Rates of Convergence of Maximum Likelihood Estimators Via Entropy Methods

Robin Evans

St Catharine's College

I declare that this essay is my own work done as part of the Part III Examination. It is the result of my own work, and except where stated otherwise, includes nothing which was performed in collaboration. No part of this essay has been submitted for a degree or any such qualification.

Signed .....

Home Address

26 Alpraham Crescent
Upton
Chester
CH2 1QX

Rates of Convergence of Maximum Likelihood Estimators Via Entropy Methods

Contents

1 Introduction

This essay explores the idea of using independent and identically distributed observations of a random variable to estimate its probability density function. In particular, we consider how maximum likelihood estimation can be extended to non-parametric contexts, and how quickly our estimate will converge to the correct density. This convergence rate is, perhaps unsurprisingly, related to the 'size' of the set of densities which we maximise over; we measure this size by introducing *entropy* in Section 2. Section 3 contains the major results, and in Section 6 we discover a slight variation which leads to better rates in some cases.

1.1 Maximum Likelihood Estimation

Maximum Likelihood estimation began with the work of R.A. Fisher, in his 1912 paper, "An absolute criterion for fitting frequency curves", [7]. It is a very simple, intuitively sensible idea: given what we have seen, which distribution was *the most likely* to have produced these observations? In parametric models we may use this idea to find parameter values which maximise the likelihood function, but we should always keep in mind that it is the density which we are interested in estimating, not just the parameters for their own sake. This motivates our first definition.

Definition 1.1
Let $X_1, \dots X_n$ be independent and identically distributed (i.i.d.) random variables with unknown probability density function p_0, and let $\mathcal{P}$ be a collection of densities which contains p_0. We say that $\hat{p}_n = \hat{p}_n(X_1, \dots, X_n) \in \mathcal{P}$ is a **maximum likelihood estimator (MLE)** for p_0 if for any $p \in \mathcal{P}$

$$\prod_{i=1}^{n} \hat{p}_n(X_i) \geq \prod_{i=1}^{n} p(X_i).$$

Notice that this definition makes no mention of parameters, but since any sensible parametrisation is identifiable, it is irrelevant whether we refer to the density or the parameter as being an MLE. We argue that it actually makes more sense to consider the MLE to be a density, as this (rather pathological) example shows.

Example 1.1
Let $X_1, \dots, X_n$ be i.i.d. normal random variables, with unknown mean μ and known variance σ^2. Suppose we make the following (bijective) reparametrisation of μ:

$$\lambda = \begin{cases} 0 & \text{for } \mu = 0 \\ \mu^{-1} & \text{otherwise.} \end{cases}$$

Then it is easy to show (and not very surprising) that an MLE for λ is

$$\hat{\lambda} = \bar{X}^{-1} \mathbb{1}_{\{\bar{X} \neq 0\}}.$$

So we reach an interesting situation: if the true value is $\lambda = 0$, then we know that $\bar{X} \to 0$ almost surely; yet since $\mathbb{P}(\bar{X} = 0) = 0$, this means that $|\lambda - \hat{\lambda}| \to \infty$ a.s.

This example is rather artificial, since no-one would ever choose to use such a parametrisation, but the point to keep in mind is that we are really interested distributions, not parameters, and consistency of parameter estimates will only be useful if the parametrisation is sensible; certainly it should be continuous. We will therefore think of a family of densities $\mathcal{P}$ as being a (pseudo)metric space $(\mathcal{P}, d)$, often under the Hellinger distance.

Definition 1.2
Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let p_1 and p_2 be densities with respect to a σ-finite measure μ. We define the **Hellinger distance** between p_1 and p_2 as

$$h(p_1, p_2) = \left(\int_\Omega \left[p_1^{1/2} - p_2^{1/2} \right]^2 d\mu \right)^{\frac{1}{2}}.$$

Remark 1.1
We can write this as

$$h(p_1, p_2) = \left\| p_1^{1/2} - p_2^{1/2} \right\|_2,$$

where $\| \cdot \|_p$ is the $L^p(\mu)$ norm.

The Hellinger distance between two densities is bounded above by $\sqrt{2}$, although some authors prefer to include a multiplicative constant which ensures that $h \leq 1$. As is pointed out by Birgé and Massart (1993) [4], this boundedness is one of the Hellinger distance's nicest properties. An alternative would be the more 'natural'[1] Kullback-Leibler information number,

$$K(p) = \mathbb{E}_{p_0} \left\{ \log \frac{p(X)}{p_0(X)} \right\},$$

but unfortunately, even if the likelihood function is uniformly bounded over all p there is no guarantee of MLE consistency under K because it is unbounded.

We can use h to define a pseudometric on $\mathcal{P}$.

1.2 Finite-Dimensional Parameter Spaces

The case of a finite-dimensional parameter space is a familiar one: let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, $\mathcal{X}$ be a sampling space, and $\mathcal{P} = \{p_\theta : \mathcal{X} \to \mathbb{R}^+\}$ be a set of densities, indexed by a finite-dimensional set Θ. We observe $X_1, X_2, \ldots$ i.i.d. random variables in $\mathcal{X}$ with density p_{θ_0} for unknown $\theta_0 \in \Theta$.

We let $p_0 = p_{\theta_0}$ and $\hat{p}_n = p_{\hat{\theta}_n}$, which is a slight abuse of notation, but recognises our new view of the MLE as a density. We are aware that under appropriate conditions[2] the parameter MLE is consistent and asymptotically efficient, so we think of it as a 'good' estimator. Does this necessarily mean that the density MLE converges to the true density? Our Example 1.1 suggests that it does not, which is not surprising. In order to compare the two properties, we first define the notion of a convergence rate for densities.

Definition 1.3
Let $\mathcal{P}$ be a space of densities with a pseudometric d, suppose we have densities $p_n, p \in \mathcal{P}$ such

[1] At least in the view of Birgé and Massart.
[2] An exponential family of densities is, for example, sufficient.

that $d(p_n, p) \to 0$. We say that the **convergence rate** of p_n to p is $O(\epsilon_n)$ if

$$d(p_n, p) = O_{\mathbb{P}}(\epsilon_n).$$

If $a_n = O(b_n)$ and $b_n = O(a_n)$, we write

$$a_n \asymp b_n.$$

To see the relationship between convergence of a sequence of densities and convergence of a sequence of parameters, we will consider the example of exponential families.

Example 1.2
Let $\mathcal{P} = \{p_\theta : \theta \in \Theta\}$ be an exponential family of densities on $\mathbb{R}$, i.e.

$$p_\theta(x) = \rho(x) \exp\{\tau(\theta) \cdot t(x) - c(\theta)\},$$

where c, τ are twice differentiable functions. Then for $p_1, p_2 \in \mathcal{P}$,

$$\begin{aligned} h(p_1, p_2)^2 &= \int_{\mathbb{R}} p_1 + p_2 - 2\sqrt{p_1 p_2}\, d\mu \\ &= 2 - 2\int_{\mathbb{R}} \rho(x) \exp\left\{\tfrac{1}{2}[\tau(\theta_1) + \tau(\theta_2)]t(x) - \tfrac{1}{2}[c(\theta_1) + c(\theta_2)]\right\} \mu(dx) \end{aligned}$$

so by Taylor expanding $\tau(\theta_2)$ and $c(\theta_2)$ around θ_1,

$$\begin{aligned} &= 2 - 2\int_{\mathbb{R}} \rho(x) \exp\left\{\tau(\theta_1)t(x) - c(\theta_1) + \frac{1}{2}(\theta_2 - \theta_1)(\tau'(\theta_1)t(x) + c'(\theta_1)) + O((\theta_2 - \theta_1)^2)\right\} \mu(dx) \\ &= 2 - 2\int_{\mathbb{R}} p_1(x) \exp\left\{\frac{1}{2}(\theta_2 - \theta_1)(\tau'(\theta_1)t(x) + c'(\theta_1)) + O((\theta_2 - \theta_1)^2)\right\} \mu(dx) \\ &= 2 - 2\int_{\mathbb{R}} p_1(x) \left[1 + \frac{1}{2}(\theta_2 - \theta_1)(\tau'(\theta_1)t(x) + c'(\theta_1)) + O((\theta_2 - \theta_1)^2)\right] \mu(dx) \\ &= -\int_{\mathbb{R}} p_1(x) \left[(\theta_2 - \theta_1)(\tau'(\theta_1)t(x) + c'(\theta_1)) + O((\theta_2 - \theta_1)^2)\right] \mu(dx) \\ &= O((\theta_2 - \theta_1)^2), \end{aligned}$$

where the linear term vanishes because it is proportional to the score function, and we have informally assumed that the error is uniform. So we see that $\hat{\theta}_n$ converging to θ_0 at rate $O_{\mathbb{P}}(n^{-\frac{1}{2}})$ implies that $\hat{p}_n$ converges to p_0 at rate $O_{\mathbb{P}}(n^{-\frac{1}{2}})$ under the Hellinger metric.

The $O(n^{-\frac{1}{2}})$ rate for density convergence can be recovered for most parametric examples; we can see this intuitively by examining a familiar result concerning the Wald statistic.

Lemma 1.1
Take as our sampling space $\mathcal{X} = \mathbb{R}^d$. Let $\mathcal{P} = \{p_\theta : \theta \in \Theta\}$, with $\Theta \subseteq \Theta^*$ where Θ^* is a finite dimensional linear space and $\dim \Theta = d$; let $X_1, \ldots, X_n$ be independent observations sampled from a distribution with density $p_0 = p_{\theta_0} \in \mathcal{P}$. Let $i_1(\theta)$ be the Fisher Information for one observation with parameter θ, and assume that it is non-singular. Then the MLE $\hat{\theta}_n$ for θ_0 based on the first n observations satisfies

$$\sqrt{n}\, i_1(\theta_0)^{\frac{1}{2}}(\hat{\theta}_n - \theta_0) \xrightarrow{d} Z \sim N_d(0, I).$$

We may ask however, what happens if we don't have a finite-dimensional parameter space? Indeed, how do we choose Θ and consequently $\mathcal{P}$ in the first place? When using real data we may be able to use common sense to choose a set of possible distributions; to give a simple example, if our data are the number of car crashes in the UK each month for two years, we could try to fit a Poisson model, since it is reasonable to assume that each individual crash is independent. This is not so easy in general, and leads us to turn to non-parametric statistics.

1.3 Non-Parametric MLEs

Readers who are familiar with non-parametric statistics will know that their two main features are: (i) robustness, and (ii) that they are not generally as powerful as their parametric counterparts. The study of MLEs proves to be no exception to this. The following example shows why it is desirable to place some restriction on the set of densities.

Example 1.3
Let $\mathcal{X}$ be a closed interval in $\mathbb{R}$. Suppose we sample $X_1, \ldots, X_n$ independently from p_0, an unknown density. If we take $\mathcal{P}$ to be the set of *all* possible densities, it is clear that an MLE for p_0 is

$$\frac{1}{n}\sum_{i=1}^{n}\delta_i,$$

where δ_i is a δ-mass centred on X_i; the likelihood is infinite in this case[3]. This is the empirical distribution, giving equal probability $\frac{1}{n}$ to each observation, and has many nice properties such as weak convergence (see for example [11]), however, if p_0 is continuous, our density estimate is zero p_0 almost everywhere and will not converge under any sensible metric. It is clear then, that the MLE does not converge to the true density in the way we want and in this sense it is not consistent.

The moral of this example is that we cannot choose to work over a set of densities which is too large. Section 2 gives us a way of measuring the 'size' of a set of densities, but before we explore this, there is one more compromise we will sometimes be forced to make.

