An exposition to the asymptotic equivalence of several nonparametric regression problems

Chun-Hei Lam
CID: 01351290
Supervisor: Alastair Young

Submitted in partial fulfillment of the requirements for the MSci degree in Mathematics (with a Year Abroad) of Imperial College London.

June 2022

Department of Mathematics
Imperial College London

Abstract

In this thesis, we provide an exposition on the concept of asymptotic equivalence between two statistical experiments due to Le Cam [1]. If two statistical experiments are closed to each other in Le Cam's sense, then given a decision rule of one experiment, one can find a decision rule of another experiment that performs almost as well as the original decision rule. Therefore, we can use the Le Cam distance to formalise the asymptotic equivalence among three common non-parametric problems in statistics: density estimation, Gaussian white noise and non-parametric regression.

This thesis provides scrutiny of some of the classical proofs for the asymptotic equivalence among non-parametric problems, including Brown and Low's proof [2] of asymptotic equivalence between non-parametric regression and Gaussian white noise and Nussbaum's proof [3] of asymptotic equivalence between density estimation and Gaussian white noise. The main ideas behind the proofs are to extract stochastic kernels between experiments and provide a uniform upper bound of the distance between the associated family of measures in the experiments. The paper also discusses theoretical and practical considerations of using the extracted stochastic kernel between experiments.

Acknowledgement

Soli Deo Gloria.

I thank my supervisor, Prof. Alastair Young, for continuous guidance and support. This topic emerged from our discussions regarding Gaussian sequence models, and I thank Prof. Young for encouraging me to develop further in this direction.

Since this thesis concludes my journey as an undergraduate for the MSci Mathematics (with a year abroad), I am extending my sincere gratitude to the Departments of Mathematics at Imperial College London and the Massachusetts Institute of Technology for providing high-quality education despite the challenges posed by the pandemics and preparing me for the upcoming journey as a researcher. In addition, I would like to thank my family and friends for their unwavering love and support.

Declaration

I pledge that this thesis and the work reported herein were our original work unless otherwise specified, and are submitted solely in partial fulfillment of the requirements of the MSci degree in Mathematics. Wherever works of others are involved, they have been cited clearly.

Chun-Hei Lam
14th June 2022

Contents

1 Overview . . . 3

2 The non-parametric statistical experiments . . . 4
2.1 Kernel Estimators and L^2 risks 4
2.2 L^∞ risks 6

3 Notion of Equivalence . . . 8
3.1 Measure-theoretic definition of a statistical experiments 8
3.2 Distances between statistical experiments 9
3.3 Bounds of Le Cam distance by other distance of measures 11
3.4 Example: Nonparametric regression and Gaussian White Noise 12
3.5 Another definition of Le Cam distance 13
3.6 Example: Density Estimation and Multinomial experiment 15
3.6.1 Step 1: Converting a multinomial experiment to a density estimation problem . . . 16
3.6.2 Step 2: Convert estimation of a discrete density to estimation of a continuous density by construction of stochastic kernel. . . . 17
3.6.3 Wrapping Up . . . 18

4 Local equivalence . . . 19
4.1 Setup: Constructing appropriate experiments 20
4.2 Application for the KMT Inequality, First Attempt 25
4.3 Divide and Conquer 26
4.3.1 Properties of λ_{f,f_0,A_j} . . . 28
4.3.2 Breaking up $\mathrm{Expt}_{1,n}(f_0)$ . . . 30
4.4 Poissonisation 35
4.5 Local equivalence for other experiments 37

5 Global Equivalence . . . 41
5.1 Proposal Estimator 41
5.2 Completing the proof by product experiments 42

6 Discussion . . . 45
6.1 Applicablity of stochastic kernels for constructing new decision rules 45
6.2 Necessity of assumptions of asymptotic equivalence 45
6.3 Conclusion and looking forward 46

7 Appendix . . . 47
7.1 Conditional Distribution 47
7.2 Distance between measures 47
7.3 Girsanov Theorem 50

1 Overview

The main duty of statistics is to reveal some characteristics of the underlying model of a data. In this paper, we are looking at the following three problems/experiment [1]:

- *Density estimation:* $\text{Expt}_{0,n}$ In this problem, we observe data simulated from an unknown probability density. The task is to estimate the underlying density function.
- *Non-parametric regression (Gaussian sequence)* $\text{Expt}_{\square,n}$: In this problem, you observe a signal (or function) contaminated by some noise, evaluated at *some* time-points. We assume for simplicity that the time-points are equally spaced. We would like to get an estimate of the original signal (function).
- *Gaussian white noise:* $\text{Expt}_{1,n}$ In this problem, you observe a signal (or function) contaminated by some noise, evaluated at *every* time-point. It is often difficult to work with the contaminated signals, so in the problem we try to get an estimate of the original signal (function).

Here n represents the size of an observation. A common feature of the above statistical problems is that they possess a family of kernel-type estimators that "smoothen" the noisy observations by convolving them with a kernel function. As we shall see, the kernel-type estimators in all the above problems have a similar performance when n is large. We will briefly discuss this in the next chapter.

