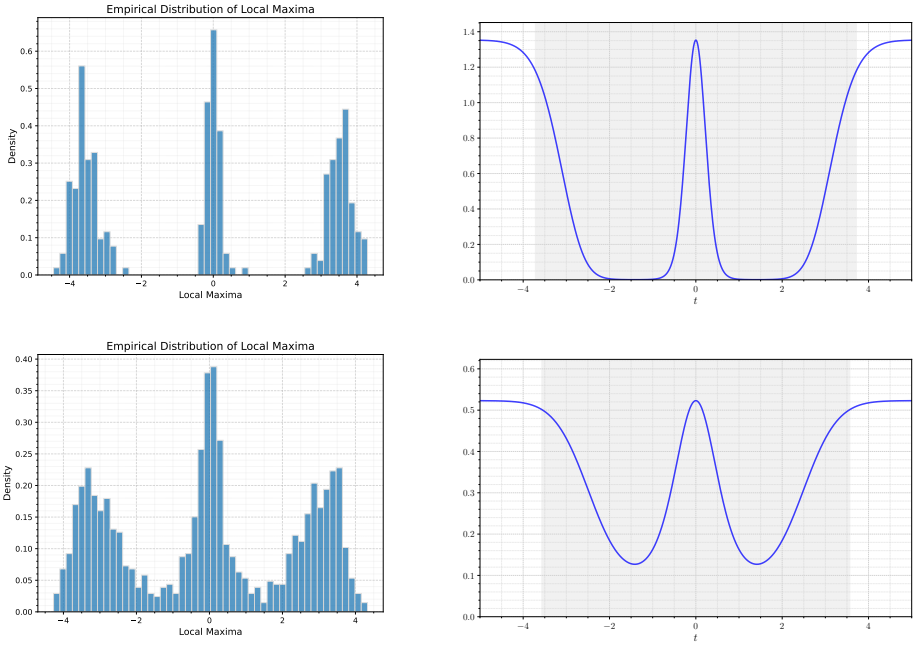


Figure 4: $n = 10^5$ is fixed throughout. (Left) Empirical distribution of the modes of $\widehat{P}_n$ over T for $\beta = 100$ (top) and $\beta = 300$ (bottom). (Right) The function $t \mapsto \sqrt{\beta}\exp(-A_t)$ for $\beta = 100$ (top) and $\beta = 300$ (bottom), which, due to the Kac-Rice formula, is an approximation for the distribution of the number of modes of $\widehat{P}_n$ in T. Shaded in grey is the interval T. (Code available at github.com/KimiSun18/2024-gauss-kde-attention.)

We compute (2.1) under the approximation (2.8). By (2.9) and (2.5), we have that

$$\int_S \int_0^\infty y(\det \Sigma_t)^{-\frac{1}{2}} \varphi\Big(\Sigma_t^{-\frac{1}{2}}[(0,y)-\mu_t]\Big)\,\mathrm{d}y\,\mathrm{d}t \asymp \sqrt{\beta}\int_S e^{-A_t}\,\mathrm{d}t \tag{2.10}$$

for any measurable $S \subset \mathbb{R}$. Assuming validity of the Gaussian approximation (see Section 3), it follows from the Kac-Rice formula that the density of modes at $t \in \mathbb{R}$ is proportional to $\sqrt{\beta}e^{-A_t}$. We plot this density in Figure 4 with the same choice of n and β as in the empirical distribution. We see that they match on the highlighted interval T, but not outside of T where the Gaussian approximation—see Remark 3.3.

We compute (2.10) explicitly for $S = T$ and $S = T'$.

Lemma 2.4. *If $n^c \lesssim \beta \lesssim n^{2-c}$ for some $c > 0$, then*

$$\int_T \int_0^\infty y(\det \Sigma_t)^{-\frac{1}{2}} \varphi\Big(\Sigma_t^{-\frac{1}{2}}[(0,y)-\mu_t]\Big)\,\mathrm{d}y\,\mathrm{d}t \asymp \sqrt{\beta\log\beta},$$
$$\int_{T'} \int_0^\infty y(\det \Sigma_t)^{-\frac{1}{2}} \varphi\Big(\Sigma_t^{-\frac{1}{2}}[(0,y)-\mu_t]\Big)\,\mathrm{d}y\,\mathrm{d}t \lesssim \sqrt{\beta}.$$

Proof of Lemma 2.4. Recall A_t from Lemma 2.3. By (2.10), it suffices to show that

$$\int_T e^{-A_t}\,\mathrm{d}t \asymp \sqrt{\log\beta} \quad \text{and} \quad \int_{T'} e^{-A_t}\,\mathrm{d}t \lesssim 1. \tag{2.11}$$

As $A_t > 0$, the integral is at most the length of T, which is $O(\sqrt{\log\beta})$ by (1.2). Let $t_s := \sqrt{2\log n - s\log\beta}$. As the integrand is positive, for constants $C, C' > 0$

$$\begin{aligned}
\int_T e^{-A_t}\,\mathrm{d}t &\geqslant \int_{t_2}^{t_{5/2}} \exp\Big(-C\beta^{-\frac{3}{2}}nt^2e^{-\frac{t^2}{2}}\Big)\,\mathrm{d}t \\
&\geqslant \big(t_{5/2}-t_2\big)\exp\Big(-C\beta^{-\frac{3}{2}}nt_{5/2}^2e^{-\frac{t_2^2}{2}}\Big) \\
&\gtrsim \sqrt{\log\beta}\exp\Big(-C'\beta^{-\frac{1}{2}}\log n\Big) \\
&\gtrsim \sqrt{\log\beta}
\end{aligned}$$

as $n, \beta \to \infty$ with $\log n \asymp \log\beta$. Now if $t \in T' = [-t_3, t_3]$, we have

$$e^{-\frac{t^2}{2}} \geqslant e^{-\frac{t_3^2}{2}} = \beta^{\frac{3}{2}}n^{-1}.$$

Hence

$$\int_{T'} e^{-A_t}\,\mathrm{d}t \lesssim \int_0^{t_3} \exp\Big(-C\beta^{-\frac{3}{2}}nt^2e^{-\frac{t^2}{2}}\Big)\,\mathrm{d}t \leqslant \int_0^{t_3} e^{-Ct^2}\,\mathrm{d}t \lesssim 1. \qquad \square$$

3 Leveraging the Edgeworth expansion

In this section, we show the approximation of p_t by φ is valid in T by showing

$$\int_T \int_0^\infty (\det \Sigma_t)^{-\frac{1}{2}} y |q_t - \varphi| \left(\Sigma_t^{-\frac{1}{2}} [(0, y) - \mu_t] \right) \mathrm{d}y \, \mathrm{d}t \ll \sqrt{\beta \log \beta}. \tag{3.1}$$

One natural idea is to use some asymptotic series to expand p_t around φ, e.g. the Edgeworth expansion, to argue that $|p_t - \varphi| \ll \varphi$ in the sense of the integral over y. We discuss the two major obstacles we have to overcome in order to implement this approach.