Definition 1.4
Suppose $X_1, \ldots, X_n$ are i.i.d. with density p_0 and let $\eta_n \to 0$ be a positive real sequence. We say that $\hat{p}_n = \hat{p}_n(X_1, \ldots, X_n) \in \mathcal{P}$ is an $\boldsymbol{\eta_n}$**-MLE** for p_0 if for any $p \in \mathcal{P}$

$$\prod_{i=1}^{n}\hat{p}_n(X_i) \geq \prod_{i=1}^{n}p(X_i) - \eta_n.$$

The motivation for this definition is clear: if we work over an infinite dimensional set of densities, it is possible that there is no exact MLE, or that we are unable to calculate it exactly. An η_n-MLE will allow us to approximate the true MLE and some of its properties.

[3]Formally speaking this is not a density, but for continuous p_0 we could find a sequence of proper densities $\tilde{p}_m$ with $h(\tilde{p}_m, p_0) \to \sqrt{2}$ as $m \to \infty$ and where the likelihood ratio is arbitrarily large; to see this, take triangular peaks around the observations and let them get thinner and taller. The limit of this process is the δ-masses.

2 Entropy

2.1 Entropy

Intuitively one would imagine that the rate of convergence of any estimator will be related to the 'size' of the set of densities over which we work; this is indeed the case, and in order to define an appropriate measure of this size, we introduce the concept of entropy, proceeding in a similar manner to Kolmogorov and Tihomirov [10].

Definition 2.1
Let (U, d) be a metric space.

$V \subseteq U$ is called an **ϵ-net** for U if for every $u \in U$, we can find $v \in V$ with $d(u, v) \leq \epsilon$.

For each i in some indexing set I, let $W_i \subseteq U$ such that the diameter of each W_i is at most 2ϵ. We call $\mathcal{W} = \{W_i : i \in I\}$ an **ϵ-covering** of U if

$$\bigcup_{i \in I} W_i = U.$$

We say U is **centred** if for any $A \subseteq U$ with diameter $2k$, there exists a point $x \in U$ such that $B_k(x)$, the closed ball of radius k around x, contains A.

Remark 2.1
It is worth taking a few moments to reflect on these definitions, which are best understood heuristically. An ϵ-net is a collection of points in the space U such that we can never be further than ϵ from one of them; an ϵ-covering is merely a covering with sets no 'bigger' than 2ϵ. Clearly we can form a covering from a net simply by taking closed balls of radius ϵ around each point in the net.

Where U is centred, we can form a net from a covering: given $\mathcal{W}$, we can find points $x_i \in W_i$ so that the balls $B_\epsilon(x_i)$ also form a covering; then $\{x_i\}$ is an ϵ-net.

Not every space is centred: consider the unit circle in $\mathbb{R}^2$ under the Euclidean metric. The circle itself has diameter 1, so we can trivially form a 1-covering of order 1, but we cannot form a 1-net of order 1 since every point *on the circle* is exactly 2 units from the point opposite it[4].

The definitions are perfectly valid if d is only a pseudometric on U.

Definition 2.2
We let $N(\epsilon, U, d)$ be the minimal cardinality of an ϵ-covering of U (possibly infinite). Then the **ϵ-entropy** of U is defined by

$$H(\epsilon, U, d) = \log N(\epsilon, U, d).$$

If $N(\epsilon, U, d) < \infty$ for every $\epsilon > 0$, we say that (U, d) is **totally bounded**.

Remark 2.2
For convenience, when U is centred (often the case) we can use the cardinality of minimal ϵ-nets to calculate the entropy.

[4]Elementary geometry shows us that a minimal 1-net is of order 3.

Both $N(\epsilon, U, d)$ and $H(\epsilon, U, d)$ are non-increasing and right continuous. The first of these properties is easy to see, a proof of the second can be found in Section 1 of [10].

An obvious first question arising from this definition is why we choose to use $H(\epsilon, U, d)$ rather than $N(\epsilon, U, d)$. Suppose we have some random variables $\{Z_i\}$ which are in some sense 'exponentially bounded', i.e.

$$\mathbb{P}(Z_i \geq a) \leq \exp(-f(a)),$$

for some increasing function f, commonly $f(a) = a^2$. Then

$$\begin{aligned} \mathbb{P}\left(\max_{i=1,\ldots,N} Z_i \geq a\right) &\leq N \exp(-f(a)) \\ &= \exp(\log N - f(a)), \end{aligned}$$

and thus we are more likely to be interested in the logarithm of N, since we require $\log N < f(a)$ for the inequality to hold any meaning. Here we start to see the connection between maximum likelihood and entropy: let Z_i be a likelihood function for some density p_i, so

$$Z_i = Z(p_i) = \prod_{j=1}^{n} \frac{p_i(X_j)}{p_0(X_j)}.$$

Suppose that $\mathbb{P}(Z_i \geq a) \leq \exp(-f(a))$ when p_i is in some sense 'not very close' to p_0. Then the sketched argument above gives us a way of asking how likely it is that the maximum of all p_i 'not very close' to p_0 is large. Then since we know $Z(\hat{p}_n) \geq 1$, this inequality gives us information about how likely it is that $\hat{p}_n$ is among those densities which are 'not very close' to p_0, and if it is unlikely, then the convergence of $\hat{p}_n$ to p_0 is fast. It also suggests that we cannot choose a space of densities which is too large or the bound will become greater than 1, and thus trivial.

A more detailed explanation of the intuition behind entropy can be found in Kolmogorov and Tihomirov's 1959 paper on the subject [10]. The concept comes from information theory, and is defined by them as $\log_2 N(\epsilon, U, d)$. The idea is related to the number of binary digits required to relay a finite set of distinct messages with a noisy signal; we use the natural logarithm, but this differs only by a multiplicative constant.

We will often treat the metric d as implicit, and write

$$N(\epsilon, U) = N(\epsilon, U, d), \qquad H(\epsilon, U) = H(\epsilon, U, d).$$

For most of this essay, d will be the Hellinger distance, $h(\cdot, \cdot)$.

Example 2.1
We introduce two relatively simple examples of the explicit calculation of entropy, both taken from Kolmogorov and Tihomirov (1959) [10].

(i) Let $U = [a, b]$ and d be the Euclidean Norm. Then

$$N(\epsilon, U) = \left\lceil \frac{b-a}{\epsilon} \right\rceil.$$

(ii) Let $\mathcal{F}$ be the set of real-valued functions f on $[a, b]$, such that $f(a) = 0$, and f is Lipschitz continuous with constant L. Working under the supremum norm,

$$H(\epsilon, \mathcal{F}) = \left\lceil \frac{(b-a)L}{\epsilon} - 1 \right\rceil \log 2.$$

Proof of Example 2.1
(i) is easy, and we leave it to the reader.

For (ii), first notice that there is an isometry under the supremum norm from $\mathcal{F}$ to the set of real-valued Lipschitz continuous functions on $\Delta = [0, (b-a)L]$ with constant 1. This new space, $\mathcal{F}'$ say, is centred, so we can proceed by constructing a minimal ϵ-net (see Remark 2.2). Let

$$n = \left\lceil \frac{(b-a)L}{\epsilon} \right\rceil,$$

and let $\mathcal{G}$ be a set containing the 2^{n-1} distinct functions defined by

$$g(0) = g(\epsilon) = 0 \qquad g(k\epsilon) = g((k-1)\epsilon) \pm \epsilon$$

for each integer $2 \leq k \leq n$, and linear in between. Figure 2.1 shows a few example elements of $\mathcal{G}$; the solid triangle shows the edge of possible values which an element of $\mathcal{F}'$ could take, given the Lipschitz condition which constrains it.

Now consider any $f \in \mathcal{F}'$; there is at least one $g \in \mathcal{G}$ such that $|f(k\epsilon) - g(k\epsilon)| \leq \epsilon$ at each integer $k \leq n$, due to the way we constructed $\mathcal{G}$. Figure 2.1 shows an example g with its 'ϵ-corridor' shaded in grey, and one segment slightly darker. Suppose that for some $f \in \mathcal{F}'$, $|f(3\epsilon) - g(3\epsilon)| \leq \epsilon$ and $|f(4\epsilon) - g(4\epsilon)| \leq \epsilon$ — i.e. it takes values on the left and right hand sides of the dark segment. Then it follows that $|f(x) - g(x)| \leq \epsilon$ for all $x \in [3\epsilon, 4\epsilon]$: the upper and lower boundaries of the darker shaded area have gradient 1, therefore it is impossible for f to 'leave' the shaded area below because in order to 're-enter', it would violate the Lipschitz condition. Similarly, it cannot 'leave' the shaded area above, because this would also violate the Lipschitz condition.

Hence for any $f \in \mathcal{F}'$, we can find $g \in \mathcal{G}$ with

$$\sup_{x \in [0, L(b-a)]} |f(x) - g(x)| \leq \epsilon,$$

so $\mathcal{G}$ is an ϵ net of order 2^{n-1}. This gives the required result. □

One can prove that this is a minimal net by introducing ϵ-separated sets. We leave this for the reader to investigate herself (see [10]).

2.2 Entropy With Bracketing

There is a stronger form of an ϵ-net which we will need to make use of.

Definition 2.3
Let d be a metric defined upon a set U of real valued functions, and let

$$\mathcal{W}_\epsilon = \{(f_i^l, f_i^u) : f_i^l \leq f_i^u,\ d(f_i^l, f_i^u) \leq \epsilon \quad \text{for all } i \in I\}$$

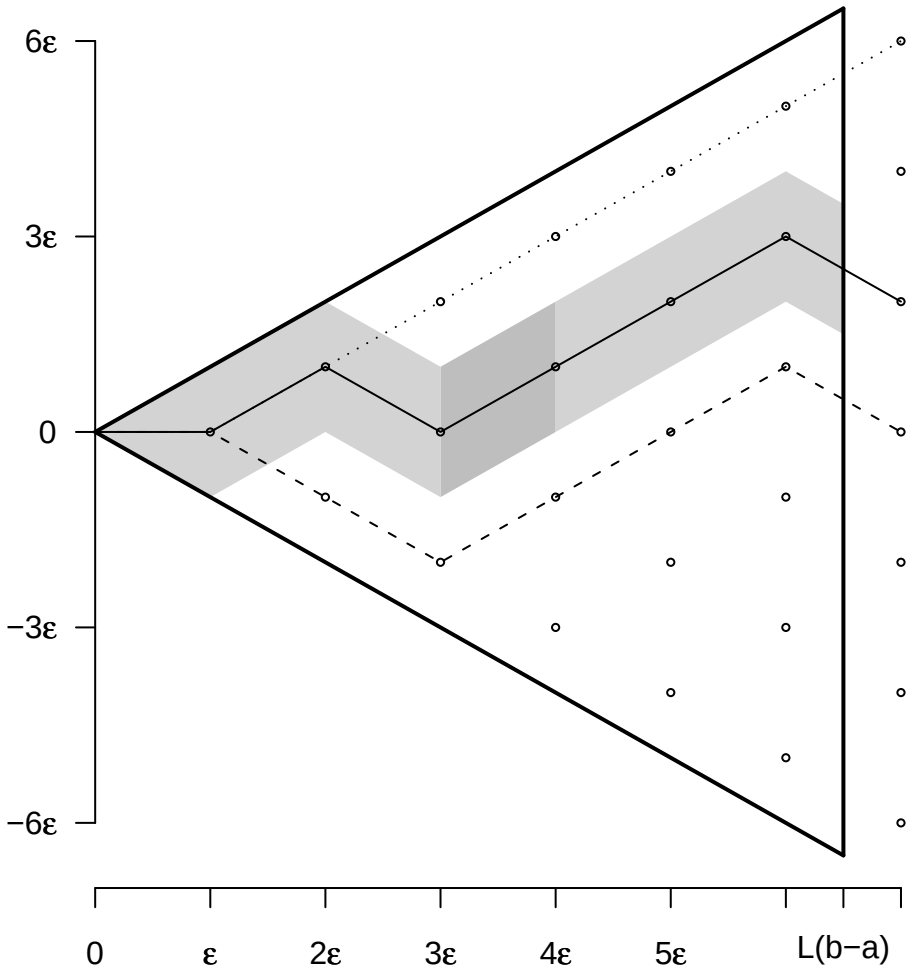

Figure 1: Examples of g.

be a set of pairs of functions in U. We say that $\mathcal{W}$ **ϵ-brackets** U if for any $g \in U$, we can find $i \in I$ with $f_i^l \leq g \leq f_i^u$.