The above fact motivates us to determine if the above problems are asymptotically equivalent, and if so, in what sense? We will address this in chapter 3 by defining Le Cam's notion of distance between two experiments, first introduced in his seminal paper on asymptotic sufficiency [1]. The main idea is the following: if one experiment is close with another in Le Cam's sense, then one can find a stochastic kernel which maps a decision rule of an experiment to another decision rule of another experiment having comparable performance. In the same chapter, we also discuss how we can use total variation (L^1) and Hellinger distances to provide an upper bound of the Le Cam distances before providing some simple illustrations.

We will then prove our main result in chapter 4 and 5 by closely following the arguments from Nussbaum's seminal paper [3]. The main result states that, if the target density $f(t)$ is continuous over $[0,1]$ ($f \in C^0[0,1]$) that satisfies some additional regularity conditions, the statistical problem of estimating $f(t)$ from data having same distribution is asymptotically equivalent with Gaussian white noise experiment with drift $\sqrt{f(t)}$ and Gaussian noise with intensity $1/2\sqrt{n}$. We divide the proof into two main steps. Chapter 4 is about the first step, that is, to establish asymptotic equivalence between these two experiments for the case when we restrict the parameter space to a small neighborhood of fixed density f_0. In particular, we will provide further justifications on different tricks utilised in the proof, which are not present in the original paper.

In chapter 5, we show that one can use a "proposal" estimator in either density estimation or Gaussian white noise experiment to identify a local likelihood for which there is a high probability that the target density would lie in. We will formalise the principle by looking at suitable product experiments.

We will end our discussion in chapter 6 by further commenting on applicability of asymptotic results, as well as summarise recent efforts in understanding the necessity of the regularity conditions in the above theorem.

[1] We will use the word "statistical problem" interchangeably with "statistical experiment."

2 The non-parametric statistical experiments

The aim of this chapter is to review the definition of suitable kernel-type estimators for each of the problems above, for which they enjoy similar bounds for its L^2 loss and L^∞ loss. We will not formally define the experiments using measure-theoretic definitions until the next chapter.

Let f be a continuous function on $[0,1]$ that satisfies the following smooth conditions: f is a α-Hölder continuous ($f \in C^\alpha[0,1]$) with $\alpha > 1/2$, that is bounded above by ϵ. That means there is a constant $M > 0$ such that for all $x, y \in [0,1]$, we have

$$|f(x) - f(y)| \leq M|x-y|^\alpha, \quad \alpha > 1/2, \tag{2.1}$$

and

$$f(x) \geq \epsilon \quad \forall x \in [0,1]. \tag{2.2}$$

We write $f \in \Sigma_{\alpha,M,\epsilon}$ for our convenience. Note in our case f is bounded above and below. If f_n is a sequence of kernel-type estimator such that the index n represents how many observations the kernel-type estimators are based on, then we can show that the L^2 loss of f_n is of order $n^{-\alpha/(2\alpha+1)}$, and the L^∞ loss of f_n is of order $(\ln n/n)^{\alpha/2\alpha+1}$. Some statisticians hence use this property to determine if an experiment is non-parametric. The discussions will be based around [4, 5].

2.1 Kernel Estimators and L^2 risks

Before delving into the discussion, let us define some notations. We begin by considering a kernel $K : \mathbb{R} \to \mathbb{R}$ such that it integrates to 1 over $\mathbb{R}$. For practical purposes, we assume that K has compact support and being continuous, so that K admits a maximum. We can then scale the kernel by the following:

$$K_h(x) = \frac{1}{h} K\left(\frac{x}{h}\right) \tag{2.3}$$

which also integrates to 1. We recall the the L^p norm of K over $\mathbb{R}$ is defined as

$$\|K\|^p_{L^p(\mathbb{R})} = \int_{\mathbb{R}} |K(u)|^p \, du \tag{2.4}$$

With this, we are ready to define different kernel-type estimators for different experiments.

Density estimation. Given observations $y := (y_1, ..., y_n) \in (\mathbb{R}, \mathscr{B}(\mathbb{R}))^{\otimes n}$, so that each entries are generated independently from the density $f(x)$. Given a kernel K, then one can define the *kernel density estimator* for this experiment:

$$\hat{f}_n^{(\mathsf{KDE})}(x) = \frac{1}{n} \sum_{i=1}^n K_h(y_i - x). \tag{2.5}$$

Non-parametric regression (Gaussian Sequence). Given observations $y := (y_1, ..., y_n)$ where $y_i = f(i/n) + \sigma Z_i$ with $Z_i \overset{\text{iid}}{\sim} \mathsf{N}(0,1)$. The *Nadaraya-Watson estimator* is defined as

$$\hat{f}_n^{(\mathsf{NW})}(x) = \frac{\sum_{i=1}^n K_h(i/n - x) y_i}{\sum_{i=1}^n K_h(i/n - x)} \tag{2.6}$$

It can be approximated by the following *Priestley-Chao* form:

$$\hat{f}_n^{(\mathsf{PC})}(x) = \frac{\sum_{i=1}^n K_h(i/n - x) y_i}{n} \tag{2.7}$$

It can be generalised to the following *local polynomial estimator*, with the following form presented in [4]

$$\hat{f}_n^{(\mathsf{LP})}(x) = (U(0))^\top \hat{\theta}_n(x), \tag{2.8}$$

where

$$U(u) = \left(1, u, ..., u^\ell/\ell!\right)^\top \tag{2.9}$$

$$\theta(x) = \left(f(x), f'(x)h, ..., f^{(\ell)}(x)h^\ell\right)^\top \tag{2.10}$$

$$\hat{\theta}_n(x) = \text{argmin}_\theta \sum_{i=1}^{n} \left(y_i - \theta^\top U\left(\frac{i/n - x}{h}\right)\right) hK_h(i/n - x) \tag{2.11}$$

all of the estimators above are appropriate kernel-type estimators for non-parametric regression problem.

