Minimax properties of gamma kernel density estimators under L^p loss and β-Hölder smoothness of the target

Frédéric Ouimet[1]

[a] *Département de mathématiques et d'informatique, Université du Québec à Trois-Rivières, Trois-Rivières, Canada*

Abstract

This paper considers the asymptotic behavior in β-Hölder spaces, and under L^p loss, of the gamma kernel density estimator introduced by Chen [Ann. Inst. Statist. Math. 52 (2000), 471–480] for the analysis of nonnegative data, when the target's support is assumed to be upper bounded. It is shown that this estimator can achieve the minimax rate asymptotically for a suitable choice of bandwidth whenever $(p, \beta) \in [1, 3) \times (0, 2]$ or $(p, \beta) \in [3, 4) \times ((p - 3)/(p - 2), 2]$. It is also shown that this estimator cannot be minimax when either $p \in [4, \infty)$ or $\beta \in (2, \infty)$.

Keywords: Density estimation, gamma kernel, L^p loss, minimax estimation, nonnegative data, nonparametric estimation.
2020 MSC: Primary: 62G07; Secondary: 62G05, 62G20

1. Introduction

Let $X_1, \ldots, X_n$ be independent and identically distributed (iid) random variables taking values in $[0, \infty)$ with unknown density f. Estimating f nonparametrically is a classical problem, and kernel density estimators (KDEs) in the sense of ?] and ?] remain a standard tool; see, e.g., [? ? ?]. When the support of f is constrained (for instance $[0, \infty)$ or $[0, 1]$), the use of symmetric kernels in the usual KDE creates a boundary bias near the endpoints, which can significantly deteriorate global L^p risks. A large literature addresses this issue using reflection and pseudodata mechanisms [? ? ?], transformation approaches [?], or various boundary correction techniques [? ? ? ? ? ?].

An alternative approach is to use *asymmetric* kernels whose support matches that of the target density and whose shape adapts with the evaluation point. A prominent contribution is due to ?], who introduced *beta kernel* estimators for densities supported on $[0, 1]$, and to ?], who proposed a *gamma kernel* estimator for densities supported on $[0, \infty)$. In addition to density estimation, ?] developed beta-kernel smoothers for regression curves, and ?] studied local linear regression smoothers based on asymmetric kernels, including gamma-type constructions. Many extensions and refinements of gamma kernel estimators have appeared since then; see, for example, the semiparametric density estimation approach of ?] based on copulas, applications in econometrics [? ? ? ? ?], and many other relevant references such as [? ? ? ? ? ? ? ? ? ?].

Some of these techniques have notably been adapted to more complex supports such as the simplex [? ? ? ? ? ? ? ? ?], product spaces like the unit hypercube or positive orthant [? ? ? ? ?], half-spaces [?], and the cone of positive definite matrices [? ?]. Parallel to these specific developments, general frameworks for multivariate asymmetric kernels (called associated kernels) have been proposed and studied; see, e.g., [? ? ? ?] and references therein.

While pointwise asymptotic expansions (bias, variance, limit distributions) for asymmetric kernel density estimators are by now well documented, their *minimax* performance under integrated losses is more delicate. The difficulty stems from the fact that the smoothing induced by such kernels is spatially inhomogeneous, which complicates the control of global risk. In particular, depending on the support of the target density and on the loss under consideration, the variance or higher-order moments of the estimator may fail to be uniformly integrable. This lack of global moment control can prevent attainment of the classical minimax rate even when boundary bias is substantially reduced. This phenomenon was analyzed for beta kernel density estimators by **?**], who identified regimes of smoothness and loss for which minimax optimality holds or fails. Those conclusions were later extended to Dirichlet kernel density estimators on the simplex by **?**].

The purpose of this paper is to establish parallel minimax and non-minimax results for the gamma kernel density estimator introduced by **?**]. To isolate the boundary behavior at 0 and to avoid additional tail phenomena in global L^p risks, we work over Hölder-type classes of densities supported on a compact interval (throughout, on $[0, 1]$, noting that any density supported on $[0, C]$ for $C \in (0, \infty)$ can be rescaled to this domain). For such classes, the benchmark minimax rate under L^p loss is of order $n^{-\beta/(2\beta+1)}$; see, e.g., **?**]. Our main results show that, with a suitable bandwidth choice, the gamma kernel density estimator achieves this rate for all $p \in [1, 3)$ and $\beta \in (0, 2]$, and also for a nontrivial subset of (p, β) with $p \in [3, 4)$ and $\beta \in (0, 2]$. On the other hand, we prove two complementary non-minimaxity statements: the estimator cannot be minimax for any p when $\beta \in (2, \infty)$, and it cannot be minimax when $p \in [4, \infty)$ even if $\beta \in (0, 2]$. These conclusions mirror those obtained for beta and Dirichlet kernel density estimators.