Let $N_B(\epsilon, U, d)$ be the minimal cardinality of a set I such that this holds, then

$$H_B(\epsilon, U, d) = \log N_B(\epsilon, U, d)$$

is called the **ϵ-entropy with bracketing** for U.

We will sometimes write $H_B(\epsilon, U)$ for $H_B(\epsilon, U, d)$ where it has been clearly stated what metric we are working in. The idea here is that we require all our functions to be squeezed between a pair in our bracketing set, although in fact we will not always require that f_i^l and f_i^u are members of the class $\mathcal{F}$.

Example 2.2
Consider the set $\mathcal{F}'$ from Example 2.1 again, still using the supremum norm. Looking at Figure 2.1, it is easy to see that we can take as our pairs of functions the top and bottom of each ϵ-corridor, giving us the result

$$H_B(2\epsilon, \mathcal{F}') \leq H(\epsilon, \mathcal{F}').$$

Indeed, it is easy to show that the above result is true for general $\mathcal{F}'$ when working under the supremum norm.

Lemma 2.1
Let $\mathcal{G}_{m,\alpha}(M)$ be the set of functions $g : [0,1] \to \mathbb{R}$ such that g is m-times differentiable, and the mth derivative of g is α-Hölder with constant M. Then for every $\mathcal{H} \subseteq \mathcal{G}_{m,\alpha}(M)$ and such that $\mathcal{H}$ is bounded with respect to the L^2 norm,

$$H_B(\delta, \mathcal{H}, \|\cdot\|_2) = O(\delta^{-\frac{1}{m+\alpha}}).$$

Furthermore this bound cannot be improved, in the sense that there exists a bounded set $\mathcal{H} \subseteq \mathcal{G}_{m,\alpha}(M)$ with

$$\delta^{\frac{1}{m+\alpha}} = O\left(H_B(\delta, \mathcal{H}, \|\cdot\|_2)\right).$$

We will prove this result for $m = 0$ so as to give the reader a flavour of how one can show that an entropy achieves a minimum order, but first we need to take a combinatorial diversion: the following lemma is found in Birgé (1983) [3].

Lemma 2.2
Let $\mathcal{T} = \{0,1\}^n$ with a metric d defined on $t, u \in \mathcal{T}$ by

$$d(t,u) = \sum_{i=1}^{n} |t_i - u_i|,$$

i.e. the number of places at which t and u differ. Further, given $m < \frac{1}{8}n$, let $\mathcal{U}_m \subset \mathcal{T}$ be a maximal subset such that $d(u,u') \geq m$ for all $u, u' \in \mathcal{U}_m$. Then for sufficiently large n

$$\log |\mathcal{U}_m| \geq 0.3n.$$

Proof
Under d, a closed ball of radius $m-1$ has order

$$K_{m-1} = \sum_{i=0}^{m-1} \binom{n}{i}.$$

If $m < \frac{1}{8}n$ then for $i < m$

$$\binom{n}{i} \leq \frac{n-i}{7(i+1)} \binom{n}{i} = \frac{1}{7} \binom{n}{i+1}$$

giving

$$K_{m-1} \leq \frac{7}{6} \binom{n}{m-1}.$$

By maximality, closed balls of radius $m-1$ around each point in $\mathcal{U}_m$ must cover $\mathcal{T}$, and hence

$$|\mathcal{U}_m| \geq \frac{2^n}{K_{m-1}} \geq \frac{6}{7} 2^n \frac{(m-1)!(n-m+1)!}{n!}.$$

A version of Stirling's formula,

$$\sqrt{2\pi n}\, n^n e^{-n} \exp\left(\tfrac{1}{12n+1}\right) \leq n! \leq \sqrt{2\pi n}\, n^n e^{-n} \exp\left(\tfrac{1}{12n}\right),$$

gives us (for $m < \frac{1}{8}n$)

$$\begin{aligned} |\mathcal{U}_m| &\geq \frac{6}{7} 2^n \sqrt{2\pi \tfrac{(m-1)(n-m+1)}{n}} \frac{(m-1)^{m-1}(n-m+1)^{n-m+1}}{n^n} \underbrace{e^{\frac{1}{12(m-1)+1} + \frac{1}{12(n-m+1)+1} - \frac{1}{12n}}}_{\geq 1} \\ &\geq \frac{6}{7} 2^n \sqrt{2\pi p(1-p)n}\, p^{pn} (1-p)^{(1-p)n} \end{aligned}$$

where we have set $p = \frac{m-1}{n}$. By taking logarithms and differentiating it is easy to see that for large n this expression has a unique local minimum which tends towards $p = \frac{1}{2}$, and thus for sufficiently large n, $p = \frac{1}{8}$ will give us the minimum on $p \in [0, \frac{1}{8}]$. So, since $\frac{m-1}{n} < \frac{1}{8}$,

$$|\mathcal{U}_m| \geq \frac{3}{4} \sqrt{\frac{2\pi n}{7}} \frac{7^{\frac{7n}{8}}}{4^n},$$

and

$$\log |\mathcal{U}_m| \geq \frac{7n}{8} \log 7 - n \log 4 > 0.316n.$$

□

Proof of Lemma 2.1
Let $m = 0$. A proof for entropy (without bracketing) when working with the supremum norm can be found in Section 5 of Kolmogorov and Tihomirov (1959) [10]; this immediately gives the first part of the theorem, since

$$H_B(\delta, \mathcal{F}, \|\cdot\|_2) \leq H(2\delta, \mathcal{F}, \|\cdot\|_\infty).$$

For the second part, we turn to a proof used in van de Geer [14]; consider

$$\mathcal{H} = \{g \in \mathcal{G}_{0,\alpha}(1) : |g| \leq 1\}.$$

Then take a series of numbers

$$0 = a_0 < a_1 < \cdots < a_N = 1,$$

with $a_k = k\delta^{\frac{1}{\alpha}}$ for $k = 1, \ldots, N-1$ and $N\delta^{\frac{1}{\alpha}} \geq 1$. Define $\psi_k(x)$ to be a triangular function between a_{k-1} and a_k for $k = 1, \ldots, N-1$, each with height $\delta/2$ (see Figure 2).

Given $\zeta \in \{-1, 1\}^{N-1}$, let

$$g_\zeta(x) = \sum_{k=1}^{N-1} \zeta_k \psi_k(x);$$

an example is shown in Figure 3. Since $\psi_k \in \mathcal{H}$ for each k, and no two of them are non-zero at the same point, then $g_\zeta \in \mathcal{H}$ for every ζ. Also notice $\int \psi_k^2 = \frac{1}{12}\delta^{2+\frac{1}{\alpha}}$, so if ζ and ζ' differ in at least $\frac{N-1}{8}$ places then

$$\begin{aligned} \int (g_\zeta - g_{\zeta'}) &> \frac{N-1}{8}\frac{1}{6}\delta^{2+\frac{1}{\alpha}} \\ &> \frac{N}{8^2}\delta^{2+\frac{1}{\alpha}} \\ &\geq \left(\frac{\delta}{8}\right)^2, \end{aligned}$$

where we have assumed $N > 4$ for our second inequality.

Now let $\mathcal{U}_N \subset \{-1, 1\}^{N-1}$ be maximal such that any two elements of $\mathcal{U}_N$ differ in at least $N/8$ places, and let $\mathcal{H} = \{g_\zeta : \zeta \in \mathcal{U}_N\}$; any two elements of $\mathcal{H}$ are at least $\delta/8$ apart under the L^2 norm. Then by Lemma 2.2, for sufficiently large N we have

$$|\mathcal{H}| = |\mathcal{U}_N| \geq e^{0.3(N-1)}.$$

Hence for some $A > 0$,

$$H_B\left(\frac{\delta}{8}, \mathcal{H}, \|\cdot\|_2\right) \geq A\delta^{-\frac{1}{\alpha}}$$

for sufficiently small δ. □

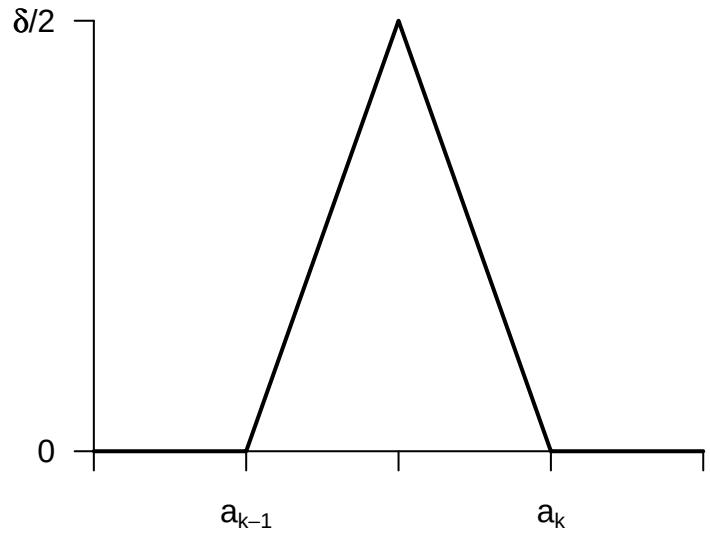

Figure 2: $\psi_k(x)$.

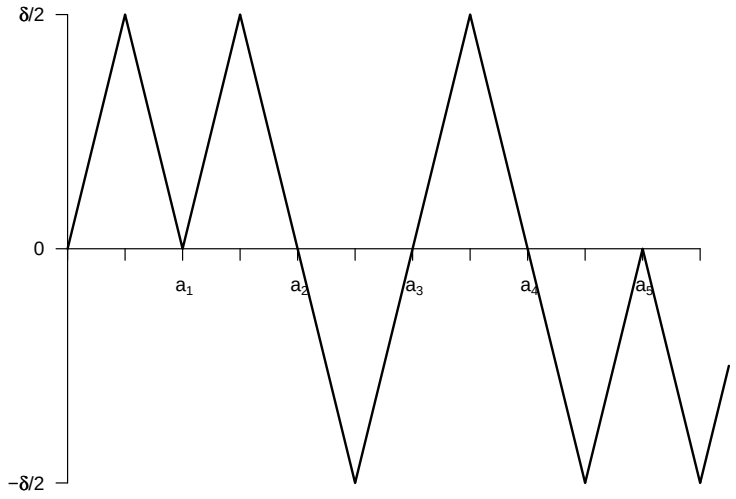

Figure 3: $g_\zeta(x)$ for $\zeta = (1, 1, -1, 1, -1, -1, \ldots)$.

3 Convergence

Now that we have defined a suitable notion of 'size', we need to establish theorems which will link this new concept with the rate of convergence. Most of what follows comes from Wong and Shen (1995) [18], though it builds on work done by Shen and Wong (1994) [12], Birgé and Massart (1993) [4], and van de Geer (1993) [13], amongst others.