Gaussian white noise. Given observation $y(t)$, which draws from a weak solution of the following SDE

$$dy(t) = f(t)\,dt + \sigma n^{-1/2} dW_t. \tag{2.12}$$

We impose the assumption that $f(t)$ is 1-periodic, i.e. $f(t+1) = f(t)$, as suggested in [5] for the convenience of computation. We may then form the periodised kernel

$$K_h^\circ(t) = \sum_{j\in\mathbb{Z}} K_h(t+j), \tag{2.13}$$

and define the *kernel estimator* as

$$\hat{f}_h^{(\mathsf{K})}(s) = \int_0^1 K_h^\circ(s-t)\,dy(t). \tag{2.14}$$

The above kernel-type estimators all enjoy the following properties related to bias and variance for sufficiently small h.

- Define the pointwise bias at x_0 as $b(x_0) = \mathbb{E}[\hat{f}(x_0)] - f(x_0)$ and $\sigma^2(x_0) = \mathbb{E}[\hat{f}_h(x_0) - \mathbb{E}[\hat{f}_h(x_0)]]$. Then one have

$$|b(x_0)| \lesssim |h|^\alpha,$$
$$|\sigma^2(x_0)| \lesssim \frac{1}{nh},$$

  where $\lesssim$ represents less than up to a constant, i.e. $a \lesssim b \iff \exists C > 0$ such that $a \le Cb$. Therefore there exists constant $C_1, C_2 > 0$ (which depends on α, M from assumption of parameter space and the kernels K) such that the mean squared error (MSE) satisfies

$$\sup_{f\in\Sigma_{\alpha,M,\epsilon}} \sup_{x\in\mathbb{R}} \mathbb{E}[|\hat{f}_n(x) - f(x)|^2] \le C_1 h^{2\alpha} + \frac{C_2}{nh}. \tag{2.15}$$

- Going through further calculations, we see that

$$\sup_{f\in\Sigma_{\alpha,M,\epsilon}} \mathbb{E}\left[\left\|\hat{f}_n - f\right\|^2_{L^2(0,1)}\right] \le C_1' h^{2\alpha} + \frac{C_2'}{nh}. \tag{2.16}$$

  where C_1', C_2' are constants that are not necessary equal to the above C_1 and C_2 respectively, and the mean integrated squared error (MISE) satisfies

$$\|f_n - f\|^2_{L^2(0,1)} = \int_0^1 |\hat{f}_n(x) - f(x)|^2\,dx. \tag{2.17}$$

For detailed proofs see chapter 1 of [4] and chapter 3 of [5]. We can therefore find the optimal h in terms of n such that the right hand side is minimised. We note the following lemma:

Lemma 2.1 — Bias-Variance Trade-off. Consider the function

$$f(h) = C_1 h^{2\alpha} + \frac{C_2}{nh}, \quad h > 0 \tag{2.18}$$

This function has a global minimiser at $h^* = (C_2/2\alpha C_1 n)^{1/(2\alpha+1)}$. The global minimiser of f is $Cn^{-2\alpha/(2\alpha+1)}$ for some constant C.

Proof. Note that $f'(h) = 2\alpha C_1 h^{2\alpha-1} - C_2/(nh^2)$, and $f'(h) = 0$ at $h = h^* = (C_2/2\alpha C_1 n)^{1/(2\alpha+1)}$, and if $h < h^*$ we have $f'(h) < 0$ and if $h > h^*$ we have $f'(h) > 0$. ■

We therefore have

Theorem 2.2

$$\sup_{f\in\Sigma_{\alpha,M,\epsilon}} \sup_{x\in\mathbb{R}} \mathbb{E}[|\hat{f}_n(x) - f(x)|^2] \lesssim n^{-2\alpha/(2\alpha+1)}, \tag{2.19}$$

$$\sup_{f\in\Sigma_{\alpha,M,\epsilon}} \mathbb{E}\left[\left\|\hat{f}_n - f\right\|^2_{L^2(0,1)}\right] \lesssim n^{-2\alpha/(2\alpha+1)}, \tag{2.20}$$

which is attained by choosing $h = O(n^{1/2\alpha+1})$.