- Firstly, all known results on validity of asymptotic series such as Edgeworth expansions treat densities q_t and φ that are independent of n. As β grows in n, we will need to re-derive these results and carefully track the dependence on β. This will give extra constraints on (t, β, n) for the validity of such an asymptotic series. Fortunately, this will be satisfied precisely when $t \in T$.
- Secondly, even without the n-dependence of β, expanding q_t to the order φ plus error in for example [BR10, Theorem 19.2] gives error roughly

  $$|p_t - \varphi|(\mathbf{x}) \lesssim \frac{1}{1 + \|\mathbf{x}\|^2} \qquad \text{for all } \mathbf{x} \in \mathbb{R}^2,$$

  but then the integral over y in (3.1) fails to converge. Therefore, we will need to go to the next term ψ in the Edgeworth series. We control its integral over y and t in (3.1) manually in Section 3.1. Then, we control the higher order terms in Section 3.2 using the approach motivated above, so our error now converges upon integrating over y, i.e. roughly

  $$\left| p_t - \varphi - n^{-\frac{1}{2}} \psi \right| (\mathbf{x}) \lesssim \frac{1}{1 + \|\mathbf{x}\|^3} \qquad \text{for all } \mathbf{x} \in \mathbb{R}^2.$$

3.1 Bounding the third order error

Recall Y from (2.6). For a multi-index $\alpha \in \mathbb{Z}^2_{\geqslant 0}$, let κ_t^α be the cumulant of Y indexed by α, which depends on t. Let H_t^α denote the standard Hermite polynomials of with index α. The next term in the Edgeworth series is $n^{-1/2}\psi$ with

$$\psi(\mathbf{x}) := \frac{\varphi(\mathbf{x})}{6} \sum_{k=0}^{3} \kappa_t^{(k,3-k)} H^{(k,3-k)}(\mathbf{x}) \quad \text{where} \quad H^\alpha := (-1)^{|\alpha|} \frac{\partial^\alpha \varphi}{\varphi} \tag{3.2}$$

If $|\alpha| = 3$, κ_t^α are the third moments of Y, which we bound in Appendix A.1.

Fact 3.1. *Let η_3 be the largest third cumulant of Y. Then*

$$\eta_3 := \max \left\{ \kappa_t^{(k,3-k)} : k \in \{0, 1, 2, 3\} \right\} \lesssim \beta^{\frac{1}{4}} e^{\frac{t^2}{4}}.$$

We bound $H^{(k,3-k)}\left(\Sigma_t^{-1/2}[(0,y)-\mu_t]\right)$ by a polynomial in $\tilde{y} = \Theta(\beta^{1/2}e^{t^2/4}y)$. Now, by a similar method as Lemma 2.3, we obtain the following bound. It says that when we integrate the Edgeworth series $p_t = \varphi + n^{-1/2}\psi + \dots$ over y and t, the contribution φ dominates $n^{-1/2}\psi$, hinting at the validity of the approximation.

Lemma 3.2. *Recall T from (1.2) and A_t from Lemma 2.3. Then*

$$\int_T \int_0^\infty y(n\det\Sigma_t)^{-\frac{1}{2}}\psi\left(\Sigma_t^{-\frac{1}{2}}[(0,y)-\mu_t]\right)\mathrm{d}y\,\mathrm{d}t \lesssim e^{-\frac{\omega(\beta)}{4}}\sqrt{\beta\log\beta}.$$

Proof of Lemma 3.2. By the proof of Lemma 2.3, (3.2), and Fact 3.1

$$\begin{aligned}
&\int_T \int_0^\infty y(n\det\Sigma_t)^{-\frac{1}{2}}\psi\left(\Sigma_t^{-\frac{1}{2}}[(0,y)-\mu_t]\right)\mathrm{d}y\,\mathrm{d}t\\
&= \frac{1}{6}\sum_{k=0}^{3}\int_T (n\det\Sigma_t)^{-\frac{1}{2}}\kappa_t^{(k,3-k)}\int_0^\infty y\left[\varphi H^{(k,3-k)}\right]\left(\Sigma_t^{-\frac{1}{2}}[(0,y)-\mu_t]\right)\mathrm{d}y\,\mathrm{d}t\\
&\asymp \int_T (n\det\Sigma_t)^{-\frac{1}{2}}\eta_3 e^{-A_t}\int_0^\infty y\left(\beta^{\frac{1}{4}}e^{\frac{t^2}{4}}y\right)^{O(1)} e^{-\Theta\left(\beta^{\frac{1}{2}}e^{\frac{t^2}{2}}\right)y^2}\mathrm{d}y\,\mathrm{d}t\\
&\asymp \int_T (n\det\Sigma_t)^{-\frac{1}{2}}e^{-A_t}\eta_3\beta^{-\frac{1}{2}}e^{-\frac{t^2}{2}}\left(\int_0^\infty \tilde{y}^{O(1)}e^{-\frac{\tilde{y}^2}{2}}d\tilde{y}\right)\mathrm{d}t\\
&\lesssim n^{-\frac{1}{2}}\beta^{\frac{3}{4}}\sup_{t\in T}\left(e^{\frac{t^2}{4}}\right)\int_T e^{-A_t}\,\mathrm{d}t\\
&\lesssim e^{-\frac{\omega(\beta)}{4}}\sqrt{\beta\log\beta}
\end{aligned}$$

where the last step follows (2.11) and the definition (1.2) of T. □

Remark 3.3. *One can see at this is actually an asymptotic equality by checking the Gaussian integrals in the proof above are of their typical order (i.e. no cancellation of leading terms). Hence, the decay is only a factor of $e^{-\omega(\beta)/4}$. For $t \notin T$, even $t = \sqrt{2\log n - 0.99\log\beta}$, the last inequality in Lemma 3.2 fails and we get a bound of polynomially larger that $\sqrt{\beta}$. Then, as the third order error is asymptotically larger than the contribution of the Gaussian approximation, so the normal approximation is no longer valid. This can be seen by comparing the plots in Figure 4.*

3.2 Bounding higher order errors

We follow the classical proof about the validity of the Edgeworth expansion as an asymptotic series to show bound the higher order pointwise error of density function as follows.

Lemma 3.4. *Assume that p_t is bounded almost everywhere for any fixed $t \in T$. If $n^c \lesssim \beta \lesssim n^{2-c}$ for some $c > 0$, then for any $t \in T$ and $\mathbf{x} \in \mathbb{R}^2$, we have that*

$$\left(1+\|\mathbf{x}\|^3\right)\left|q_t - \varphi - n^{-\frac{1}{2}}\psi\right|(\mathbf{x}) \lesssim e^{-\frac{\omega(\beta)}{4}}. \tag{3.3}$$

We defer the discussion of the assumption to Assumption 4.2 and defer the proof to Appendix A.3. Here, we comment on the differences with [BR10, Theorem 19.2] for $s = 3$, which we follow in spirit but modify to allow n dependence in q_t in the form of β. There, it is shown that the left hand side is $o(n^{-1/2})$ provided third moments $\eta_3 = O(1)$. With the bound Fact 3.1 on η_3 in our case which is not constant, the error from the asymptotic series decays not in powers of $n^{-1/2}$ but powers of $O(n^{-1/2}e^{t^2/4}\beta^{1/4}) = O(\exp(-\omega(\beta)/2))$.