3.1 A Link Between MLE Convergence and Entropy

Theorems 3.1 and 3.2 allow us to connect the concepts of entropy and convergence; the proofs involve many technical intermediate results, and can be found in Section 8. We assume that $Y_1, \ldots, Y_n$ are i.i.d. random variables from a distribution with density p_0.

Theorem 3.1
Let $\mathcal{P}$ be a classes of densities containing p_0, and let $H_B(\delta, \mathcal{P})$ be the Hellinger δ-entropy with bracketing for $\mathcal{P}$. There exist positive constants c_1, c_2, c_3, c_4 such that if

$$\int_{\delta^2/2^8}^{\sqrt{2}\delta} H_B^{1/2}(u/c_3, \mathcal{P})\, du \leq c_4 n^{\frac{1}{2}} \delta^2, \tag{1}$$

then for sufficiently large n

$$\mathbb{P}^* \left(\sup_{\{p \in \mathcal{P} : \|p^{1/2} - p_0^{1/2}\|_2 \geq \delta\}} \prod_{i=1}^{n} \frac{p(Y_i)}{p_0(Y_i)} \geq \exp(-c_1 n \delta^2) \right) \leq 4 \exp(-c_2 n \delta^2). \tag{2}$$

Here $\mathbb{P}^*$ is the outer probability operator for $\mathbb{P}$, since we cannot be certain the above event is measurable.

Remark 3.1
The essence of the theorem is that the likelihood ratio is uniformly exponentially small outside a Hellinger ball of radius δ around the true density. Thus since we know that

$$\prod_{i=1}^{n} \frac{\hat{p}_n(Y_i)}{p_0(Y_i)} \geq 1 > \exp(-c_1 n \delta^2),$$

this suggests that for large n, $\hat{p}_n$ is likely to lie inside a Hellinger ball of radius δ around p_0.

Theorem 3.2
Let $\delta_n \to 0$ be a sequence such that the integral condition (1) is satisfied with $\delta = \delta_n$ for each n. Let $\hat{p}_n$ be an η_n-MLE for p_0, with $\eta_n \leq c_1 \delta_n^2$. Then for sufficiently large n,

$$\mathbb{P}\left(h(\hat{p}_n, p_0) \geq \delta_n\right) \leq 4 \exp(-c_2 n \delta_n^2). \tag{3}$$

Remark 3.2
Thus the convergence rate is $O(\delta_n)$.

The combination of these two theorems essentially says that the convergence of the MLE is determined by the equation

$$\int_{\delta^2}^{\delta} H_B^{1/2} = n^{\frac{1}{2}}\delta^2;$$

we will see this put into practice in Example 4.1.

3.1.1 An Envelope Condition

Though the result is a nice one and serves as the heart of this essay, it does suffer from one major drawback: that we are imposing an **envelope condition** on the densities. In general, given a set of functions $\mathcal{G}$, if we insist that

$$H_B(u, \mathcal{G}, \|\cdot\|_p) < \infty$$

for all $u > 0$, we are implicitly requiring that

$$\int G(x)\,dx < \infty$$

where $G = \sup_{\mathcal{G}} |g|$. To see this, notice first that we must have

$$\sup_{\mathcal{G}} \|g\|_1 \leq R$$

for some constant R. So let $\left\{[g_j^L, g_j^U]\right\}_{j=1}^N$ be a u-bracketing of $\mathcal{G}$. Then if $g_j^L \leq g \leq g_j^U$, we have that

$$|g| \leq |g_j^L| + |g_j^U - g_j^L| \leq \sum_{j=1}^{N} \left(|g_j^L| + |g_j^U - g_j^L|\right)$$

and this last bound is uniform over $\mathcal{G}$. Hence

$$\|G\|_1 \leq \sum_{j=1}^{n} \left(\|g_j^L\|_1 + \|g_j^U - g_j^L\|_1\right) \leq N(R+u).$$

3.2 Variations

3.2.1 Entropy Conditions

Van de Geer [14] gives a result equivalent to Theorem 3.2, but which imposes entropy conditions on a modified set of densities. Let $\mathcal{P}$ be a class of densities containing the 'true' density p_0, and define

$$\bar{\mathcal{P}}^{\frac{1}{2}} = \left\{ \left(\frac{p+p_0}{2}\right)^{\frac{1}{2}} : p \in \mathcal{P} \right\};$$

let $\hat{p}_n$ be the maximum likelihood estimator of p_0 based upon the first n observations. We write $H_B(\delta, \mathcal{P}, \|\cdot\|_2)$ to mean the δ-entropy with bracketing using the L^2 norm, and $\mu_0 = \mu \mathbb{1}_{\{p_0>0\}}$.

Theorem 3.3
Suppose that δ_n is a sequence satisfying

$$\Psi(\delta_n) := \delta_n \vee \int_{\delta_n^2/2^{13}}^{\delta_n} H_B^{1/2}(u, \bar{\mathcal{P}}^{\frac{1}{2}}, \|\cdot\|_2)\, du \leq c_1^{-1} n^{1/2} \delta_n^2$$

for some constant c_1. Then provided $\Psi(\delta)/\delta^2$ is non-increasing,

$$\mathbb{P}(h(\hat{p}_n, p_0) > \delta_n) \leq c_2 \exp\left[-\frac{n\delta_n^2}{c_2^2}\right]$$

for some universal constant c_2.

Remark 3.3
Since

$$H_B(u, \bar{\mathcal{P}}^{\frac{1}{2}}, \|\cdot\|_2) = H_B(u, \bar{\mathcal{P}}, h),$$

and one can show that

$$h\left(\frac{p+p_0}{2}, p_0\right) \leq \frac{1}{\sqrt{2}}\, h(p, p_0) \leq 2\sqrt{2}\, h\left(\frac{p+p_0}{2}, p_0\right),$$

this result is simply a restatement of Theorems 3.1 and 3.2.

The reason for presenting this slightly different form of the theorem is to emphasise how difficult it can be to calculate the entropy; if we know the L^2 entropy (or even the L^∞ entropy) of $\mathcal{P}$, it tells us nothing about the Hellinger entropy of $\mathcal{P}$, because for small p_1, p_2 then $|\sqrt{p_1} - \sqrt{p_2}|$ may be much larger than $|p_1 - p_2|$. Essentially this is because $\sqrt{\cdot}$ is not differentiable at 0. This leads many of the examples on the subject to look slightly contrived, either because we do not allow the densities to become small (unless they are zero), or because we impose conditions on their square roots.

3.2.2 Convexity

Suppose now that $\mathcal{P}$ is a convex class of densities, and define

$$\mathcal{P}^* = \left\{\frac{2pp_0}{p+p_0} : p \in \mathcal{P}\right\}.$$

Again from van de Geer [14], we have the following result.

Theorem 3.4
Suppose that δ_n is a sequence satisfying

$$\Psi(\delta_n) := \delta_n \vee \int_{\delta_n^2/c_1}^{\delta_n} H_B^{1/2}(u, \mathcal{P}^*, \|\cdot\|_2)\, du \leq c_1^{-1} n^{1/2} \delta_n^2$$

for some constant c_1. Then provided $\Psi(\delta)/\delta^2$ is non-increasing,

$$\mathbb{P}(h(\hat{p}_n, p_0) > \delta_n) \leq c_2 \exp\left[-\frac{n\delta_n^2}{c_2^2}\right]$$

for some universal constant c_2.

Remark 3.4
The advantage of this version of the theorem is that it does not entail an envelope condition, and we will see an example in which this theorem gives a rate which cannot be shown using Theorem 3.2.

4 Examples

Example 4.1
Let $\mathcal{G} = \left\{g : g \leq K,\ \int g^2 = 1,\ g \in \mathcal{G}_{m,\alpha}(1)\right\}$, where $\mathcal{G}_{m,\alpha}(M)$ is the set of real functions whose mth derivative exists and is α-Hölder with constant M. Then $\mathcal{P} = \{p = g^2 : g \in \mathcal{G}\}$ is a class of densities. We know from Lemma 2.1 that

$$H_B(\delta, \mathcal{G}, \|\cdot\|_2) = O(\delta^{-\frac{1}{m+\alpha}}),$$

and since

$$h(p_1, p_2) = \left\| p_1^{1/2} - p_2^{1/2} \right\|_2 = \|g_1 - g_2\|_2$$

we have $H_B(\delta, \mathcal{P}, h) = O(\delta^{-\frac{1}{m+\alpha}})$. The integral from (1) is then

$$\begin{aligned} \int_{\delta^2/2^8}^{\sqrt{2}\delta} H_B^{1/2}(u/c_3, \mathcal{P}_n)\, du &= \int_{\delta^2/2^8}^{\sqrt{2}\delta} O(u^{-\frac{1}{2(m+\alpha)}})\, du \\ &= O(\delta^{1-\frac{1}{2(m+\alpha)}} + \delta^{2-\frac{1}{m+\alpha}}) \\ &= O(\delta^{1-\frac{1}{2(m+\alpha)}}). \end{aligned}$$

since we are interested in behaviour as $\delta \to 0$. Then we wish to choose the fastest rate δ_n such that (1) is satisfied; that is to say

$$\delta^{1-\frac{1}{2(m+\alpha)}} \asymp n^{\frac{1}{2}}\delta^2.$$

Rearranging gives

$$\delta_n = O(n^{-\frac{m+\alpha}{2(m+\alpha)+1}}).$$

Remark 4.1
This method is typical for finding convergence rates using the integral conditions: the constants involved can be largely ignored, and we concern ourselves with finding the fastest rate δ such that

$$\int_{\delta^2}^{\delta} H_B^{1/2}(u, \mathcal{P})\, du \leq n^{\frac{1}{2}}\delta^2.$$

Example 4.2
Let $\mathcal{F} = \{f : \mathbb{R}_+ \to [0,1],\ f \text{ decreasing}\}$. Then by adapting results found in Birman and Solomjak (1967) [6], it can be shown that

$$H_B(\delta, \mathcal{F}, \|\cdot\|_2) \leq A\delta^{-1}$$

for some constant A. Then since $\mathcal{F} = \mathcal{F}^{\frac{1}{2}}$, we have

$$H_B(\delta, \mathcal{F}, h) \leq A\delta^{-1},$$

and so if we let $\mathcal{P} = \{p : \mathbb{R}_+ \to [0, B],\ \int p = 1,\ p \text{ decreasing}\}$, for some B, we can get an MLE convergence rate of

$$\delta_n = O(n^{-\frac{1}{3}}).$$

This is a nice result, as there are many real life situations where we would choose a density which implies things are always more likely to happen earlier than later, and this gives us great generality.

Example 4.3
Let $\mathcal{P} = \left\{p : [0,1] \to \mathbb{R}_+,\ \int p\, d\mu = 1,\ |p'| \leq M < \infty\right\}$. This time we make no specific condition on the square roots of the densities, so there is no way to bound the bracketing entropy of $\mathcal{P}$ under the Hellinger distance. However, it is shown in van de Geer [14] that subject to some conditions on p_0, the class

$$\mathcal{P}^* = \left\{ \frac{2\,pp_0}{p + p_0} : p \in \mathcal{P} \right\}$$

is such that $H_B(\delta, \mathcal{P}^*, \|\cdot\|_2) < A\delta^{-1}$ for some constant A. Then by applying Theorem 3.4, we can show that

$$h(\hat{p}_n, p_0) = O_{\mathbb{P}}(n^{-\frac{1}{3}}).$$

This is another good result, since our set of densities is quite large, and it demonstrates the value of being able to work with $\mathcal{P}^*$. However, the conditions which we have to impose on p_0 for this to work are fairly constrictive; for example, it is sufficient that for some $n > 1$

$$\limsup_{x\to\infty} x^n p_0(x) < \infty, \quad \limsup_{x\to 0} x^{\frac{1}{n}} p_0(x) < \infty.$$

5 Limitations

5.1 The Parametric Rate

As discussed in the introduction, the rate of convergence for MLEs in most parametric contexts is $O(n^{-\frac{1}{2}})$, and it would seem reasonable to try and extract this rate from the theorems introduced in Section 3. In fact, the δ-entropy for most finite dimensional spaces is $O(\log \delta^{-1})$, from which Theorem 3.2 only gives an MLE convergence rate of $O(n^{-\frac{1}{2}} \log n)$.