We therefore see that the L^2 risk satisfies

$$\mathbb{E}\left[\left\|\hat{f}_n - f\right\|_{L^2[0,1]}\right] \lesssim n^{-\alpha/2\alpha+1} \tag{2.21}$$

The rate $\psi_n := n^{-\alpha/2\alpha+1} = o(n^{-1/4})$ is therefore viewed as a key feature of a *non-parametric* statistical problem. This rate is, in fact, *optimal* in the following sense as noted in [4]: define the squared L^2 minimax risk

$$\mathscr{R}^*_n := \inf_{\hat{f}_n} \sup_{f\in\Sigma_{\alpha,M,\epsilon}} \mathbb{E}\left[\|f_n - f\|^2_{L^2[0,1]}\right], \tag{2.22}$$

then there are constants $0 < c, C < \infty$ such that

$$c \le \liminf_{n\to\infty} \psi_n^{-2}\mathscr{R}^*_n \le \limsup_{n\to\infty} \psi_n^{-2}\mathscr{R}^*_n \le C. \tag{2.23}$$

2.2 L^∞ risks

Consider the L^∞ risk for the above non-parametric experiments [2]:

$$\mathbb{E}\left[\|f_n - f\|_{L^\infty(0,1)}\right], \quad \|f_n - f\|_{L^\infty(0,1)} = \|f_n - f\|_\infty = \sup_{x\in(0,1)} |f_n(x) - f(x)|$$

For all of the above non-parametric experiments, the L^∞ risk possess the rate $\psi_n := (\ln n/n)^{-\alpha/2\alpha+1}$. It is optimal in the following minimax sense: if we define the squared L^∞ minimax risk:

$$\mathscr{R}^*_n := \inf_{\hat{f}_n} \sup_{f\in\Sigma_{\alpha,M,\epsilon}} \mathbb{E}\left[\|f_n - f\|^2_{L^\infty[0,1]}\right], \tag{2.24}$$

then there are constants $0 < c, C < \infty$ such that

$$c \le \liminf_{n\to\infty} \psi_n^{-2}\mathscr{R}^*_n \le \limsup_{n\to\infty} \psi_n^{-2}\mathscr{R}^*_n \le C. \tag{2.25}$$

We omit the details of proof. If interested, please refer to [6] for the case of density estimation, and [**?**] for the case of Gaussian white noise model. [4] not only proves that ψ_n is optimal for the case

[2] Here the supremum can be replaced by essential supremum, since f_n and f are continuous.

of non-parametric regression, but also shows that the rate is attained by a suitable choice of local polynomial estimators.

One should finally note that the above result implies the following: for density estimation, non-parametric regression or Gaussian white noise, there is always a sequence of estimator $\hat{f}_n$ and constant $C > 0$ such that

$$\sup_{f \in \Sigma_{\alpha,M,\epsilon}} \mathbb{P}\left(\left\| \hat{f}_n - f \right\|_\infty \geq C\psi_n \right) \overset{n\to\infty}{\to} 0. \tag{2.26}$$

This forms a key step in proving asymptotic equivalence between density estimation and Gaussian white noise in chapter 5.

3 Notion of Equivalence

3.1 Measure-theoretic definition of a statistical experiments

We now formally introduce the measure-theoretic notion of statistical experiments, as introduced in [5]:

Definition 3.1 — Statistical Experiments/Problems. A statistical experiment Expt is a tuple $(\mathscr{Y}, \mathscr{F}, \{\mathbb{P}_\theta\}_{\theta\in\Theta}, \mathscr{A})$, where

- $(\mathscr{Y}, \mathscr{F})$ represents the measure space of observables,
- $\{\mathbb{P}_\theta\}_{\theta\in\Theta}$ is a family of probability measures on $(\mathscr{Y}, \mathscr{F})$, parametrised by θ in some space of parameters Θ, and
- $\mathscr{A}$ is the action space containing all actions an experimeter will take based on the observations.

Here we assume $(\mathscr{Y}, \mathscr{F})$ is a Polish space; that means, it is a complete separable metric space. Assume that $\mathscr{A}$ can also be metrized as a Polish space. The action space will depend on what decision the experimenter would like to make. If one wants to estimate the parameter, he can choose $\mathscr{A}$ to be a set containing Θ, and if one wants to test a hypothesis, he can choose $\mathscr{A} = \{0, 1\}$ representing whether the hypothesis is true or not. In our project we allow random decisions. To characterise this, we define the notion of stochastic kernel:

Definition 3.2 — Transitional, Stochastic kernel. Let $(\Omega_1, \mathscr{F}_1)$, $(\Omega_2, \mathscr{F}_2)$ be Polish spaces. A map $Q : \Omega_1 \times \mathscr{F}_2 \to [0, \infty]$ is a *transitional kernel* if

- when $A_2 \in \mathscr{F}_2$ is fixed, $Q(., A_2)$ is a $\mathscr{F}_1$-measurable function.
- when $\omega_1 \in \Omega_1$ is fixed, the set function $Q(\omega_1, .)$ is a measure on $(\Omega_2, \mathscr{F}_2)$

In addition if $\forall\omega_1 \in \Omega_1$, $Q(\omega_1, \Omega_2) = 1$ (i.e. $Q(\omega_1, .)$ is a probability measure), then Q is a *(Markov) stochastic kernel*.