The paper is organized as follows. Section **??** introduces the gamma kernel density estimator, the risk criteria, the definition of β-smoothness, and other relevant notational conventions. Section **??** states the main results, i.e., the regions of (p, β) where we have minimaxity and non-minimaxity. Section **??** briefly examines the regularity of mirrored gamma densities truncated to be supported on $[0, 1]$ within the proposed functional classes. Section **??** outlines several directions for future work. Proofs are gathered in Section **??**, with proofs of some technical lemmas relegated to Appendix **??**.

2. Definitions and notation

Let $X_1, \ldots, X_n$ be a sequence of independent and identically distributed (iid) random variables with an unknown density f supported on $[0, \infty)$. The goal is to estimate f from the sample.

For a smoothing parameter (bandwidth) $b \in (0, \infty)$ and a point $x \in [0, \infty)$, define the *gamma kernel*

$$K_b(x, t) \equiv K_{x/b+1,b}(t) = \frac{t^{x/b} e^{-t/b}}{b^{x/b+1}\Gamma(x/b + 1)} \mathbb{1}_{[0,\infty)}(t), \quad t \in [0, \infty), \tag{1}$$

where $\Gamma(\cdot)$ is Euler's gamma function. The associated gamma kernel density estimator is

$$\hat{f}_{n,b}(x) = \frac{1}{n} \sum_{i=1}^{n} K_{x/b+1,b}(X_i), \quad x \in [0, \infty). \tag{2}$$

The family $\{\hat{f}_{n,b} : b > 0\}$ is indexed by the smoothing parameter b, which is typically taken to depend on the sample size n.

Let ξ_x be a gamma random variable with density $t \mapsto K_{x/b+1,b}(t)$ on $[0, \infty)$. Then

$$\mathsf{E}[\hat{f}_{n,b}(x)] = \int_0^\infty K_{x/b+1,b}(t) f(t) \mathrm{d}t = \mathsf{E}[f(\xi_x)].$$

Moreover, for this parametrization, one has

$$\mathsf{E}(\xi_x) = x + b, \quad \mathsf{Var}(\xi_x) = xb + b^2.$$

Hence, the smoothing induced by $\hat{f}_{n,b}$ is *spatially inhomogeneous*: the typical fluctuation scale of ξ_x around x is of order $\sqrt{xb}$ when x/b is large. This feature makes the choice of the functional class (and the control of global L^p risks) more delicate than in compact-support settings.

For $p \in [1, \infty)$ and a measurable function $g : [0, \infty) \to \mathbb{R}$, write

$$\|g\|_p = \left(\int_0^\infty |g(x)|^p \mathrm{d}x \right)^{1/p}.$$

For an estimator f_n of f, define its L^p risk at f by

$$R_n(f_n, f) = \left\{ \mathsf{E}\big(\|f_n - f\|_p^p\big) \right\}^{1/p},$$

whenever the expectation exists. For a class of densities $\mathcal{F}$, define the maximal risk

$$R_n(f_n, \mathcal{F}) = \sup_{f \in \mathcal{F}} R_n(f_n, f),$$

and the minimax risk

$$r_n(\mathcal{F}) = \inf_{f_n} R_n(f_n, \mathcal{F}),$$

where the infimum is over all estimators based on $(X_1, \ldots, X_n)$.

For $\beta \in (0, \infty)$, let

$$m = \sup\{\ell \in \mathbb{N}_0 : \ell < \beta\},$$

and define the β-Hölder class $\Sigma(\beta, L)$ as the set of all densities f supported on $[0, 1]$ that are m-times differentiable on $(0, \infty)$ (so a jump discontinuity is allowed at 0 but not at 1) and such that

$$\max_{0 \le k \le m} \sup_{u \in (0,\infty)} |f^{(k)}(u)| \le L \quad \text{and} \quad \sup_{\substack{u,v \in (0,\infty) \\ u \neq v}} \frac{|f^{(m)}(u) - f^{(m)}(v)|}{|u - v|^{\beta - m}} \le L,$$

where $f^{(k)}$ denotes the kth derivative of f (with $f^{(0)} = f$).