Corollary 3.5. *If $n^c \lesssim \beta \lesssim n^{2-c}$ for $c > 0$, then asymptotically in $n, \beta \to \infty$*

$$\int_T \int_0^\infty (\det \Sigma_t)^{-\frac{1}{2}} y \left| q_t - \varphi - n^{-\frac{1}{2}}\psi \right| \left(\Sigma_t^{-\frac{1}{2}}[(0,y) - \mu_t] \right) \mathrm{d}y \, \mathrm{d}t \lesssim e^{-\frac{\omega(\beta)}{4}} \sqrt{\beta \log \beta}.$$

Proof of Corollary 3.5. By Lemma 3.4 and computations in the proof of Lemma 2.3

$$\begin{aligned}
&\int_T \int_0^\infty (\det \Sigma_t)^{-\frac{1}{2}} y \left| q_t - \varphi - n^{-\frac{1}{2}}\psi \right| \left(\Sigma_t^{-\frac{1}{2}}[(0,y) - \mu_t] \right) \mathrm{d}y \, \mathrm{d}t \\
&\lesssim e^{-\frac{\omega(\beta)}{4}} \int_T \int_0^\infty (\det \Sigma_t)^{-\frac{1}{2}} y \left(1 + \left\| \Sigma_t^{-\frac{1}{2}}[(0,y) - \mu_t] \right\|^3 \right)^{-1} \mathrm{d}y \, \mathrm{d}t \\
&\lesssim e^{-\frac{\omega(\beta)}{4}} \int_T \int_0^\infty (\det \Sigma_t)^{-\frac{1}{2}} y \left(1 + \Theta\left(\beta^{-\frac{3}{2}} n t^2 e^{-\frac{t^2}{2}} + \beta^{\frac{1}{2}} e^{\frac{t^2}{2}} y^2 \right)^{\frac{3}{2}} \right)^{-1} \mathrm{d}y \, \mathrm{d}t \\
&\lesssim e^{-\frac{\omega(\beta)}{4}} \int_T (\det \Sigma_t)^{-\frac{1}{2}} \left(\beta^{\frac{1}{4}} e^{\frac{t^2}{4}} \right)^{-2} \int_0^\infty \frac{\tilde{y}}{1 + \tilde{y}^3} \mathrm{d}\tilde{y} \\
&\asymp e^{-\frac{\omega(\beta)}{4}} \sqrt{\beta \log \beta}
\end{aligned}$$

for $\tilde{y} \asymp \beta^{1/4} e^{t^2/4} y$, where we note that T has length $\Theta(\sqrt{\log \beta})$ by (1.2) and

$$\int_0^\infty \frac{\tilde{y}}{1 + \tilde{y}^3} \mathrm{d}\tilde{y} = \frac{2\pi}{3\sqrt{3}} = O(1). \qquad \square$$

4 Proof of Theorem 1.2

We prove Propositions 1.6 and 1.7 by checking (2.3), thereby proving Theorem 1.2.

4.1 Proof of Proposition 1.6

To prove Proposition 1.6 we seek to apply Theorem 2.1 to $\mathbb{E}U_0(F_n, T)$. This in turn requires checking all the assumptions of Theorem 2.1. We have

Proposition 4.1. *Let β, n be as in Theorem 1.2, and fix $t \in T$. Let μ_t denote the law of $(F_n(t), F_n'(t))$ defined in (2.2). Then μ_t admits a density $p_t \in \mathscr{C}^0(\mathbb{R}^2)$ satisfying $p_t(\mathbf{x}) \to 0$ as $\|\mathbf{x}\| \to \infty$. Moreover, conditions 1, 2 in Theorem 2.1 also hold for $\Psi(t) = F_n(t)$.*

We defer the proof to Appendix A.4. We work under assumption

Assumption 4.2. *Consider the setting of Proposition 4.1. We assume that condition 4 of Theorem 2.1 holds for* $\Psi(t) = F_n(t)$.

To check conditions on moduli of continuity such as *4* in Theorem 2.1 in the Gaussian setting, one usually resorts to using results such as the Borell-TIS inequality [AT09, Theorem 2.1.1]. Checking the validity of this assumption in the present, non-Gaussian, setting does not appear straightforward.

With Proposition 4.1 and Assumption 4.2, we deduce

Lemma 4.3. *With the notation as in Theorem 2.1,*

$$\mathbb{E}U_0(F_n, T) = \int_T \int_0^\infty y p_t(0, y)\, \mathrm{d}y\, \mathrm{d}t. \tag{4.1}$$

Proof of Proposition 1.6. Combining Lemmas 2.4 and 3.2 and Corollary 3.5 gives

$$\begin{aligned}
\mathbb{E}U_0(F_n, T) &= \int_T \int_0^\infty y p_t(0, y)\, \mathrm{d}y\, \mathrm{d}t \\
&= \int_T \int_0^\infty (\det \Sigma_t)^{-\frac{1}{2}} y q_t\Big(\Sigma_t^{-\frac{1}{2}}[(0, y) - \mu_t]\Big)\, \mathrm{d}y\, \mathrm{d}t \\
&= \int_T \int_0^\infty (\det \Sigma_t)^{-\frac{1}{2}} y \varphi\Big(\Sigma_t^{-\frac{1}{2}}[(0, y) - \mu_t]\Big)\, \mathrm{d}y\, \mathrm{d}t \\
&\quad + \int_T \int_0^\infty (n \det \Sigma_t)^{-\frac{1}{2}} y \psi\Big(\Sigma_t^{-\frac{1}{2}}[(0, y) - \mu_t]\Big)\, \mathrm{d}y\, \mathrm{d}t \\
&\quad + \int_T \int_0^\infty (\det \Sigma_t)^{-\frac{1}{2}} y \Big[q_t - \varphi - n^{-\frac{1}{2}}\psi\Big]\Big(\Sigma_t^{-\frac{1}{2}}[(0, y) - \mu_t]\Big)\, \mathrm{d}y\, \mathrm{d}t \\
&\asymp \sqrt{\beta \log \beta}
\end{aligned}$$

for $t \in T$, as the first summand is $\Theta(\sqrt{\beta \log \beta})$ while the last two are $o(\sqrt{\beta \log \beta})$. Replacing T by $T' = [-\sqrt{2 \log n - 3 \log \beta}, \sqrt{2 \log n - 3 \log \beta}]$, the first summand is $o(\sqrt{\beta \log \beta})$ by Lemma 2.4. By positivity of the integrand, we bound the last two integrals over T' by those over T, which are themselves $o(\sqrt{\beta \log \beta})$. Now, all three summands are $o(\sqrt{\beta \log \beta})$, as desired in Proposition 1.6. □

4.2 Proof of Proposition 1.7

In this section, we prove Proposition 1.7.

Lemma 4.4. *For any* $a > 0$ *and* $X_1, \dots, X_n \in \mathbb{R}$, *the number of modes of* $\widehat{P}_n$ *in* (a, ∞) *is at most* $|I|$ *where* $I = \{i \in [n] : X_i \geqslant a\}$.