Remark 3.3 A stochastic kernel induces the following two maps:

- a linear contraction from an essentially bounded function f on $(\Omega_2, \mathscr{F}_2)$ (i.e. $f \in L^\infty(\Omega_2)$) to another essentially bounded function on $(\Omega_1, \mathscr{F}_1)$ by mapping

$$f \in L^\infty(\Omega_2) \mapsto [Qf](\omega_1) := \int_{\Omega_2} f(\omega_2)\, Q(\omega_1, d\omega_2), \tag{3.1}$$

  which satisfies the following contraction property:

$$\|Qf\|_{L^\infty(\Omega_1)} \leq \|f\|_{L^\infty(\Omega_2)}\,. \tag{3.2}$$

- a map on Borel measures (measures that take finite value when evaluated on a compact set) of Ω_1 to Borel measures of Ω_2:

$$\mu \mapsto Q^\vee\mu(A) := \int_{\Omega_1} Q(\omega_1, A)\, \mu(d\omega_1). \tag{3.3}$$

Further introductions to the notion of stochastic kernel can be seen in [7, 8].

Then a decision rule is defined as followed:

Definition 3.4 — Decision Rule. Consider an experiment $(\mathscr{Y}, \mathscr{F}, \{\mathbb{P}_\theta\}_{\theta\in\Theta}, \mathscr{A})$. A decision rule $\delta(C|y) := \delta(y, C)$ is a stochastic kernel from $\mathscr{Y}$ to $\mathscr{A}$.

In particular, any point statistics as measurable maps from $\mathscr{Y}$ to $\mathscr{A}$ is a decision rule, since they are δ-measures on $\mathscr{A}$ when y is fixed.

Assuming loss function $L : \mathscr{A} \times \Theta \to [0, \infty)$, then the risk function of a decision rule is defined as

$$r_L(\delta, \theta) = \int_{\mathscr{Y}} \int_{\mathscr{A}} L(a, \theta)\, \delta(da|y)\, \mathbb{P}_\theta(dy) = \int_{\mathscr{Y}} \delta L(y, \theta) \mathbb{P}_\theta(dy). \tag{3.4}$$

If $|L|$ is integrable over the space $\mathscr{Y} \times \mathscr{A}$, e.g. L is essentially bounded, then by Fubini theorem, we may change of integration order, so that

$$r_L(\delta, \theta) = \int_{\mathscr{A}} L(a, \theta)\, \delta^\vee \mathbb{P}_\theta(da). \tag{3.5}$$

Remark 3.5 The choice of action space $\mathscr{A}$ does not affect our analysis below, so we may omit it when defining an experiment.

3.2 Distances between statistical experiments

To characterise the notion of asymptotic equivalence between two statistical problems, we have to introduce some notions of distance. Recall a statistical experiment Expt is a tuple $(\mathscr{Y}, \mathscr{F}, \{\mathbb{P}_\theta\}_{\theta\in\Theta}, \mathscr{A})$ as specified in the previous chapter. We can now formalise the notion of equivalence between two statistical experiments by defining the Le Cam's distance between two experiments:

Definition 3.6 — Le Cam's deficiency and distance. (Definition from (5, 9)). Consider two experiments

$$\mathsf{Expt}_i = (\mathscr{Y}_i, \mathscr{F}_i, \{\mathbb{P}_{i,\theta}\}_{\theta\in\Theta}, \mathscr{A}), i = 0, 1$$

with *same parameter space* Θ. We define the *Le Cam's deficiency* of Expt_0 with respect to Expt_1 as

$$\Delta_d(\mathsf{Expt}_0, \mathsf{Expt}_1) = \sup_{L:\|L\|_\infty \le 1} \left(\sup_{\delta_1} \inf_{\delta_0} \sup_{\theta\in\Theta} |r_{0,L}(\delta_0, \theta) - r_{1,L}(\delta_1, \theta)| \right) \tag{3.6}$$

where $\|L\|_\infty = \sup_{a\in\mathscr{A},\theta\in\Theta} L(a, \theta)$, and that $r_{i,L}$ is the risk function for the loss function L for experiment $i = 0, 1$. The *Le Cam's distance* between Expt_0 and Expt_1 is then defined as

$$\Delta(\mathsf{Expt}_0, \mathsf{Expt}_1) = \max(\Delta_d(\mathscr{P}_0, \mathscr{P}_1), \Delta_d(\mathscr{P}_1, \mathscr{P}_0)) \tag{3.7}$$

Let us look at the implication of $\Delta_d(\mathsf{Expt}_0, \mathsf{Expt}_1) < \epsilon$. Unfolding the first supremum, we see that ϵ is an upper bound of the set

$$U = \left\{ \epsilon \;\middle|\; \forall L, \|L\|_\infty \le 1,\ \sup_{\delta_1} \inf_{\delta_0} \sup_{\theta\in\Theta} |r_L(\delta_0, \theta) - r_L(\delta_1, \theta)| < \epsilon \right\} \tag{3.8}$$

Therefore, for all losses L with $\|L\|_\infty \le 1$, and for all decision rules δ_1 in experiment Expt_1, we have

$$\inf_{\delta_0} \sup_{\theta\in\Theta} |r_L(\delta_0, \theta) - r_L(\delta_1, \theta)| < \epsilon$$

which means that ϵ is not a lower bound of the set $\{\delta_0 \mid \sup_{\theta\in\Theta} |r_L(\delta_0, \theta) - r_L(\delta_1, \theta)|\}$, and hence there exists a decision rule δ_0 such that for all θ, one have

$$|r_L(\delta_0, \theta) - r_L(\delta_1, \theta)| < \epsilon \implies r_L(\delta_0, \theta) < \epsilon + r_L(\delta_1)$$

We can repeat the above argument to understand the implication of $\Delta_d(\mathsf{Expt}_1, \mathsf{Expt}_0) < \epsilon$. To conclude, if $\Delta(\mathsf{Expt}_0, \mathsf{Expt}_1) < \epsilon$, then

- for all decision rule δ_1 of experiment Expt_1, there is a decision rule δ_0 of experiment Expt_0 such that for all θ we have $r_{0,L}(\delta_0, \theta) < r_{1,L}(\delta_1, \theta) + \epsilon$
- for all decision rule δ_0 of experiment Expt_0, there is a decision rule δ_1 of experiment Expt_1 such that for all θ we have $r_{1,L}(\delta_1, \theta) < r_{0,L}(\delta_0, \theta) + \epsilon$.