Remark 1. Here, the definition of $\Sigma(\beta, L)$ is stated for densities supported on $[0, 1]$ for simplicity. Indeed, a new space, say $\widetilde{\Sigma}(\beta, L)$, could be defined analogously for densities supported on $[0, C]$ for $C \in (0, \infty)$. All results in the paper would still be valid, given that $\tilde{f} \in \widetilde{\Sigma}(\beta, \tilde{L})$ if and only if $\tilde{f}(\cdot) = C f(C\, \cdot)$ for some $f \in \Sigma(\beta, L)$.

The main question is whether the family $\{\hat{f}_{n,b} : b > 0\}$ can achieve the minimax rate over $\Sigma(\beta, L)$ under L^p loss, for an appropriate choice of $b = b_n$.

Throughout the paper, expectation is taken with respect to the joint law of the mutually independent copies $X_1, \dots, X_n$ of X. The notation $u = \mathcal{O}(v)$ means that $\limsup |u/v| < B < \infty$ as $n \to \infty$ or $b \to 0$, depending on the context. The positive constant B may depend on the risk exponent p, the smoothness parameter β, the Lipschitz constant L, and the target density f, but no other variable unless explicitly written as a subscript. Similarly, throughout the proofs, $c, C \in (0, \infty)$ denote generic positive constants whose value may change from expression to expression and which may depend on p, β, L, and f, but not on n, b. If both $u = \mathcal{O}(v)$ and $v = \mathcal{O}(u)$ hold, then one writes $u \asymp v$. Similarly, the notation $u = \mathrm{o}(v)$ means that $\lim |u/v| = 0$ as $n \to \infty$ or $b \to 0$. Subscripts indicate which parameters the convergence rate can depend on. If f_n is any estimator of f, then $\mathsf{E}[f_n]$ is a shorthand for the map $x \mapsto \mathsf{E}[f_n(x)]$. The gamma distribution always has the shape/scale parametrization.

3. Main results

For one-dimensional density estimation under β-Hölder smoothness assumptions, the minimax rate under L^p loss is

$$r_n\{\Sigma(\beta, L)\} \asymp n^{-\beta/(2\beta+1)}; \tag{3}$$

see, e.g., **?** , Theorem 5.1] and **?** , Theorem 2.8].

The following theorem states a minimax property of the gamma kernel density estimator (**??**) when the bandwidth is tuned according to the smoothness parameter β.

Theorem 1. *Let $L > 0$ be given. Define*

$$\mathcal{S} = \left\{ (p, \beta) \in [3, 4) \times (0, 2] : \frac{p-3}{p-2} < \beta \leq 2 \right\}.$$

Assume that $(p, \beta) \in [1, 3) \times (0, 2]$ or that $(p, \beta) \in \mathcal{S}$. Let $b_n = c\, n^{-2/(2\beta+1)}$ for all $n \in \mathbb{N}$ and some constant $c \in (0, \infty)$. Then

$$\limsup_{n \to \infty} \frac{R_n\{\hat{f}_{n,b_n}, \Sigma(\beta, L)\}}{r_n\{\Sigma(\beta, L)\}} < \infty,$$

i.e., the sequence $\{\hat{f}_{n,b_n} : n \in \mathbb{N}\}$ achieves the minimax rate over $\Sigma(\beta, L)$ under L^p loss.

Remark 2. The somewhat restrictive definition of $\Sigma(\beta, L)$ (smoothness on $(0, \infty)$ even though $\mathrm{supp}(f) \subseteq [0, 1]$) implicitly enforces a *compatibility condition at the upper endpoint* $x = 1$: since $f(x) = 0$ for $x > 1$ and the derivatives are assumed to exist and be bounded on $(0, \infty)$, one necessarily has $f(1) = 0$ and, when $\beta > 1$, also $f'(1) = 0$, etc. This prevents an additional "endpoint leakage" bias coming from the fact that the gamma kernel has support on $[0, \infty)$ and is not truncated at 1.