To recover the correct rate, we have to replace condition (1) by a slightly weaker one. Instead of considering the entropy with bracketing of $\mathcal{P}$, we concern ourselves with the **local entropy with bracketing**, i.e. (1) becomes

$$\int_{\delta^2/2^8}^{\sqrt{2}\delta} H_B^{1/2}(u/c_3, \mathcal{P}(\delta))\, du \leq c_4 n^{\frac{1}{2}} \delta^2,$$

where

$$\mathcal{P}(\delta) = \{p \in \mathcal{P} : h(p, p_0) \leq \delta\}.$$

The proof that Theorem 3.1 still holds when we use the local entropy follows similarly to that of the original; more details can be found in [14] and [18].

5.2 A Problem

Let $\mathcal{G} \subseteq \mathcal{G}_{0,\alpha}(K)$ be bounded, where $\mathcal{G}_{0,\alpha}(K)$ is the collection of all α-Hölder functions with constant K on $[0,1]$; then we know from Lemma 2.1 that there exists a constant A such that for sufficiently small $\delta > 0$

$$H_B(\delta, \mathcal{G}) \leq A\delta^{-\frac{1}{\alpha}}.$$

Thus

$$J(\delta) = \int_{\delta^2}^{\delta} H_B^{1/2}(u, \mathcal{G})\, du \leq \int_{\delta^2}^{\delta} A u^{-\frac{1}{2\alpha}}\, du,$$

giving

$$J(\delta) = \begin{cases} O(\delta^{1-\frac{1}{2\alpha}}) & \alpha > \frac{1}{2} \\ O(\log \delta^{-2}) & \alpha = \frac{1}{2} \\ O(\delta^{2-\frac{1}{\alpha}}) & \alpha < \frac{1}{2}. \end{cases}$$

Applying Theorems 3.1 and 3.2, we get

$$h(\hat{p}_n, p_0) = \begin{cases} O_{\mathbb{P}}(n^{-\frac{\alpha}{2\alpha+1}}) & \alpha > \frac{1}{2} \\ O_{\mathbb{P}}(n^{-\frac{1}{4}} \log^{\frac{1}{2}} n) & \alpha = \frac{1}{2} \\ O_{\mathbb{P}}(n^{-\frac{\alpha}{2}}) & \alpha < \frac{1}{2}. \end{cases}$$

In the next section we will, by using sieves, see that this is not the fastest rate that can be shown for MLE convergence; the central problem here is that the parameter space is too 'large'

for $\alpha \leq \frac{1}{2}$, in the sense that the integral

$$\int_{\delta}^{1} H_B^{1/2}(u, \mathcal{G})\, du$$

diverges as $\delta \to 0$. It should be noted that we cannot immediately tell the reason for this non optimal rate: it could either be that $\hat{p}_n$ is not optimal, or because Theorem 3.1 does not give the true convergence rate of $\hat{p}_n$. A more detailed discussion in Birgé and Massart (1993) [4] shows that in a similar case it is indeed the fault of the choice of estimator, and that we can improve things by choosing to maximise over a smaller space.

6 Sieves

In order to deal with the problem of our parameter space being too large, we introduce the concept of a sieve.

Definition 6.1
Let $\epsilon_n \to 0$ be a real sequence, and let (U, d) be a metric space, with a sequence of subsets $U_n \subseteq U$. Suppose that for each n, U_n is an ϵ_n-net for U; then we call U_n an $\boldsymbol{\epsilon_n}$**-sieve**. So for each $u \in U$ we can find $u' \in U_n$ such that $d(u, u') \leq \epsilon_n$.

Remark 6.1
The idea here is to ensure that if our parameter space is 'too large', as discussed in Example 1.3, we can use a sequence of smaller spaces which approximate the full space. This can be used to prevent the integral of the square root entropy from diverging, which may give us suboptimal rates of convergence. We do not necessarily require U_n to be a subset of U.

It turns out to be most appropriate to define a new 'index of discrepancies' for sieves.

Definition 6.2
Given the true density p_0 and another density p, we set

$$\rho_\alpha(p_0, p) = \begin{cases} \dfrac{1}{\alpha}\left[\mathbb{E}\left(\dfrac{p_0}{p}\right)^\alpha - 1\right] & \text{for } \alpha \in [-1,1]\backslash\{0\} \\ \\ \mathbb{E}_{p_0} \log\left(\dfrac{p_0}{p}\right) & \alpha = 0. \end{cases}$$

Remark 6.2
We have some significant special cases: $\alpha = -\frac{1}{2}$ gives us the squared Hellinger distance and $\alpha = 0$ is the Kullback-Leibler information.

The following theorem is also from Wong and Shen (1995) [18].

Theorem 6.1
Let $\mathcal{P}$ be a class of densities, and let $\mathcal{P}_n$ be a sequence of subsets of $\mathcal{P}$ such that

$$\inf_{p \in \mathcal{P}_n} \rho_\alpha(p_0, p) \leq \epsilon_n < \frac{1}{\alpha},$$

for some $\alpha \in (0, 1]$. With δ_n as the smallest value of δ satisfying (1) for each n, define

$$\delta_n^* = \begin{cases} \delta_n & \text{if } \epsilon_n < \frac{1}{4} c_1 \delta_n^2 \\ \\ (4\epsilon_n / c_1)^{\frac{1}{2}} & \text{otherwise.} \end{cases}$$

Then if $\hat{p}$ is an η_n-MLE, where $\eta_n < \frac{1}{2} c_1 (\delta_n^*)^2$,

$$\mathbb{P}\left(h(\hat{p}_n, p_0) \geq \delta_n\right) \leq C \exp\left(Dn(\delta_n^*)^2\right)$$

for some constants C, D.

Remark 6.3
Our inequality echoes Theorem 3.2, but the definition of δ_n^* means that our rate cannot be better than the square root of our sieve's convergence rate. C and D depend upon α, but this does not affect the convergence rate.

Wong and Shen also prove a slightly weaker result if we only have control for $\alpha = 0$ (the Kullback-Leibler information), but we will not have need of it.

Example 6.1
A natural choice of sieve, and one which allows us to choose the rate, is simply to use the functions constructed in order to define entropy. Let $\tau_n \to 0$ be a real sequence, and suppose $\mathcal{P}$ has finite ϵ-Hellinger entropy with bracketing for every $\epsilon > 0$. Then let $\mathcal{G}_n = \{(p_{n,i}^l, p_{n,i}^u),\ i \in I\}$ be τ_n-bracketing of order $H_B(\tau_n, \mathcal{P}, h)$ for $\mathcal{P}$ and define

$$\mathcal{P}_n = \left\{ \frac{p_{n,i}^u}{\int p_{n,i}^u} : i \in I \right\}.$$

Letting $\bar{p} = p^u / \int p^u$, where the pair (p^l, p^u) brackets p_0, notice first that

$$\begin{aligned}
\frac{p_0}{\bar{p}} &\leq \int p^u \\
&= h(0, p^u)^2 \\
&\leq \left(h(0,p) + h(p, p^u)\right)^2 \\
&\leq (1+\tau_n)^2 \\
&= 1 + O(\tau_n)
\end{aligned}$$

so[5] we obtain

$$\begin{aligned}
\rho_1(p_0, \bar{p}) &= \int \left(\frac{p_0}{\bar{p}} - 1 \right) p_0 \\
&= \int \left(\frac{p_0}{\bar{p}} - 1 - 1 + \frac{\bar{p}}{p_0} \right) p_0 \\
&= \int \frac{(p_0 - \bar{p})^2}{\bar{p}} \\
&= \int \left(\frac{p_0}{\bar{p}} - 1 \right)^2 \bar{p} \\
&= \int O(\tau_n^2) \bar{p} \\
&= O(\tau_n^2) \int \bar{p} \\
&= O(\tau_n^2),
\end{aligned}$$

since the convergence is uniform. So $\mathcal{P}_n$ constitutes a τ_n^2-sieve with respect to ρ_1; it is clear that for $\delta \leq \frac{1}{4}\tau_n$ we have $H_B(\delta, \mathcal{P}_n) = H_B(\tau_n, \mathcal{P})$, so substituting into (1) and choosing δ_n as the minimal value for which the inequality is satisfied, we get

$$\delta_n = O\left(\max(n^{-\frac{1}{2}} H_B^{1/2}(\tau_n, \mathcal{P}),\ \tau_n) \right).$$

[5]We defined the Hellinger distance between two *densities*, but it is perfectly valid on general non-negative functions, and one can easily check that it is still a pseudometric.

We can minimise this rate by choosing τ_n such that the two arguments in the maximum are of the same order:

$$H_B(\tau_n, \mathcal{P}) \asymp n\tau_n^2,$$

in which case $\delta_n = O(\tau_n)$.

Remark 6.4
This is a powerful result, and its most immediate use is to improve the suboptimal rate observed in Section 5. We have

$$\begin{aligned} H_B(\tau_n, \mathcal{P}) &\asymp n\tau_n^2 \\ \tau_n^{-\frac{1}{\alpha}} &\asymp n\tau_n^2 \\ \tau_n &\asymp n^{-\frac{\alpha}{2\alpha+1}}, \end{aligned}$$

so

$$\delta_n = O(n^{-\frac{\alpha}{2\alpha+1}}),$$

for all $\alpha \in (0, 1]$.

7 Conclusion

We have seen that maximum likelihood estimation can sensibly be extended to non-parametric contexts, and that although the results are less powerful than those achieved in parametric models, they are nevertheless very useful. We conclude with a glance at some possible generalisations of the work we have seen.

7.1 Penalised Maximum Likelihood

Let $X_1, X_2, \ldots$ be i.i.d. observations of a random variable with distribution function P. Suppose that we seek a compromise between the two goals of choosing a density which gives a large likelihood, and choosing one which is in some way 'natural'. We can formulate this trade off explicitly by using **penalised maximum likelihood**: consider trying to maximise

$$\frac{1}{n}\sum_{i=1}^{n} \log p(X_i) - \lambda_n I(p)^2,$$

for some *penalty function* I^2, and *smoothing parameter* λ_n. A common choice is

$$I(p)^2 = \int_{\mathcal{X}} p^{(m)}(x)^2 \, dx,$$

especially with $m = 2$. This is an intuitively sensible way of proceeding: we want our estimator to fit the data as well as possible, but we introduce a 'cost' for a good fit if we have to have a very 'rough' density. Penalised maximum likelihood comes under a much larger category of estimators called M-estimators.

7.2 M-Estimators

Definition 7.1
Let $X_1, X_2, \ldots$ be i.i.d. observations of a random variable with distribution function P and let $\{\gamma_\theta : \mathcal{X} \to \mathbb{R}, \theta \in \Theta\}$ be a collection of loss functions. Let θ_0 be the unique θ which minimises the value of

$$\int_{\mathcal{X}} \gamma_\theta \, dP$$

over $\theta \in \Theta$. An **M-estimator** of θ_0 based upon n-observations is given by any θ which minimises

$$\frac{1}{n}\sum_{i=1}^{n} \gamma_\theta(X_i)$$

over θ. We denote it by $\hat{\theta}_n$.