We typically have $r_L \geq 0$, so if two experiments have small Le Cam distance, then for any estimator in one problem, we can find another estimator in another problem which has comparably small risk. From this, we can define the notion of asymptotic equivalence

Definition 3.7 — Asymptotic Equivalence. Assume for $i = 0, 1$, $\text{Expt}_{i,n} = (\mathscr{Y}_{i,n}, \mathscr{F}_{i,n}, \{\mathbb{P}_{i,\theta_n}\}, \mathscr{A}_n)$ are two sequences of experiments indexed by $\mathbb{Z}_{\geq 1}$, then we say they are *asymptotically equivalent* if $\Delta(\text{Expt}_0, \text{Expt}_1) \to 0$ as $n \to \infty$.

Let us prove an important property for Le Cam distance regarding sufficient statistic of an experiment, which is defined as followed:

Definition 3.8 — Sufficient statistic. Consider an experiment $\text{Expt} = (\mathscr{Y}, \mathscr{F}, \{\mathbb{P}_\theta\}_{\theta\in\Theta}, \mathscr{A})$. A statistic is a map $S : (\mathscr{Y}, \mathscr{F}) \to (E, \mathscr{E})$ such that S is $\mathscr{F}$-measurable. The statistic S is *sufficient* if there is a regular conditional distribution of Y given S is independent of θ, with Y being the identity map on $(\mathscr{Y}, \mathscr{F})$.

Sufficient statistic can be viewed as a change of variable that preserves information in an experiment. Let $Q(s, C)$ be a regular conditional distribution of Y given S, and define $\mathbb{Q}_\theta$ be the push forward measure $S^*\mathbb{P}_\theta(C) = \mathbb{P}_\theta(S \in C)$. Then almost surely we have for all $C \in \mathscr{F}$,

$$\int_E Q(s, C)\, \mathbb{Q}_\theta(ds) = \int_{\mathscr{Y}} Q(S(y), C)\mathbb{P}_\theta(dy) = \int_{\mathscr{Y}} \mathbb{E}[\mathbb{I}_C \mid S](y)\, \mathbb{P}_\theta(dy) = \int_{\mathscr{Y}} \mathbb{I}_C(y)\, \mathbb{P}_\theta(dy) = \mathbb{P}_\theta(C) \tag{3.9}$$

For more thorough discussion of regular conditional distribution, see section 7.1 for overview. We therefore have the following:

Lemma 3.9 — Equivalence by sufficiency. Consider the experiment $\text{Expt}_0 = (\mathscr{Y}, \mathscr{F}, \{\mathbb{P}_\theta\}_\theta, \mathscr{A})$, and let S be a sufficient statistic from $(\mathscr{Y}, \mathscr{F})$ to $(E, \mathscr{E})$, and $Q(s, C)$ is a regular conditional distribution as defined above. Consider the experiment $\text{Expt}_1 = (E, \mathscr{E}, \{\mathbb{Q}_\theta\}_{\theta\in\Theta}, \mathscr{A})$, then the Le Cam distance between these two experiments is zero.

This is intuitive, since transforming an experiment by a sufficient statistic does not lose information, so for a decision in one experiment, we can map it to another decision in another experiment with same risk. We formalise this argument by considering the regular conditional distribution:

Proof. Let $\delta_1(.|s)$ be a decision rule of Expt_1, then the stochastic kernel $\tilde{\delta}_1(.|S(y))$ is a decision rule of Expt_0. Moreover, for all bounded loss function L,

$$r_{0,L}(\tilde{\delta}_1, \theta) = \int_{\mathscr{Y}} \int_{\mathscr{A}} L(a, \theta)\, \tilde{\delta}_1(da|S(y))\, \mathbb{P}_\theta(dy) = \int_E \int_{\mathscr{A}} L(a, \theta)\, \delta_1(da|s)\, \mathbb{Q}_\theta(ds) = r_{1,L}(\delta_1, \theta)$$