Proof of Lemma 4.4. Note that $\widehat{P}_n(t) = \sum_{i=1}^n g_i(t)$ where for $i \in [n]$ we define

$$g_i(t) := \sqrt{\frac{\beta}{2\pi n^2}} \mathsf{K}_{\beta^{-1/2}}(t - X_i). \tag{4.2}$$

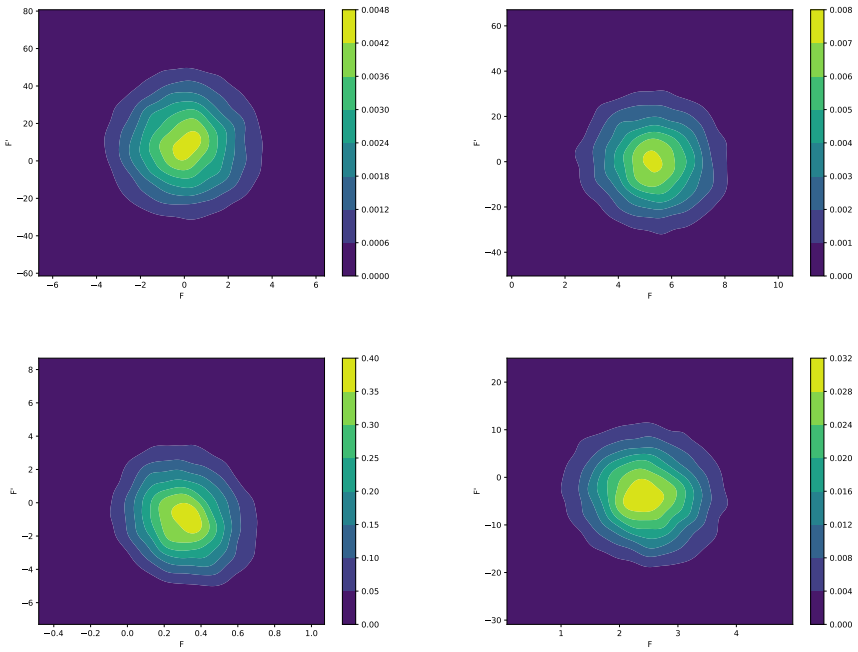


Figure 5: An estimate of the density $p_t = p_t(x, y)$ of $(F_n(t), F_n'(t))$ for $t = 0, 1, 2, 3$ (clockwise from top left), where $\beta = 81$ and $n = 6500$, so that $\sqrt{2 \log n - \log \beta} \approx 3$. (Code available at github.com/KimiSun18/2024-gauss-kde-attention.)

For $i \notin I$, g_i is monotonically decreasing on $[X_i, \infty) \supset (a, \infty)$, so $\sum_{i \notin I} g_i(t)$ has no modes in (a, ∞). To this Gaussian mixture, we add in $g_i(t)$ for $i \in I$ one-by-one. By [CPW03, Theorem 2], each time the number of modes in (a, ∞) increases by at most one. In $|I|$-many steps, there are at most $|I|$ such modes. □

Remark 4.5. *This argument crucially relies on the KDE being Gaussian. As discussed in [CPW03], the Gaussian kernel is the only kernel where for any fixed samples the number of modes of the KDE is non-increasing in the bandwidth h. For other kernels, we do suspect the analog of Lemma 4.4 to hold, but a different argument is needed. In particular, [MMF92, Mam95, KM97] avoids this problem by counting modes on compact sets.*

Proof of Proposition 1.7. By Lemma 4.4, symmetry of T in (1.2) around $t = 0$, lin-

earity of expectations, and the tail bound $\mathbb{P}(|X| \geqslant a) \leqslant 2e^{-a^2/2}$ for $X \sim N(0,1)$

$$\begin{aligned}\mathbb{E}U_0(F_n, \mathbb{R} \setminus T) &\leqslant \mathbb{E}|\{i : X_i \notin T\}| \\ &= n\mathbb{P}(X \notin T) \\ &\leqslant 2n \exp\left(-\frac{2\log n - \log\beta - \omega(\beta)}{2}\right) \\ &= 2\sqrt{\beta \exp(\omega(\beta))} \\ &\ll \sqrt{\beta \log \beta}\end{aligned} \tag{4.3}$$

by the definition of $\omega(\beta)$, proving Proposition 1.7. □

This concludes the proof of Theorem 1.2.

5 Concluding remarks

We showed that the expected number of modes of a Gaussian KDE with bandwidth $\beta^{-\frac{1}{2}}$ of $n \geqslant 1$ samples drawn iid from $N(0,1)$ is of order $\Theta(\sqrt{\beta \log \beta})$ for $n^c \lesssim \beta \lesssim n^{2-c}$, where $c > 0$ is arbitrarily small. We also provide a precise picture of where the modes are located.

The question in the higher-dimensional case, as well as on the unit sphere $\mathbb{S}^{d-1}$ with uniformly distributed samples, remains open.

Acknowledgments

The authors would like to thank Enno Mammen for useful discussion and sharing important references. We also thank Dan Mikulincer for discussions on Gaussian approximation using Edgeworth expansions, Valeria Banica for comments on the method of stationary phase, and Alexander Zimin for providing Figure 2.

Funding. P.R. was supported by NSF grants DMS-2022448, CCF-2106377, and a gift from Apple. Y.S. was supported by the MIT UROP and MISTI France Programs.

A Additional proofs

A.1 Proof of Fact 2.2

In this section we compute the first two moments of (G, G') to prove Fact 2.2. Note that if $n^c \lesssim \beta \lesssim n^{2-c}$ for some $c > 0$, and for $t \in T$, then we have $\exp\Theta(t^2/\beta) \to 1$. This implies that exponentials in the moments are asymptotically $e^{-t^2/2}$. We frequently use the following fact about Gaussian integrals both in exact and asymptotic forms.

Fact A.1. *Let Γ denote the Gamma function. For any $\alpha > 0$ and integer $n \geqslant 0$,*

$$\int_0^\infty z^n e^{-\alpha u^2}\,\mathrm{d}u = \frac{1}{2}\Gamma\left(\frac{n+1}{2}\right)\alpha^{-\frac{n+1}{2}}.$$

We first compute μ_t. Completing the square gives

$$\frac{\beta}{2}z^2 + \frac{1}{2}(z-t)^2 = \frac{\beta+1}{2}u^2 + \frac{\beta t^2}{2(\beta+1)} \quad \text{where} \quad u = z - \frac{t}{\beta+1}.$$