Remark 7.1
The M-estimator was first introduced by Huber in 1964 [8]: 'M' stands for minimum. If $\gamma_\theta = -l(\theta)$ we recover an estimator which is equivalent to the maximum likelihood estimator. Other common non-parametric techniques, such as least squares regression and penalised maximum likelihood, can also be incorporated under the M-estimator umbrella. Analytical solutions to an M-estimator problem are rare, and we usually proceed numerically: see, for example, [9].

8 Technical Proofs

This section is designed as an appendix to prove Theorems 3.1 and 3.2. Wong and Shen's proof of Theorem 3.1 works by considering lower truncated log-likelihood ratios, something we have not touched upon in the main section of the essay.

Definition 8.1
Given a log-likelihood ratio function

$$Z_p(Y) = \log \frac{p(Y)}{p_0(Y)},$$

and a constant $\tau > 0$, we define the **lower truncated log-likelihood ratio** as

$$\tilde{Z}_p = \max\left(Z_p, -\tau\right).$$

We write $\tilde{\mathcal{L}} = \{\tilde{Z}_p : p \in \mathcal{P}\}$ for the class of truncated log-likelihood ratios (from one observation).

This definition will guarantee that we have nice behaviour of $\tilde{Z}_p$, and since we are only interested in densities p which cause Z_p to be large, lower truncation will not prevent us from getting the results we want. Next we find a way to control the L^2 entropy with bracketing of the class of log-likelihood ratios, using the Hellinger entropy with bracketing of $\mathcal{P}$.

Lemma 8.1
We have

$$H_B(\delta, \tilde{\mathcal{L}}, \|\cdot\|_2) \leq H_B\left(\tfrac{1}{2}\delta e^{-\frac{\tau}{2}}, \mathcal{P}, h\right).$$

Proof
Let $p, q \in \mathcal{P}$, and let $\tilde{p} = \max(p,\, p_0 e^{-\tau})$; then $\tilde{Z}_p = \log(\tilde{p}/p_0)$. It is easy to show using the mean value theorem, that for x, y on some set A,

$$|\log x - \log y| \leq |x - y| \sup_{z \in A} z^{-1}.$$

Hence

$$\begin{aligned}
|\tilde{Z}_p - \tilde{Z}_q| &= 2\left|\log\left(\frac{\tilde{p}}{p_0}\right)^{\frac{1}{2}} - \log\left(\frac{\tilde{q}}{p_0}\right)^{\frac{1}{2}}\right| \\
&\leq 2\left(\frac{e^\tau}{p_0}\right)^{\frac{1}{2}} |\tilde{p}^{\frac{1}{2}} - \tilde{q}^{\frac{1}{2}}| \\
&\leq 2\left(\frac{e^\tau}{p_0}\right)^{\frac{1}{2}} |p^{\frac{1}{2}} - q^{\frac{1}{2}}|,
\end{aligned}$$

where the last inequality follows because we can never increase the distance between p and q by truncating them below at the same point. Thus

$$\mathbb{E}_{p_0}(\tilde{Z}_p - \tilde{Z}_q)^2 \leq 4e^\tau \int_\Omega |p^{\frac{1}{2}} - q^{\frac{1}{2}}|^2 \, d\mu.$$

which gives the result. □

The following Lemma, a form of Bernstein's Inequality, can be found in Bennett [2].

Lemma 8.2
Let $Y_1, \ldots, Y_n$ be i.i.d. random variables, and let f be a real valued function with $|f| \leq T < \infty$ and $\text{Var}\{f(Y_1)\} = \sigma^2$. Then for all $M > 0$,

$$\mathbb{P}\left(\nu_n(f) \geq M\right) \leq \exp\left[-\frac{M^2}{2(\sigma^2 + MT/3n^{1/2})}\right], \tag{4}$$

where $\nu_n(f) = n^{-\frac{1}{2}} \sum_{i=1}^{n} \left(f(Y_i) - \mathbb{E}f(Y_i)\right)$.

Remark 8.1
Set $\kappa = 2e^{-\frac{\tau}{2}}/(1 - e^{-\frac{\tau}{2}})^2$. It can be shown (see Lemmas 4, 5 and 6 in [18]) that

$$\mathbb{E}\tilde{Z}_p \leq (1 - \kappa)\, h(p, p_0)^2$$

and

$$\mathbb{P}\left(n^{-\frac{1}{2}} \sum_{i=1}^{n} \left(\tilde{Z}_p - \mathbb{E}\tilde{Z}_p\right) \geq t\right) \leq \exp\left[-\frac{t^2}{8(8c_0 \|p^{\frac{1}{2}} - p_0^{\frac{1}{2}}\|_2^2 + 2t/n^{\frac{1}{2}})}\right].$$

for some $c_0 > 0$.

The following result is Lemma 7 from Wong and Shen (1995) [18]; it allows us to connect the entropy of a set of densities $\mathcal{P}$ with the behaviour of log-likelihood ratios.[6]

Lemma 8.3
For any $0 < t \leq \sqrt{2}$, $0 < \epsilon < 1$ and $M > 0$, let

$$\psi(M, t^2, n) = \frac{M^2}{8(8c_0 t^2 + Mn^{-\frac{1}{2}})},$$

where c_0 is as before. Assume that

$$M \leq \frac{\epsilon n^{\frac{1}{2}} t^2}{4} \tag{5}$$

and

$$\int_{\epsilon M/(32\sqrt{n})}^{t} H_B^{1/2}\left(\tfrac{1}{2} u e^{-\frac{\tau}{2}}, \mathcal{P}\right)\, du \leq \frac{M\epsilon^{\frac{3}{2}}}{2^7(8c_0 + 1)}. \tag{6}$$

Then

$$\mathbb{P}^*\left(\sup_{p \in \mathcal{H}_t} \nu_n(\tilde{Z}_p) \geq M\right) \leq 3\exp(-(1 - \epsilon)\psi(M, t^2, n)) \tag{7}$$

where $\mathcal{H}_t = \{p \in \mathcal{P} : \|p^{1/2} - p_0^{1/2}\|_2 \leq t\}$, i.e. a Hellinger ball in $\mathcal{P}$ of radius t around p_0.

[6] I am indebted to Richard Samworth for his assistance in unravelling this proof.

Proof
We split the proof into sections to make it easier to follow.

Bounding The Bracketing Entropy

Firstly, we show that

$$H_B(t, \tilde{\mathcal{L}}, \|\cdot\|_2) \leq \frac{\epsilon}{4}\psi(M, t^2, n). \tag{8}$$

From Lemma 8.1, we have that $H_B(u, \tilde{\mathcal{L}}, \|\cdot\|_2) \leq H_B(\frac{1}{2}ue^{\frac{\tau}{2}}, \mathcal{P})$, so (6) gives us

$$\int_{\epsilon M/(32\sqrt{n})}^{t} H_B^{1/2}(u, \tilde{\mathcal{L}}, \|\cdot\|_2)\, du \leq \frac{M\epsilon^{\frac{3}{2}}}{2^7(8c_0+1)}.$$

Now, since H_B is a non-increasing function, the integrand achieves its minimum at t and so

$$\begin{aligned}\int_{\epsilon M/(32\sqrt{n})}^{t} H_B^{1/2}(u, \tilde{\mathcal{L}}, \|\cdot\|_2)\, du &\geq \left(t - \frac{\epsilon M}{32\sqrt{n}}\right) H_B^{1/2}(t, \tilde{\mathcal{L}}, \|\cdot\|_2) \\ &\geq \left(t - \frac{\epsilon^2 t^2}{2^7}\right) H_B^{1/2}(t, \tilde{\mathcal{L}}, \|\cdot\|_2),\end{aligned}$$

where the second inequality is achieved using (5). So combining with the previous inequality and squaring both sides gives

$$\begin{aligned}\frac{M^2\epsilon^3}{2^{14}(8c_0+1)^2} &\geq \left(t - \frac{\epsilon^2 t^2}{2^7}\right)^2 H_B(t, \tilde{\mathcal{L}}, \|\cdot\|_2) \\ &\geq t^2\left(1 - \frac{\epsilon^2 t}{2^7}\right)^2 H_B(t, \tilde{\mathcal{L}}, \|\cdot\|_2) \\ &\geq \frac{t^2}{2} H_B(t, \tilde{\mathcal{L}}, \|\cdot\|_2),\end{aligned}$$

where this last inequality uses the fact that $t \leq \sqrt{2}$. Hence

$$H_B(t, \tilde{\mathcal{L}}, \|\cdot\|_2) \leq \frac{M^2\epsilon^3}{2^{13}(64c_0^2+16c_0+1)t^2} \leq \frac{M^2\epsilon^3}{2^{13}(16c_0+1)t^2}.$$

Now, $\epsilon < 1$ and (5) give $M \leq \frac{1}{4}t^2\sqrt{n}$, so by introducing a negative term into the denominator, we get

$$\begin{aligned}H_B(t, \tilde{\mathcal{L}}, \|\cdot\|_2) &\leq \frac{\epsilon^3}{4}\frac{M^2}{2^{11}(16c_0+1)t^2 + 8M/\sqrt{n} - 2\epsilon t^2} \\ &\leq \frac{\epsilon}{4}\frac{M^2}{8(8c_0t^2 + M/\sqrt{n}) + (2^{11} - 2\epsilon)t^2} \\ &\leq \frac{\epsilon}{4}\frac{M^2}{8(8c_0t^2 + M/\sqrt{n})} \\ &= \frac{\epsilon}{4}\psi(M, t^2, n),\end{aligned}$$

which is (8).

Defining A Bracketing

We have showed that the L^2 bracketing entropy of $\tilde{\mathcal{L}}$ is finite, so for any $\delta_0 > \delta_1 > \ldots > \delta_N > 0$, there exist $\tilde{\mathcal{L}}_j$, $j = 0, \ldots, N$, with $\tilde{\mathcal{L}}_j$ a δ_j bracketing of $\tilde{\mathcal{L}}$ of order $N_B(\delta_j, \tilde{\mathcal{L}})$. Then for each $\tilde{Z}_p \in \tilde{\mathcal{L}}$, define $(f_j^L(\tilde{Z}_p), f_j^U(\tilde{Z}_p))$ as any pair in $\tilde{\mathcal{L}}_j$ which δ_j-brackets $\tilde{Z}_p$ in L^2. Now let

$$u_k(\tilde{Z}_p) = \min_{j \le k} f_j^U(\tilde{Z}_p) \quad \text{and} \quad l_k(\tilde{Z}_p) = \max_{j \le k} f_j^L(\tilde{Z}_p).$$

Then $(l_k(\tilde{Z}_p), u_k(\tilde{Z}_p))$ is sequence of pairs of functions in L^2 which δ_k-bracket $\tilde{Z}_p$. Notice that as $\tilde{Z}_p$ varies over $p \in \mathcal{P}$ there are at most $\prod_{j=1}^k |\tilde{\mathcal{L}}_j|$ distinct functions $u_k(\tilde{Z}_p)$.