On contrary, let $\delta_0(.|y)$ be a decision rule of Expt_0. Define the decision rule

$$\tilde{\delta}_0(.|s) = \int_{\mathscr{Y}} \delta_0(.|y) Q(s, dy) \tag{3.10}$$

Then it is a decision rule of Expt_1. Moreover, for all integrable loss function L

$$\begin{aligned} r_{1,L}(\tilde{\delta}_1, \theta) &= \int_E \int_{\mathscr{A}} L(a, \theta)\, \tilde{\delta}_0(da|s)\, \mathbb{Q}_\theta(ds) \\ &= \int_E \int_{\mathscr{A}} \int_{\mathscr{Y}} L(a, \theta)\, \delta_0(da|y)\, Q(s, dy)\, \mathbb{Q}_\theta(ds) \end{aligned}$$

$$\overset{\text{(Fubini)}}{=} \int_{\mathscr{Y}} \int_{\mathscr{A}} L(a,\theta)\delta_0(da|y) \int_E Q(s,dy)\mathbb{Q}_\theta(ds)$$
$$= \int_{\mathscr{Y}} \int_{\mathscr{A}} L(a,\theta)\delta_0(da|y)\mathbb{P}_\theta(ds) = r_{0,L}(\delta_0,\theta)$$

■

Example 3.10 Any bijective map $S : \mathscr{Y} \to \mathscr{E}$ is sufficient. To see this, we note that

$$\mathbb{P}_\theta(C) = \int_{\mathscr{Y}} \mathbb{I}_C(y)\,\mathbb{P}_\theta(dy) = \int_{\mathscr{Y}} \mathbb{I}_C(S^{-1}(s))\mathbb{Q}_\theta(ds) \tag{3.11}$$

so the regular conditional distribution $\mathbb{I}_C(S^{-1}(s))$ is independent of θ. ■

We will see an example of asymptotic equivalence later, but let us develop more tools that allow us to bound the Le Cam distance between experiments.

3.3 Bounds of Le Cam distance by other distance of measures

We further assume two experiments Expt_0 and Expt_1 have same space of observables $(\mathscr{Y}, \mathscr{F})$ and space of actions $\mathscr{A}$, and assume all measures $\mathbb{P}_{0,\theta}$, $\mathbb{P}_{1,\theta}$ are absolutely continuous with respect to a common measure ν. Writing $d\mathbb{P}_{0,\theta}/d\nu = p_{0,\theta}$ and $d\mathbb{P}_{1,\theta}/d\nu = p_{1,\theta}$, we can define the following additional notions of distance between experiments.

Definition 3.11 — More notions of distance between experiments. Under the above settings, define

- The L^1 distance $L^1(\mathsf{Expt}_0, \mathsf{Expt}_1) = \sup_{\theta\in\Theta} L^1(\mathbb{P}_{0,\theta}, \mathbb{P}_{1,\theta})$, where

$$L^1(\mathbb{P}_{0,\theta}, \mathbb{P}_{1,\theta}) = \sup_{A\in\mathscr{F}} |\mathbb{P}_{0,\theta}(A) - \mathbb{P}_{1,\theta}(A)| = \int_{\mathscr{Y}} |p_{0,\theta}(y) - p_{1,\theta}(y)|\nu(dy) \tag{3.12}$$

  is the L^1 (total variational) distance between the measures $\mathbb{P}_{0,\theta}$ and $\mathbb{P}_{1,\theta}$.

- The (Hellinger) distance $H(\mathsf{Expt}_0, \mathsf{Expt}_1) = \sup_{\theta\in\Theta} H^2(\mathbb{P}_{0,\theta}, \mathbb{P}_{1,\theta})$, where

$$H^2(\mathbb{P}_{0,\theta}, \mathbb{P}_{1,\theta}) := (H(\mathbb{P}_{0,\theta}, \mathbb{P}_{1,\theta}))^2 = \int_{\mathscr{Y}} \left|\sqrt{p_{0,\theta}(y)} - \sqrt{p_{1,\theta}(y)}\right|^2 \nu(dy) \tag{3.13}$$

  is the squared Hellinger distance between the measures $\mathbb{P}_{0,\theta}$ and $\mathbb{P}_{1,\theta}$.

We then have the following immediate inequality

Lemma 3.12 — L^1 bound of Le Cam distance.

$$\Delta(\mathsf{Expt}_0, \mathsf{Expt}_1) \le L^1(\mathsf{Expt}_0, \mathsf{Expt}_1) \tag{3.14}$$

Proof. We show that if δ is a decision rule in Expt_1 with risk $r_{0,L}(\delta,\theta)$ then it is also a decision rule in Expt_0 with risk at most $r_{0,L}(\delta,\theta) + L^1(\mathsf{Expt}_0, \mathsf{Expt}_1)$ (here $\|L\|_\infty \le 1$). To show this, note that

$$|r_{0,L}(\delta,\theta) - r_{1,L}(\delta,\theta)| = \left|\int_{\mathscr{Y}} \int_{\mathscr{A}} L(a,\theta)\delta(da|y)(p_{0,\theta}(y) - p_{1,\theta}(y))\,\nu(dy)\right|$$
$$\le \left|\int_{\mathscr{Y}} (p_{0,\theta}(y) - p_{1,\theta}(y))\,\nu(dy)\right|$$
$$\le \int_{\mathscr{Y}} |p_{0,\theta}(y) - p_{1,\theta}(y)|\,\nu(dy) \le L^1(\mathsf{Expt}_0, \mathsf{Expt}_1)$$