Hence, using Fact A.1 we compute

$$\begin{aligned}
\mathbb{E}G(t) = \int_{-\infty}^{\infty} z e^{-\frac{\beta}{2}z^2} \,\mathrm{d}\gamma_{t,1}(z) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z e^{-\frac{\beta}{2}z^2 - \frac{1}{2}(z-t)^2} \,\mathrm{d}z \\
&= \frac{e^{-\frac{\beta}{2(\beta+1)}t^2}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \Big(u + \frac{t}{\beta+1}\Big) e^{-\frac{\beta+1}{2}u^2} \,\mathrm{d}u \\
&= \frac{e^{-\frac{\beta}{2(\beta+1)}t^2}}{\sqrt{2\pi}} \Big(\frac{t}{\beta+1}\Big) \frac{\sqrt{\pi}}{\big(\frac{\beta+1}{2}\big)^{\frac{1}{2}}} \\
&= \frac{e^{-\frac{\beta}{2(\beta+1)}t^2} t}{(\beta+1)^{\frac{3}{2}}},
\end{aligned}$$

as well as

$$\begin{aligned}
\mathbb{E}G'(t) &= \int_{-\infty}^{\infty} (1-\beta z^2) z^{-\frac{\beta}{2}z^2} \,\mathrm{d}\gamma_{t,1}(z) \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \Big(1-\beta z^2\Big) e^{-\frac{\beta}{2}z^2 - \frac{1}{2}(z-t)^2} \,\mathrm{d}z \\
&= \frac{e^{-\frac{\beta}{2(\beta+1)}t^2}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \Big[1 - \beta\Big(u + \frac{t}{\beta+1}\Big)^2\Big] e^{-\frac{\beta+1}{2}u^2} \,\mathrm{d}u \\
&= \frac{e^{-\frac{\beta}{2(\beta+1)}t^2}}{\sqrt{2\pi}} \left[\left(1 - \frac{\beta t^2}{(\beta+1)^2}\right) \frac{\sqrt{\pi}}{\big(\frac{\beta+1}{2}\big)^{\frac{1}{2}}} - \frac{\beta\sqrt{\pi}}{2\big(\frac{\beta+1}{2}\big)^{\frac{3}{2}}}\right] \\
&= \frac{e^{-\frac{\beta}{2(\beta+1)}t^2}}{(\beta+1)^{\frac{5}{2}}} \Big((\beta+1)^2 - \beta t^2 - \beta(1+\beta)\Big) \\
&= \frac{e^{-\frac{\beta}{2(\beta+1)}t^2}}{(\beta+1)^{\frac{5}{2}}} \Big(1 + \beta - \beta t^2\Big).
\end{aligned}$$

From these computations, and the remark after Fact A.1, we readily obtain the asymptotics of μ_t as in Fact 2.2 upon multiplying by $\sqrt{n}$.

We now compute Σ_t. Completing the square gives

$$z^2 + \frac{1}{2}(z-t)^2 = \frac{2\beta+1}{2}u^2 + \frac{\beta t^2}{2\beta+1} \quad \text{where} \quad u = z - \frac{t}{2\beta+1}.$$

Hence using Fact A.1 we compute

$$\begin{aligned}
\mathbb{E}G^2(t) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2 e^{-\beta z^2 - \frac{1}{2}(z-t)^2} \, \mathrm{d}z \\
&= \frac{e^{-\frac{\beta}{(2\beta+1)}t^2}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \left(u + \frac{t}{2\beta+1}\right)^2 e^{-\frac{1+2\beta}{2}u^2} \, \mathrm{d}u \\
&= \frac{e^{-\frac{\beta}{(2\beta+1)}t^2}}{\sqrt{2\pi}} \left[\left(\frac{t}{2\beta+1}\right)^2 \frac{\sqrt{\pi}}{\left(\frac{2\beta+1}{2}\right)^{\frac{1}{2}}} + \frac{\sqrt{\pi}}{2\left(\frac{2\beta+1}{2}\right)^{\frac{3}{2}}} \right] \\
&= \frac{e^{-\frac{\beta}{(2\beta+1)}t^2}}{(2\beta+1)^{\frac{5}{2}}} \Big(t^2 + 2\beta + 1\Big),
\end{aligned}$$

as well as

$$\begin{aligned}
&\mathbb{E}\big[G(t)G'(t)\big] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z(1-\beta z^2) e^{-\beta z^2 - \frac{1}{2}(z-t)^2} \, \mathrm{d}z \\
&= \frac{e^{-\frac{\beta}{(2\beta+1)}t^2}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \left(u + \frac{t}{2\beta+1} - \beta\left(u + \frac{t}{2\beta+1}\right)^3\right) e^{-\frac{1+2\beta}{2}u^2} \, \mathrm{d}u \\
&= \frac{e^{-\frac{\beta}{2\beta+1}t^2}}{\sqrt{2\pi}} \left[\left(\frac{t}{2\beta+1} - \beta\left(\frac{t}{2\beta+1}\right)^3\right) \frac{\sqrt{\pi}}{\left(\frac{2\beta+1}{2}\right)^{\frac{1}{2}}} - \left(\frac{3t\beta}{2\beta+1}\right) \frac{\sqrt{\pi}}{2\left(\frac{2\beta+1}{2}\right)^{\frac{3}{2}}} \right] \\
&= \frac{e^{-\frac{\beta}{2\beta+1}t^2}}{(2\beta+1)^{\frac{7}{2}}} \Big[t(2\beta+1)^2 - \beta t^3 - 3t\beta(2\beta+1) \Big] \\
&= \frac{e^{-\frac{\beta}{2\beta+1}t^2}}{(2\beta+1)^{\frac{7}{2}}} \Big(-2\beta^2 t + \beta t - \beta t^3 + t \Big),
\end{aligned}$$

and, finally,

$$\begin{aligned}
\mathbb{E}G'^2(t) &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}(1-\beta z^2)^2 e^{-\beta z^2-\frac{1}{2}(z-t)^2}\,\mathrm{d}z \\
&= \frac{e^{-\frac{\beta}{2\beta+1}t^2}}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\left(1-\beta\left(u+\frac{t}{2\beta+1}\right)^2\right)^2 e^{-\frac{1+2\beta}{2}u^2}\,\mathrm{d}u \\
&= \frac{e^{-\frac{\beta}{2\beta+1}t^2}}{\sqrt{2\pi}}\left[\left(1-\frac{\beta t^2}{(2\beta+1)^2}\right)^2\frac{\sqrt{\pi}}{\left(\frac{2\beta+1}{2}\right)^{\frac{1}{2}}}\right. \\
&\qquad\left. + \left(\frac{6\beta^2t^2}{(2\beta+1)^2}-2\beta\right)\frac{\sqrt{\pi}}{2\left(\frac{2\beta+1}{2}\right)^{\frac{3}{2}}} + \beta^2\cdot\frac{3\sqrt{\pi}}{4\left(\frac{2\beta+1}{2}\right)^{\frac{5}{2}}}\right] \\
&= \frac{e^{-\frac{\beta}{2\beta+1}t^2}}{(2\beta+1)^{\frac{9}{2}}}\left[((2\beta+1)^2-\beta t^2)^2+(2\beta+1)6\beta^2t^2-2\beta(2\beta+1)^3+3\beta^2(2\beta+1)^2\right] \\
&= \frac{e^{-\frac{\beta}{2\beta+1}t^2}}{(2\beta+1)^{\frac{9}{2}}}\Big(12\beta^4+4\beta^3(t^2+5)+\beta^2(t^4-2t^2+15)-2\beta(t^2-3)+1\Big).
\end{aligned}$$

We check that entries of Σ_t are asymptotically the corresponding second moments. Indeed, surpressing the dependence on t, we have that

- $\mathbb{E}G^2 \asymp \beta^{-\frac{5}{2}}(t^2+\beta)e^{\frac{t^2}{2}} \gg \beta^{-3}t^2E_t^2 \asymp (\mathbb{E}G)^2$,
- $\mathbb{E}GG' \asymp -\beta^{-\frac{7}{2}}e^{\frac{t^2}{2}}(\beta^2t+\beta t^3) \gg \beta^{-4}E_t^2t(\beta-\beta t^2) \asymp (\mathbb{E}G)(\mathbb{E}G')$,
- $\mathbb{E}G'^2 \asymp \beta^{-\frac{9}{2}}e^{\frac{t^2}{2}}(\beta^4+\beta^2t^4) \gg \beta^{-5}E_t^2\beta^2(1+t^4) \asymp (\mathbb{E}G')^2$,

where we bound $e^{\frac{t^2}{2}} \leqslant 1$. From these computations, and the remark after Fact A.1, we readily obtain the asymptotics of Σ_t as indicated in Fact 2.2. □