Splitting The Inequality

Now let $\{a_0, \ldots, a_{N-1}\}$ be a sequence of strictly decreasing numbers to be chosen later. Define

$$\begin{aligned}
B_0 &= \{(u_0(\tilde{Z}_p) - l_0(\tilde{Z}_p) \ge a_0\} \\
B_k &= \{(u_k(\tilde{Z}_p) - l_k(\tilde{Z}_p) \ge a_k\} \cap \left(\bigcup_{j=0}^{k-1} B_j\right)^C \quad \text{for } k = 1, \ldots, N-1 \\
B_N &= \left(\bigcup_{j=0}^{N-1} B_j\right)^C .
\end{aligned}$$

We can see that the B_j form a partition of $\tilde{\mathcal{L}}$, so simple manipulation gives

$$\begin{aligned}
\tilde{Z}_p &= u_0 + \sum_{k=0}^{N} (u_k \mathbb{1}_{B_k} - u_0 \mathbb{1}_{B_k}) + \left(\tilde{Z}_p - \sum_{k=0}^{N} u_k \mathbb{1}_{B_k}\right) \\
&= u_0 + \sum_{k=1}^{N} \sum_{j=1}^{k} (u_j - u_{j-1}) \mathbb{1}_{B_k} + \sum_{k=0}^{N} (\tilde{Z}_p - u_k) \mathbb{1}_{B_k} \\
&= u_0 + \sum_{j=1}^{N} (u_j - u_{j-1}) \mathbb{1}_{\cup_{k \ge j} B_k} + \sum_{k=0}^{N} (\tilde{Z}_p - u_k) \mathbb{1}_{B_k} \\
&= u_0 + \underbrace{\sum_{j=1}^{N} (u_j - u_{j-1}) \mathbb{1}_{\cup_{k \ge j} B_k}}_{:=\alpha(\tilde{Z}_p)} + \underbrace{\sum_{k=0}^{N-1} (\tilde{Z}_p - u_k) \mathbb{1}_{B_k}}_{:=\beta(\tilde{Z}_p)} + \underbrace{(\tilde{Z}_p - u_N) \mathbb{1}_{B_N}}_{:=\gamma(\tilde{Z}_p)},
\end{aligned}$$

where we have removed the explicit dependence of the u_k upon $\tilde{Z}_p$ for convenience. So

$$\begin{aligned}
\mathbb{P}^* \left(\sup_{\mathcal{H}_t} \nu_n(\tilde{Z}_p) > M \right) &= \mathbb{P}^* \left(\sup_{\mathcal{H}_t} \nu_n \left(u_0 + \alpha(\tilde{Z}_p) + \beta(\tilde{Z}_p) + \gamma(\tilde{Z}_p) \right) > M \right) \\
&= \mathbb{P}^* \left(\sup_{\mathcal{H}_t} \left[\nu_n(u_0) + \nu_n(\alpha) + \nu_n(\beta) + \nu_n(\gamma) \right] > M \right) \\
&\le \mathbb{P}^* \left(\sup_{\mathcal{H}_t} \nu_n(u_0) + \sup_{\mathcal{H}_t} \nu_n(\alpha) + \sup_{\mathcal{H}_t} \nu_n(\beta) + \sup_{\mathcal{H}_t} \nu_n(\gamma) > M \right).
\end{aligned}$$

Defining Some Constants

We now define δ_j, N, a_j and another sequence η_j. We take

$$\begin{aligned}
\delta_0 &= \inf\left\{x : H_B(x, \tilde{\mathcal{L}}, \|\cdot\|_2) \leq \frac{\epsilon}{4}\psi(M, t^2, n)\right\} \\
\delta_{j+1} &= \frac{\epsilon M}{8n^{1/2}} \vee \sup\left\{x \leq \tfrac{1}{2}\delta_j : H_B(x, \tilde{\mathcal{L}}, \|\cdot\|_2) \geq 4H_B(\delta_j, \tilde{\mathcal{L}}, \|\cdot\|_2)\right\} \\
N &= \min\left\{n \in \mathbb{N} : \delta_n = \frac{\epsilon M}{8n^{1/2}}\right\},
\end{aligned}$$

so δ_j is a sequence which decreases by a factor of at least 2 each time, but stops when we reach $\frac{\epsilon M}{8\sqrt{n}}$. Notice that (8) implies that $\delta_0 \leq t$. Next let

$$\eta_j = \frac{4\delta_{j-1}}{\sqrt{\epsilon}}\left(\sum_{l\leq j} H_B(\delta_l, \tilde{\mathcal{L}}, \|\cdot\|_2)\right)^{\frac{1}{2}} \quad \text{and} \quad a_j = \frac{8\sqrt{n}\delta_{j-1}^2}{\eta_j}.$$

Then

$$\begin{aligned}
\sum_{j=1}^{N}\eta_j &= \frac{4}{\sqrt{\epsilon}}\sum_{j=1}^{N}\delta_{j-1}\left(\sum_{l\leq j} H_B(\delta_l, \tilde{\mathcal{L}}, \|\cdot\|_2)\right)^{\frac{1}{2}} \\
&\leq \frac{8}{\sqrt{\epsilon}}\sum_{j=1}^{N}\delta_{j-1}H_B^{1/2}(\delta_j, \tilde{\mathcal{L}}, \|\cdot\|_2) \\
&\leq \frac{2^6}{\sqrt{\epsilon}}\int_{\epsilon M/32\sqrt{n}}^{\delta_0} H_B^{1/2}(u, \tilde{\mathcal{L}}, \|\cdot\|_2)\, du
\end{aligned}$$

where this last inequality uses Lemma 3.1 in Alexander (1984) [1] – the proof is pure analysis, applies to all decreasing functions, and does not interest us. So by (6),

$$\begin{aligned}
\sum_{j=1}^{N}\eta_j &\leq \frac{M\epsilon^{\frac{3}{2}}}{2(8c_0+1)} \\
&\leq \frac{M\epsilon}{8}
\end{aligned}$$

where we assume $c_0 \geq 1$; in fact it can be chosen to be > 3, see [18] for details.

Splitting The Inequality, Part II

Now, returning to the main argument, let us define the following quantities:

$$\begin{aligned}
\mathbb{P}_1 &= \mathbb{P}^*\left(\sup_{\mathcal{H}_t}\nu_n(u_0) > \left(1 - \frac{3\epsilon}{8}\right)M\right) \\
\mathbb{P}_2 &= \mathbb{P}^*\left(\sup_{\mathcal{H}_t}\nu_n(\alpha) > \sum_{j=1}^{N}\eta_j\right) \\
\mathbb{P}_3 &= \mathbb{P}^*\left(\sup_{\mathcal{H}_t}\nu_n(\beta) > \tfrac{1}{2}\sum_{j=1}^{N}\eta_j\right) \\
\mathbb{P}_4 &= \mathbb{P}^*\left(\sup_{\mathcal{H}_t}\nu_n(\gamma) > \frac{\epsilon}{8}M\right).
\end{aligned}$$

Notice that from our previous inequality, the right hand sides of the equations in these four events sum to less than M. Thus

$$\mathbb{P}^*\left(\sup_{\mathcal{H}_t}\nu_n(\tilde{Z}_p) > M\right) \leq \mathbb{P}_1 + \mathbb{P}_2 + \mathbb{P}_3 + \mathbb{P}_4.$$

We proceed to bound each of these individually.

Bounding $\mathbb{P}_1$

$$\begin{aligned}
\mathbb{P}_1 &= \mathbb{P}^*\left(\sup_{\mathcal{H}_t}\nu_n(u_0) > \left(1-\frac{3\epsilon}{8}\right)M\right) \\
&\leq N_B(\delta_0, \tilde{\mathcal{L}}, \|\cdot\|_2)\sup_{\mathcal{H}_t}\left[\mathbb{P}^*\left(\nu_n(u_0) > \left(1-\frac{3\epsilon}{8}\right)M\right)\right]
\end{aligned}$$

because $u_0(\tilde{Z}_p)$ takes at most $N_B(\delta_0, \tilde{\mathcal{L}}, \|\cdot\|_2)$ distinct values as $\tilde{Z}_p$ varies. Using the second inequality mentioned in Remark 8.1 (remembering that $\mathcal{H}_t$ is a Hellinger ball of radius t) we get

$$\begin{aligned}
\mathbb{P}_1 &\leq \exp\left[H_B(\delta_0, \tilde{\mathcal{L}}, \|\cdot\|_2) - \psi\left(\left(1-\frac{3\epsilon}{8}\right)M, t^2, n\right)\right] \\
&\leq \exp\left[H_B(\delta_0, \tilde{\mathcal{L}}, \|\cdot\|_2) - \left(1-\frac{3\epsilon}{8}\right)^2\psi\left(M, t^2, n\right)\right] \\
&\leq \exp\left[\frac{\epsilon}{4}\psi\left(M, t^2, n\right) - \left(1-\frac{3\epsilon}{8}\right)^2\psi\left(M, t^2, n\right)\right] \\
&\leq \exp\left[-(1-\epsilon)\psi\left(M, t^2, n\right)\right]
\end{aligned}$$

Bounding $\mathbb{P}_2$

Next notice that for $j = 1, \ldots, N$

$$\begin{aligned}
\mathrm{Var}\left((u_j - u_{j-1})\mathbb{1}_{\cup_{k\geq j}B_k}\right) &\leq \mathbb{E}\left((u_j - u_{j-1})\mathbb{1}_{\cup_{k\geq j}B_k}\right)^2 \\
&\leq \mathbb{E}\left(u_{j-1} - l_{j-1}\right)^2 \\
&\leq \delta_{j-1}^2.
\end{aligned}$$

Using Lemma 8.2 we get

$$\mathbb{P}\left(\nu_n\left((u_j - u_{j-1})\mathbb{1}_{\cup_{k\geq j}B_k}\right) > \eta_j\right) \leq \exp\left(-\frac{\eta_j^2}{\delta_{j-1}^2 + a_j\eta_j/3n^{1/2}}\right)$$

by the definition of the sets B_j.

Using the definition of a_j,

$$\begin{aligned}
\frac{\eta_j^2}{2(\delta_{j-1}^2 + a_j\eta_j/3n^{1/2})} &= \frac{3\eta_j^2}{22\delta_{j-1}^2} \\
&= \frac{24}{11\epsilon}\sum_{l\leq j}H_B(\delta_l, \tilde{\mathcal{L}}, \|\cdot\|_2)
\end{aligned}$$

and so

$$\mathbb{P}\left(\nu_n\left((u_j - u_{j-1})\mathbb{1}_{\cup_{k\geq j}B_k}\right) > \eta_j\right) \leq \exp\left(-\frac{2}{\epsilon}\sum_{l\leq j} H_B(\delta_l, \tilde{\mathcal{L}}, \|\cdot\|_2)\right).$$

Thus

$$\begin{aligned}
\mathbb{P}_2 &= \mathbb{P}^*\left(\sup_{\mathcal{H}_t}\nu_n\left(\sum_{j=1}^{N}(u_j - u_{j-1})\mathbb{1}_{\cup_{k\geq j}}\right) > \sum_{j=1}^{N}\eta_j\right) \\
&\leq \mathbb{P}^*\left(\sum_{j=1}^{N}\sup_{\mathcal{H}_t}\nu_n\left((u_j - u_{j-1})\mathbb{1}_{\cup_{k\geq j}}\right) > \sum_{j=1}^{N}\eta_j\right) \\
&\leq \sum_{j=1}^{N}\mathbb{P}^*\left(\sup_{\mathcal{H}_t}\nu_n\left((u_j - u_{j-1})\mathbb{1}_{\cup_{k\geq j}}\right) > \eta_j\right) \\
&\leq \sum_{j=1}^{N}\prod_{r=0}^{j}|\tilde{\mathcal{L}}_r|\prod_{s=0}^{j-1}|\tilde{\mathcal{L}}_s|\sup_{\mathcal{H}_t}\mathbb{P}^*\left(\nu_n\left((u_j - u_{j-1})\mathbb{1}_{\cup_{k\geq j}}\right) > \eta_j\right).
\end{aligned}$$

So

$$\begin{aligned}
\mathbb{P}_2 &\leq \sum_{j=1}^{N}\exp\left(2\sum_{k\leq j}H_B(\delta_k, \tilde{\mathcal{L}}) - \frac{2}{\epsilon}\sum_{k\leq j}H_B(\delta_k, \tilde{\mathcal{L}})\right) \\
&\leq \sum_{j=1}^{N}\exp\left(-2\frac{1-\epsilon}{\epsilon}\sum_{k\leq j}H_B(\delta_k, \tilde{\mathcal{L}})\right) \\
&\leq \sum_{j=1}^{N}\exp\left(-2\frac{1-\epsilon}{\epsilon}4^j H_B(\delta_0, \tilde{\mathcal{L}})\right) \\
&\leq \sum_{j=1}^{N}\exp\left(-2(1-\epsilon)4^{j-1}\psi(M, t^2, n)\right) \\
&\leq 2\exp\left(-(1-\epsilon)\psi(M, t^2, n)\right),
\end{aligned}$$

where the last inequality certainly holds for sufficiently large n, since we are only really interested in the asymptotic behaviour.