To see what would happen without this restriction, imagine working instead with the more standard Hölder-type class of densities supported on $[0, 1]$ that are β-Hölder on $(0, 1)$ but with *no* matching condition at $x = 1$ (so that, for instance, $f(1^-)$ may be strictly positive). In Step 4 of the proof of Theorem **??**, the bound $|\mathsf{E}[f(\xi_x)] - f(x)| = \mathcal{O}(b^{\beta/2})$ is obtained by controlling

$|f(\xi_x) - f(x)|$ through Hölder regularity. If f is allowed to have a nonzero left limit at 1, then for x close to 1, we must also account for the event $\{\xi_x > 1\}$, on which $f(\xi_x) = 0$ while $f(x)$ can be of order one. A simple decomposition is

$$|\mathsf{E}[f(\xi_x)] - f(x)| \leq \mathsf{E}\left(|f(\xi_x) - f(x)| \mathbb{1}_{\{\xi_x \leq 1\}}\right) + f(x)\,\mathsf{P}(\xi_x > 1).$$

The first term behaves as before (of order $b^{\beta/2}$), but the second term can be much larger: when x lies within the kernel's typical fluctuation scale of the endpoint, i.e. $|1-x| = \mathcal{O}(b^{1/2})$, the tail probability $\mathsf{P}(\xi_x > 1)$ is not small (it is typically of constant order), so the pointwise bias can be $\asymp 1$ on an interval of length $\asymp b^{1/2}$ around 1 (on both sides of 1 when we integrate over $[0,\infty)$). This yields an additional integrated bias contribution of order

$$\left(\int_{1-c\,b^{1/2}}^{1+c\,b^{1/2}} |\mathsf{E}[\hat{f}_{n,b}(x)] - f(x)|^p \,\mathrm{d}x\right)^{1/p} \asymp b^{1/(2p)}.$$

Heuristically, the maximal risk bound in the proof of Theorem **??** would therefore become

$$R_n\{\hat{f}_{n,b}, \Sigma(\beta, L)\} \leq C\Big(n^{-1/2} b^{-1/4} + b^{\beta/2} + b^{1/(2p)}\Big),$$

where the extra term $b^{1/(2p)}$ comes from the lack of regularity at $x = 1$.

If $\beta \leq 1/p$, then $b^{\beta/2}$ dominates $b^{1/(2p)}$ as $b \to 0$, so this endpoint effect is negligible and the choice $b \asymp n^{-2/(2\beta+1)}$ still leads to the usual rate $n^{-\beta/(2\beta+1)}$. In contrast, if $\beta > 1/p$, then $b^{1/(2p)}$ dominates $b^{\beta/2}$, and the best achievable rate for the gamma kernel density estimator comes from balancing $n^{-1/2} b^{-1/4}$ with $b^{1/(2p)}$, which gives $b \asymp n^{-2p/(p+2)}$ and a resulting rate $n^{-1/(p+2)}$ (the rate corresponding to an "effective" smoothness $1/p$). This explains why imposing a smooth matching at $x = 1$ is important if one wants to recover the faster minimax rate for smoother targets.

The following results identify regimes of the loss exponent p and smoothness parameter β for which the gamma kernel density estimator $\hat{f}_{n,b}$ defined in (**??**) cannot achieve the minimax rate over the β-Hölder class $\Sigma(\beta, L)$, irrespective of the choice of bandwidth sequence $(b_n)_{n\in\mathbb{N}}$.

Proposition 2. *Let $p \in [1,\infty)$ and $\beta \in (2,\infty)$ be given. There exists $L > 1$ such that for every bandwidth sequence $(b_n)_{n\in\mathbb{N}} \subseteq (0,1)$, the family $\{\hat{f}_{n,b_n} : n \in \mathbb{N}\}$ satisfies*

$$\liminf_{n\to\infty} \frac{R_n\{\hat{f}_{n,b_n}, \Sigma(\beta, L)\}}{r_n\{\Sigma(\beta, L)\}} = +\infty.$$

Proposition 3. *Let $p \in [4,\infty)$ and $\beta \in (0,2]$ be given. There exists $L > 1$ such that for every bandwidth sequence $(b_n)_{n\in\mathbb{N}} \subseteq (0,1)$, the family $\{\hat{f}_{n,b_n} : n \in \mathbb{N}\}$ satisfies*

$$\liminf_{n\to\infty} \frac{R_n\{\hat{f}_{n,b_n}, \Sigma(\beta, L)\}}{r_n\{\Sigma(\beta, L)\}} = +\infty.$$