A.2 Proof of Fact 3.1

In this section, we prove Fact 3.1 on third moments of $Y = \Sigma_t^{-1/2}(G-\mathbb{E}G, G'-\mathbb{E}G')$. To upper bound, we do not need to track the leading coefficients to ensure that they do not vanish when we combine applications of Fact A.1. Recalling Σ_t^{-1} from the proof of Lemma 2.3, we upper bound asymptotically via Hölder's inequality:

$$\begin{aligned}
\eta_3 &= \max_k\left|Y^{(k,3-k)}\right| \\
&\leqslant \mathbb{E}\|Y\|^3 \\
&\leqslant \mathbb{E}\left\|(G-\mathbb{E}G, G'-\mathbb{E}G')^\intercal\Sigma_t^{-1}(G-\mathbb{E}G, G'-\mathbb{E}G')\right\|^{\frac{3}{2}} \\
&\lesssim \beta^{\frac{3}{4}}e^{\frac{3t^2}{4}}\mathbb{E}\left|\beta(G-\mathbb{E}G)^2+2t(G-\mathbb{E}G)(G'-\mathbb{E}G')+(G'-\mathbb{E}G')^2\right|^{\frac{3}{2}}
\end{aligned}$$

$$
\begin{aligned}
&\lesssim \beta^{\frac{3}{4}} e^{\frac{3t^2}{4}} \left(\beta^{\frac{3}{2}} \mathbb{E}|G|^3 + \mathbb{E}|G'|^3\right) \\
&\lesssim \beta^{\frac{3}{4}} e^{\frac{3t^2}{4}} \int_{-\infty}^{\infty} \left(\beta^{\frac{3}{2}}|z|^3 + |1-\beta z^2|^3\right) e^{-\frac{3\beta}{2}z^2 - \frac{1}{2}(z-t)^2} \,\mathrm{d}z \\
&\lesssim \beta^{\frac{3}{4}} e^{\frac{t^2}{4}} \int_0^{\infty} h\left(u + \frac{t}{3\beta+1}\right) e^{-\frac{3\beta+1}{2}u^2} \,\mathrm{d}u,
\end{aligned}
$$

where we factor out $e^{-\frac{3\beta}{2(3\beta+1)}t^2} \asymp e^{-\frac{t^2}{2}}$ of the integral by substituting

$$
h(z) = \beta^{\frac{3}{2}} z^3 + (1+\beta z^2)^3 \quad \text{and} \quad u = z - \frac{t}{3\beta+1}.
$$

By linearity of integration and Fact A.1, each monomial u^n integrates to $O(\beta^{-(n+1)/2})$. By monotonicity of h on $\mathbb{R}_{\geqslant 0}$ and since $t \in T$, for some constant $C > 0$,

$$
\begin{aligned}
\eta_3 \lesssim \beta^{\frac{3}{4}} e^{\frac{t^2}{4}} \int_0^{\infty} h\left(u + \frac{t}{3\beta+1}\right) e^{-\frac{3\beta+1}{2}u^2} \,\mathrm{d}u &\lesssim \beta^{\frac{1}{4}} e^{\frac{t^2}{4}} h\left(C\beta^{-\frac{1}{2}} + \frac{t}{3\beta+1}\right) \\
&\lesssim \beta^{\frac{1}{4}} e^{\frac{t^2}{4}} h\left(2C\beta^{-\frac{1}{2}}\right) \\
&\lesssim \beta^{\frac{1}{4}} e^{\frac{t^2}{4}}
\end{aligned}
$$

upon noting $w \mapsto h(w/\sqrt{\beta})$ has constant coefficients. This proves Fact 3.1. □

A.3 Proof of Lemma 3.4

Fix n, β sufficiently large. For a multi-index α, define

$$
\begin{aligned}
h(\mathbf{x}) &= \mathbf{x}^{\alpha}\left(q_t - \varphi - n^{-\frac{1}{2}}\psi\right)(\mathbf{x}), \\
\mathscr{F}h(\mathbf{z}) &= \partial^{\alpha}\left(\mathscr{F}(q_t) - \frac{\varphi}{6\sqrt{n}} \sum_{k=0}^{3} \kappa_t^{(k,3-k)} H^{(k,3-k)}\right)(\mathbf{z}),
\end{aligned}
$$

where

$$
(\mathscr{F}h)(\mathbf{z}) = \int_{\mathbb{R}^2} e^{-\mathrm{i}\langle \mathbf{x}, \mathbf{z}\rangle} h(\mathbf{x}) \,\mathrm{d}\mathbf{x}
$$

denotes the Fourier transform of h. By Fourier inversion, it suffices to show that for any multi-index α with order $|\alpha| \leqslant 3$,

$$
|h(\mathbf{x})| = \left|\frac{1}{(2\pi)^2} \int_{\mathbb{R}^2} e^{-\mathrm{i}\langle \mathbf{z}, \mathbf{x}\rangle} \mathscr{F}h(\mathbf{z}) \,\mathrm{d}\mathbf{z}\right| \lesssim \int_{\mathbb{R}^2} |\mathscr{F}h(\mathbf{z})| \,\mathrm{d}\mathbf{z} \tag{A.1}
$$

is $O(\exp(-\omega(\beta)/2))$. We apply [BR10, Theorem 9.12]—which is not asymptotic and has explicit constants—so we may use it even though q_t depends on n. Upon verifying the conditions on Y via Fact 2.2 and Fact 3.1, we have that

$$
|\mathscr{F}h(\mathbf{z})| \lesssim n^{-\frac{1}{2}} \eta_3^{\frac{1}{2}} \|\mathbf{z}\|^{O(1)} e^{-\frac{\|\mathbf{z}\|^2}{4}}
$$

provided $\|\mathbf{z}\| \leqslant a\sqrt{n}$ for some $a \asymp \eta_3^{-1/2}$. By Fact 3.1, we have that

$$\int_{\|\mathbf{z}\|\leqslant a\sqrt{n}} |\mathscr{F}h(\mathbf{z})|\,\mathrm{d}\mathbf{z} \lesssim n^{-\frac{1}{2}}\eta_3^{\frac{1}{2}} \int_{\mathbb{R}^2} \|\mathbf{z}\|^{O(1)} e^{-\frac{\|\mathbf{z}\|^2}{4}}\,\mathrm{d}\mathbf{z} \lesssim e^{-\frac{\omega(\beta)}{2}}. \tag{A.2}$$