Bounding $\mathbb{P}_3$

For $\mathbb{P}_3$ we have

$$\begin{aligned}
\nu_n\left((\tilde{Z}_p - u_j)\mathbb{1}_{B_j}\right) &= -n^{-\frac{1}{2}}\sum_{i=1}^{n}\left((u_j(Y_i) - \tilde{Z}_p(Y_i))\mathbb{1}_{B_j}\right) + n^{-\frac{1}{2}}\mathbb{E}\left((u_j - \tilde{Z}_p)\mathbb{1}_{B_j}\right) \\
&\leq n^{-\frac{1}{2}}\sup_{\tilde{\mathcal{L}}}\mathbb{E}\left((u_j - l_j)\mathbb{1}_{B_j}\right).
\end{aligned}$$

Now since $u_j - l_j \geq a_{j+1}$ on B_j, Markov's inequality gives

$$\mathbb{P}(B_j) \leq \frac{\mathbb{E}(u_j - l_j)^2}{a_{j+1}^2} \leq \frac{\delta_j^2}{a_{j+1}^2},$$

and using Cauchy-Schwarz

$$\begin{aligned}\sup_{\tilde{\mathcal{L}}} \mathbb{E}\left((u_j - l_j)\mathbb{1}_{B_j}\right) &\leq \sup_{\tilde{\mathcal{L}}} \left(\mathbb{E}(u_j - l_j)^2\, \mathbb{E}(\mathbb{1}_{B_j})^2\right)^{\frac{1}{2}} \\ &\leq \sup_{\tilde{\mathcal{L}}} \left(\mathbb{E}(u_j - l_j)^2\, \mathbb{P}(B_j)\right)^{\frac{1}{2}} \\ &\leq \frac{\delta_j^2}{a_{j+1}}.\end{aligned}$$

Thus

$$\sup_{\tilde{\mathcal{L}}} \nu_n\left((\tilde{Z}_p - u_j)\mathbb{1}_{B_j}\right) \leq \frac{\delta_j^2}{a_{j+1}} \leq \tfrac{1}{2}\eta_{j+1},$$

and so $\mathbb{P}_3 = 0$.

Bounding $\mathbb{P}_4$

Applying a similar argument we get

$$\begin{aligned}\nu_n\left((f - u_N)\mathbb{1}_{B_N}\right) &\leq n^{\frac{1}{2}}\mathbb{E}(u_N - l_N) \\ &\leq n^{\frac{1}{2}}\left[\mathbb{E}(u_N - l_N)^2\right]^{\frac{1}{2}} \\ &\leq n^{\frac{1}{2}}\delta_N \\ &\leq \frac{\epsilon M}{8},\end{aligned}$$

so $\mathbb{P}_4 = 0$ as well.

The End

Finally,

$$\mathbb{P}^*\left(\sup_{p\in\mathcal{H}_t} \nu_n(\tilde{Z}_p) \geq M\right) \leq \mathbb{P}_1 + \mathbb{P}_2 + \mathbb{P}_3 + \mathbb{P}_4 \leq 3\exp(-(1-\epsilon)\psi(M, t^2, n)).$$

□

Notice that the result is true for all $t > 0$, because the right hand side of the inequality is increasing in t, and two densities can never have a Hellinger distance of more than $\sqrt{2}$.

We now move to prove Theorems 3.1 and 3.2.

Proof of Theorem 3.1

For any $s > \delta$ and $\frac{1}{2} < \epsilon < 1$, we seek to apply Lemma 8.3 with $t = \sqrt{2}s$. If we choose $M = \frac{1}{2}\epsilon\sqrt{n}s^2$ then condition (5) is satisfied, and (6) becomes

$$\int_{\epsilon^2 s^2/64}^{\sqrt{2}s} H_B^{1/2}\left(\tfrac{1}{2}ue^{-\frac{\tau}{2}}, \mathcal{P}\right)\, du \leq \frac{\epsilon^{\frac{5}{2}}\sqrt{n}s^2}{2^8(8c_0+1)},$$

which is satisfied if

$$\int_{s^2/2^8}^{\sqrt{2}s} H_B^{1/2}\left(\tfrac{1}{2}ue^{-\frac{\tau}{2}}, \mathcal{P}\right)\, du \leq \frac{\epsilon^{\frac{5}{2}}\sqrt{n}s^2}{2^8(8c_0+1)}.$$

This follows from (1) if $c_3 = 2e^{\frac{\tau}{2}}$, $c_4 = \epsilon^{\frac{5}{2}}/2^8(8c_0+1)$ and by the fact that the integrand is non-increasing. Then by Lemma 8.3,

$$\mathbb{P}^*\left(\sup_{p\in\mathcal{H}_t}\nu_n(\tilde{Z}_p) \geq \tfrac{1}{2}\epsilon n^{\frac{1}{2}}s^2\right) \leq 3\exp\left[-\frac{(1-\epsilon)\epsilon^2 n s^2}{2^9c_0+16\epsilon}\right]. \tag{9}$$

Letting $\mathcal{A}(d) = \{p\in\mathcal{P} : d \leq h(p,p_0)^2 \leq 2d\}$, Remark 8.1 gives $\sup_{\mathcal{A}(s^2)}\mathbb{E}\tilde{Z}_p \leq -(1-\kappa)s^2$. Then

$$\begin{aligned}
&\left\{\sup_{\mathcal{A}(s^2)}\prod_{i=1}^n\frac{p(Y_i)}{p_0(Y_i)} \geq \exp\left(-ns^2\left(1-\kappa-\tfrac{1}{2}\epsilon\right)\right)\right\}\\
= &\left\{\sup_{\mathcal{A}(s^2)}\sum_{i=1}^n n^{-\frac{1}{2}}Z_p(Y_i) \geq -n^{\frac{1}{2}}s^2\left(1-\kappa-\tfrac{1}{2}\epsilon\right)\right\}\\
\subseteq &\left\{\sup_{\mathcal{A}(s^2)}\nu_n(\tilde{Z}_p) \geq \tfrac{1}{2}n^{\frac{1}{2}}s^2\epsilon\right\}.
\end{aligned}$$

So we use (9) to bound the probability of this first event, and we set $c_1 = 1-\kappa-\frac{1}{2}\epsilon$. Next, let L be the smallest integer such that $2^{L+1}\delta^2 \geq 2$, so that

$$\begin{aligned}
&\mathbb{P}^*\left(\sup_{\{p:\,\|p^{1/2}-p_0^{1/2}\|_2\geq\delta\}}\prod_{i=1}^n\frac{p(Y_i)}{p_0(Y_i)} \geq \exp\left(-nc_1s^2\right)\right)\\
&\qquad\leq \sum_{j=1}^L\mathbb{P}^*\left(\sup_{\mathcal{A}(\delta^2 2^j)}\prod_{i=1}^n\frac{p(Y_i)}{p_0(Y_i)} \geq \exp\left(-nc_1s^2\right)\right)\\
&\qquad\leq \sum_{j=1}^L 3\exp\left[-\frac{2^j(1-\epsilon)\epsilon^2 n\delta^2}{2^9c_0+16\epsilon}\right]\\
&\qquad\leq 4\exp\left[-\frac{(1-\epsilon)\epsilon^2 n\delta^2}{2^9c_0+16\epsilon}\right]
\end{aligned}$$

for sufficiently large n. Setting c_2 appropriately gives the result. □

Proof of Theorem 3.2
By the definition of an η_n-MLE, we have

$$\left\{\|\hat{p}^{1/2}-p_0^{1/2}\|_2 \geq \delta_n\right\} \subset \left\{\sup_{\{p:\,\|p^{1/2}-p_0^{1/2}\|_2\geq\delta_n\}}\prod_{i=1}^n\frac{p(Y_i)}{p_0(Y_i)} \geq \exp(-n\eta_n)\right\}.$$

Since $\exp(-n\eta_n) \geq \exp(-nc_2\delta_n)$, then applying Theorem 3.1 gives us

$$\mathbb{P}\left(\|\hat{p}^{1/2}-p_0^{1/2}\|_2 \geq \delta_n\right) \leq 4\exp(-c_2 n\delta_n^2).$$

□

References

[1] Alexander, K. (1984). Probability inequalities for empirical processes and a law of the iterated logarithm. *Ann. Probab. 12*, 1041–1067.

[2] Bennett, G. (1962). Probability inequalities for the sum of independent random variables. *J. Amer. Statist. Assoc. 57*, 33–45.

[3] Birgé, L. (1983). Approximation dans les espaces métriques et théorie de l'estimation. *Z. Wahrsch. Verw. Gebiete 65*, 181–237.

[4] Birgé, L. and Massart, P. (1993). Rates of convergence for minimum contrast estimators. *Probab. Theory Related Fields* **97**, 1-2, 113–150.

[5] Birgé, L. and Massart, P. (1998). Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli* **4**, 3, 329–375.

[6] Birman, M. and Solomjak, M. (1967). Piece-wise polynomial approximations of functions in the classes $\mathcal{W}_p^\alpha$. *Mathematics of the USSR Sbornik 73*, 295–317.

[7] Fisher, R. (1912). On an absolute criterion for fitting frequency curves. *Messenger of Mathematics* **41**, 1, 155–160.

[8] Huber, P. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* **35**, 1, 73–101.

[9] Huber, P. and Dutter, R. (1974). Numerical solution of robust regression problems. *Proc. Symp. Computational Statistics*, 165–172.

[10] Kolmogorov, A. and Tihomirov, V. (1959). ϵ-entropy and ϵ-capacity of sets in function spaces. *Uspekhi Mat. Nauk* **14**, 2, 3–86. [English translation in Amer. Math. Soc. Translations].

[11] Owen, A. (1949). Empirical likelihood ratio confidence regions. *Ann. Math. Statist.* **20**, 4, 595–601.

[12] Shen, X. and Wong, W. (1994). Convergence rate of sieve estimates. *Ann. Statist.* **22**, 2, 580–615.

[13] van de Geer, S. (1993). Hellinger-consistency of certain nonparametric maximum likelihood estimators. *Ann. Statist.* **21**, 1, 14–44.

[14] van de Geer, S. (2000). *Empirical Processes in M-Estimation.* Cambridge University Press.

[15] van der Vaart, A. and Wellner, J. (1996). *Weak Convergence and Empirical Processes.* Springer-Verlag.

[16] Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *Ann. Math. Statist.* **20**, 4, 595–601.

[17] Wong, W. and Severini, T. (1991). On maximum likelihood estimation in infinite dimensional parameter spaces. *Ann. Statist.* **19**, 2, 603–632.

[18] Wong, W. and Shen, X. (1995). Probability inequalities for likelihood ratios and convergence rates for sieve mles. *Ann. Statist.* **23**, 2, 339–362.