Recall that q_t is the density of $n^{-1/2}\sum_{i=1}^n Y_i$, and let f denote the density of Y. Now, we proceed as the proof of [BR10, Theorem 19.2]. As p_t is bounded, so is f, and so [BR10, Theorem 19.1] gives $\mathscr{F}f \in L^1(\mathbb{R}^2)$ and

$$\delta := \sup_{\|\mathbf{z}\|>a} |\mathscr{F}f(\mathbf{z})| < 1.$$

By properties of the Fourier transform and the product rule,

$$\begin{aligned}
\int_{\|\mathbf{z}\|>a\sqrt{n}} |\partial^\alpha \mathscr{F}q_t(\mathbf{z})|\,\mathrm{d}\mathbf{z} &\lesssim \eta_{|\alpha|} n^{\frac{|\alpha|}{2}} \delta^{n-|\alpha|-1} \int_{\mathbb{R}^2} \left|\mathscr{F}f\left(\frac{\mathbf{z}}{\sqrt{n}}\right)\right| \mathrm{d}\mathbf{z} \\
&\lesssim \left(n\beta e^{\frac{t^2}{2}}\right)^{O(1)} \delta^{n-O(1)} \\
&\lesssim e^{-\frac{\omega(\beta)}{4}}
\end{aligned} \tag{A.3}$$

for sufficiently large n. Finally, we bound similar to Lemma 3.2:

$$\begin{aligned}
&\int_{\|\mathbf{z}\|>a\sqrt{n}} \left|\partial^\alpha \frac{\varphi}{6\sqrt{n}} \sum_{k=0}^3 \kappa_t^{(k,3-k)} H^{(k,3-k)}\right|(\mathbf{z})\,\mathrm{d}\mathbf{z} \\
&\qquad \lesssim n^{-\frac{1}{2}} \sum_{k=0}^3 \kappa_t^{(k,3-k)} \int_{\mathbb{R}^2} \left|\partial^\alpha H^{(k,3-k)}\varphi\right|(\mathbf{z})\,\mathrm{d}\mathbf{z} \\
&\qquad \lesssim n^{-\frac{1}{2}}\eta_3 \int_{\mathbb{R}^2} \|\mathbf{z}\|^{O(1)} e^{-\frac{\|\mathbf{z}\|^2}{2}}\,\mathrm{d}\mathbf{z} \\
&\qquad \lesssim e^{-\frac{\omega(\beta)}{4}}.
\end{aligned} \tag{A.4}$$

Combining (A.2) to (A.4) proves (A.1) and hence Lemma 3.4. □

A.4 Proof of Proposition 4.1

Point *1* in Theorem 2.1 can readily be seen to hold because of the explicit form of both of the fields. We focus on showing Point *3*, the proof of which can be repeated essentially verbatim to deduce Point *2*.

Observe that $\mu_t = \nu_t^{*n}$, where ν_t is the law of

$$\begin{bmatrix} G(t) \\ G'(t) \end{bmatrix} = \begin{bmatrix} g(Z) \\ g'(Z) \end{bmatrix}$$

with $Z \sim N(t,1)$ and $g(z) = ze^{-\beta z^2/2}$. (Also, for $n = 1$ we have $\mu_t = \nu_t$, and ν_t cannot have a continuous density on $\mathbb{R}^2$, since both components of a drawn

random vector $(G(t), G'(t))$ are functions of the same one-dimensional Gaussian random variable.)

We first show that $\mathscr{F}(\nu_t^{*n}) = (\mathscr{F}\nu_t)^n \in L^1(\mathbb{R}^2)$ for any fixed $t \in T$, and without loss of generality we take $t = 0$. This would imply that μ_t has a density $p_t \in \mathscr{C}^0(\mathbb{R}^2)$ satisfying $p_t(\mathbf{x}) \to 0$ as $\|\mathbf{x}\| \to \infty$ by virtue of Fourier inversion and the Riemann-Lebesgue lemma. We also perform computations as if ν_t^{*n} were already a function, and all arguments can be justified by appealing to the framework of Schwarz distributions $\mathcal{S}'(\mathbb{R}^2)$ and duality.

We have

$$\begin{aligned}\int_{\mathbb{R}^2} |\mathscr{F}(\nu_t)(\xi)^n| \,\mathrm{d}\xi &= \int_{\mathbb{R}^2} |\mathscr{F}(\nu_t)(\xi)|^n \,\mathrm{d}\xi \\ &= \int_{\mathbb{R}^2} \left| \int_{\mathbb{R}} e^{-\mathrm{i}(\xi_1 g(x) + \xi_2 g'(x))} e^{-\frac{x^2}{2}} \,\mathrm{d}x \right|^n \mathrm{d}\xi.\end{aligned}$$

Recalling that $n \to \infty$ in our regime, we can suppose $n \geqslant 4$, and for the above integral to be finite, it suffices to show that

$$\left| \int_{\mathbb{R}} e^{-\mathrm{i}(\xi_1 g(x) + \xi_2 g'(x))} e^{-\frac{x^2}{2}} \,\mathrm{d}x \right| \lesssim \frac{1}{\sqrt{\xi_1 + \xi_2}} \qquad \text{as } \xi_1, \xi_2 \to \infty.$$

Observe that the critical points of g are $\pm\sqrt{\frac{1}{\beta}}$, whereas those of g' are 0 and $\pm\sqrt{\frac{3}{\beta}}$. Since $\beta \to \infty$ as well, pick $\varepsilon > 0$ sufficiently small and such that $\varepsilon > \sqrt{\frac{3}{\beta}}$. We first see that

$$\begin{aligned}&\left| \int_{|x|>\varepsilon} e^{-\mathrm{i}(\xi_1 g(x) + \xi_2 g'(x))} e^{-\frac{x^2}{2}} \,\mathrm{d}x \right| \\ &= \left| \int_{|x|>\varepsilon} \frac{1}{\mathrm{i}} \frac{1}{\xi_1 g'(x) + \xi_2 g''(x)} \frac{\mathrm{d}}{\mathrm{d}x} \Big(e^{-\mathrm{i}(\xi_1 g(x) + \xi_2 g'(x))} \Big) e^{-\frac{x^2}{2}} \,\mathrm{d}x \right| \\ &\lesssim \frac{1}{\xi_1 + \xi_2},\end{aligned}$$

where we used integration by parts to obtain the last inequality—this is in fact an elementary version of the method of non-stationary phase. For the integral over $\{|x| \leqslant \varepsilon\}$, we look to use the method of stationary phase as $\xi_1, \xi_2 \to \infty$, by distinguishing the three regimes $\xi_1 \gg \xi_2$, $\xi_2 \gg \xi_1$, and $\xi_1 \sim \xi_2$. When $\xi_1 \gg \xi_2$, we have

$$\begin{aligned}&\left| \int_{|x|\leqslant\varepsilon} e^{-\mathrm{i}(\xi_1 g(x) + \xi_2 g'(x))} e^{-\frac{x^2}{2}} \,\mathrm{d}x \right| \\ &= \left| \int_{|x|\leqslant\varepsilon} e^{-\mathrm{i}(\xi_1+\xi_2)\left(\frac{\xi_1}{\xi_1+\xi_2} g(x) + \frac{\xi_2}{\xi_1+\xi_2} g'(x) \right)} e^{-\frac{x^2}{2}} \,\mathrm{d}x \right| \\ &= \left| \int_{|x|\leqslant\varepsilon} e^{-\mathrm{i}(\xi_1+\xi_2)(g(x) + O(\xi_1 \varepsilon))} e^{-\frac{x^2}{2}} \,\mathrm{d}x \right|\end{aligned}$$

$$\lesssim \frac{1}{\sqrt{\xi_1+\xi_2}} + o\Big(\frac{1}{\sqrt{\xi_1+\xi_2}}\Big)$$

by the method of stationary phase applied to the phase g, since g'' is non-degenerate at the critical points $\pm\sqrt{\frac{1}{\beta}}$. Similarly when $\xi_2 \gg \xi_1$, we have

$$\begin{aligned}\left|\int_{|x|\leqslant\varepsilon} e^{-\mathrm{i}(\xi_1 g(x)+\xi_2 g'(x))} e^{-\frac{x^2}{2}}\,\mathrm{d}x\right| &= \left|\int_{|x|\leqslant\varepsilon} e^{-\mathrm{i}(\xi_1+\xi_2)(g'(x)+O(\xi_2\varepsilon))} e^{-\frac{x^2}{2}}\,\mathrm{d}x\right| \\ &\lesssim \frac{1}{\sqrt{\xi_1+\xi_2}} + o\Big(\frac{1}{\sqrt{\xi_1+\xi_2}}\Big)\end{aligned}$$

by the method of stationary phase applied to the phase g', since g''' is non-degenerate at the critical points 0 and $\pm\sqrt{\frac{3}{\beta}}$. The same argument then carries through when $\xi_1 \sim \xi_2$, using the phase $g+g'$. This yields the desired conclusion.

To deduce that $(t,\mathbf{x}) \mapsto p_t(\mathbf{x})$ is continuous on $T\times\mathbb{R}^2$, we note that

$$p_t(\mathbf{x}) = \frac{1}{(2\pi)^2}\int_{\mathbb{R}^2} e^{\mathrm{i}\langle \mathbf{x},\mathbf{z}\rangle} \int_{\mathbb{R}^n} \exp\left(-\mathrm{i}\left\langle \mathbf{z}, \begin{bmatrix}\sum_{j=1}^n g(\xi_j)\\ \sum_{j=1}^n g'(\xi_j)\end{bmatrix}\right\rangle\right)\gamma_t(\xi_1)\cdots\gamma_t(\xi_n)\,\mathrm{d}\xi\,\mathrm{d}\mathbf{z}.$$

We can conclude by the Lebesgue dominated convergence theorem. □

References

[AAČ13] Antonio Auffinger, Gérard Ben Arous, and Jiří Černỳ. Random matrices and complexity of spin glasses. *Communications on Pure and Applied Mathematics*, 66(2):165–201, 2013.

[AT09] Robert J Adler and Jonathan E Taylor. *Random fields and geometry*. Springer Science & Business Media, 2009.

[AW09] Jean-Marc Azaïs and Mario Wschebor. *Level sets and extrema of random processes and fields*. John Wiley & Sons, 2009.

[BPA24] Giuseppe Bruno, Federico Pasqualotto, and Andrea Agazzi. Emergence of meta-stable clustering in mean-field transformer models. *arXiv preprint arXiv:2410.23228*, 2024.

[BR10] Rabi N Bhattacharya and R Ranga Rao. *Normal approximation and asymptotic expansions*. SIAM, 2010.

[Che95] Yizong Cheng. Mean shift, mode seeking, and clustering. *IEEE transactions on pattern analysis and machine intelligence*, 17(8):790–799, 1995.

[CM02] Dorin Comaniciu and Peter Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 24(5):603–619, 2002.

[CP00] Miguel A Carreira-Perpinán. Mode-finding for mixtures of gaussian distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11):1318–1323, 2000.

[CP07] Miguel A Carreira-Perpinán. Gaussian mean-shift is an EM algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(5):767–776, 2007.

[CP15] Miguel A Carreira-Perpinán. A review of mean-shift algorithms for clustering. *arXiv preprint arXiv:1503.00687*, 2015.

[CPW03] Miguel A Carreira-Perpinán and Christopher KI Williams. On the number of modes of a gaussian mixture. In *International Conference on Scale-Space Theories in Computer Vision*, pages 625–640. Springer, 2003.

[DG85] Luc Devroye and László Györfi. *Nonparametric density estimation.* Wiley Series in Probability and Mathematical Statistics: Tracts on Probability and Statistics. John Wiley & Sons, Inc., New York, 1985. The L_1 view.

[FH75] Keinosuke Fukunaga and Larry Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on information theory*, 21(1):32–40, 1975.

[FMM21] Zhou Fan, Song Mei, and Andrea Montanari. TAP free energy, spin glasses and variational inference. *The Annals of Probability*, 49(1):1 – 45, 2021.

[GKPR24] Borjan Geshkovski, Hugo Koubbi, Yury Polyanskiy, and Philippe Rigollet. Dynamic metastability in the self-attention model. *arXiv preprint arXiv:2410.06833*, 2024.

[GLPR23] Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. A mathematical perspective on transformers. *arXiv preprint arXiv:2312.10794*, 2023.

[GLPR24] Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. The emergence of clusters in self-attention dynamics. *Advances in Neural Information Processing Systems*, 36, 2024.

[Gre00] Emmanuel Grenier. On the nonlinear instability of euler and prandtl equations. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 53(9):1067–1091, 2000.

[GRRB24] Borjan Geshkovski, Philippe Rigollet, and Domènec Ruiz-Balet. Measure-to-measure interpolation using transformers. *arXiv preprint arXiv:2411.04551*, 2024.

[KM97] V. Konakov and E. Mammen. The shape of kernel density estimates in higher dimensions. *Math. Methods Statist.*, 6(4):440–464 (1998), 1997.

[MAB20] Antoine Maillard, Gérard Ben Arous, and Giulio Biroli. Landscape complexity for the empirical risk of generalized linear models. In *Mathematical and Scientific Machine Learning*, pages 287–327. PMLR, 2020.

[Mam95] E. Mammen. On qualitative smoothness of kernel density estimates. *Statistics*, 26(3):253–267, 1995.

[MMF92] E. Mammen, J. S. Marron, and N. I. Fisher. Some asymptotics for multimodality tests based on kernel density estimates. *Probab. Theory Related Fields*, 91(1):115–132, 1992.

[RL14] Alex Rodriguez and Alessandro Laio. Clustering by fast search and find of density peaks. *Science*, 344(6191):1492–1496, 2014.

[SABP22] Michael E Sander, Pierre Ablin, Mathieu Blondel, and Gabriel Peyré. Sinkformers: Transformers with doubly stochastic attention. In *International Conference on Artificial Intelligence and Statistics*, pages 3515–3530. PMLR, 2022.

[Tsy09] Alexandre B. Tsybakov. *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York, 2009.

[VSP+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Borjan Geshkovski
Inria & Laboratoire Jacques-Louis Lions
Sorbonne Université
4 Place Jussieu
75005 Paris, France
e-mail: borjan.geshkovski@inria.fr

Philippe Rigollet
Department of Mathematics
Massachusetts Institute of Technology
77 Massachusetts Ave
Cambridge 02139 MA, United States
e-mail: rigollet@math.mit.edu

Yihang Sun
Department of Mathematics
Stanford University
450 Jane Stanford Way Building 380
Stanford, CA 94305, United States
e-mail: kimisun@stanford.edu