So we both have

$$r_{0,L}(\delta,\theta) \le r_{1,L}(\delta,\theta) + L^1(\mathsf{Expt}_0, \mathsf{Expt}_1)$$
$$r_{1,L}(\delta,\theta) \le r_{0,L}(\delta,\theta) + L^1(\mathsf{Expt}_0, \mathsf{Expt}_1)$$

so both $\Delta_d(\mathsf{Expt}_0, \mathsf{Expt}_1)$ and $\Delta_d(\mathsf{Expt}_1, \mathsf{Expt}_0)$ are bounded by $L^1(\mathsf{Expt}_0, \mathsf{Expt}_1)$, and that $\Delta(\mathsf{Expt}_0, \mathsf{Expt}_1) \le L^1(\mathsf{Expt}_0, \mathsf{Expt}_1)$ as desired. ■

We note a fundamental inequality between L^1 distance and Hellinger distance, for all measure $\mathbb{P}, \mathbb{Q}$ on $(\mathscr{Y}, \mathscr{F})$ we have

$$L^1(\mathbb{P}, \mathbb{Q}) \le 2H(\mathbb{P}, \mathbb{Q}) \tag{3.15}$$

and therefore

$$\Delta(\mathsf{Expt}_0, \mathsf{Expt}_1) \le 2H(\mathsf{Expt}_0, \mathsf{Expt}_1) \tag{3.16}$$

For proof please see lemma 7.5. We finally note a trivial result which shows that Le Cam distance is a well-defined distance

Lemma 3.13 Let Expt_i, $i = 0, 1, 2$ be three statistical experiments, then

$$\Delta(\mathsf{Expt}_0, \mathsf{Expt}_2) \le \Delta(\mathsf{Expt}_0, \mathsf{Expt}_1) + \Delta(\mathsf{Expt}_1, \mathsf{Expt}_2) \tag{3.17}$$

This shows that the Le Cam's notion of asymptotic equivalence is transitive.

3.4 Example: Nonparametric regression and Gaussian White Noise

With the above tools, we are ready to establish the asymptotic equivalence between nonparametric regression and Gaussian white noise. We compare the non-parametric regression problem $\mathsf{Expt}_{\square,n} = (\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n), (\{\mathbb{P}_{\square,f}\}_{f\in\Theta}), \mathscr{A})$, where $\mathbb{P}_{\square,f}$ is the distribution of the following Gaussian sequence

$$y_l = f(l/n) + \sigma Z_n,\ Z_n \overset{\text{iid}}{\sim} \mathsf{N}(0,1)\ \ell = 1, 2, ..., n; \tag{3.18}$$

against the Gaussian white noise experiment $\mathsf{Expt}_{1,n} = (C^0[0,1], \mathscr{B}(C^0[0,1]), (\{\mathbb{P}_{1,f}\}_{f\in\Theta}), \mathscr{A})$, where $\mathbb{P}_{1,f}$ is the distribution of the stochastic integral as a weak solution to the following SDE

$$dy(t) = f(t)\,dt + \sigma n^{-1/2} dW_t. \tag{3.19}$$

The action space can be taken as space of any functions on $[0,1]$, since it doesn't really matter in our discussion. We show that if f is regular enough, then these two (sequences of) experiments are asymptotically equivalent. Specifically,

Theorem 3.14 — Brown and Low Equivalence ((2), Theorem 4.1). Let $\Theta = C^\alpha[0,1]$, the collection of α-Hölder continuous functions on $[0,1]$ such that for all $f \in \Theta$, there is an $M \ge 0$ such that $|f(x) - f(y)| \le M|x-y|^\alpha$. If $\alpha > 1/2$, then $\mathsf{Expt}_{\square,n}$ and $\mathsf{Expt}_{1,n}$ are asymptotic equivalent.

Proof. We mainly follow the original proof from Brown and Low, see [5] and [2]. To show this, we introduce an intermediate experiment: $\overline{\mathsf{Expt}_{1,n}} = (C^0[0,1], \mathscr{B}(C^0[0,1]), \{\overline{\mathbb{P}}_{1,f}\}_{f\in\Theta}, \mathscr{A})$, where $\mathbb{P}_{1,f}$ is the distribution of the stochastic integral as a weak solution to the following SDE

$$d\bar{y}(t) = \bar{f}(t)\,dt + \sigma n^{-1/2} dW_t, \tag{3.20}$$

where $\bar{f}(t)$ represents a mapping from $f(t)$ to the piecewise constant function

$$\bar{f}(t) = \sum_{l=1}^{n} f(l/n)\mathbb{I}_{[(l-1)/n, l/n)}(t) + f(1)\mathbb{I}_{\{1\}}(t). \tag{3.21}$$

Notice that for all f, we have

$$L^1(\mathbb{P}_{1,f}, \overline{\mathbb{P}}_{1,f}) = 2\left(1 - 2\tilde{\Phi}(D_n/2)\right), \tag{3.22}$$

where $\tilde{\Phi}(z) = \int_z^\infty \exp(-t^2/2)/\sqrt{2\pi}$ is the hazard function of a standard Gaussian distribution, and that D_n satisfies:

$$D_n^2 = \frac{n}{2\sigma^2}\int_0^1 (f(t) - \bar{f}(t))^2\,dt = \frac{n}{2\sigma^2}\sum_{l=1}^n \int_{(l-1)/n}^{l/n} (f(t) - f((l-1)/n))^2\,dt. \tag{3.23}$$

For proof please see chapter 7.3. For all $t \in [(l-1)/n, l/n)$ we know that by α-Hölder continuity we know that $|f(t) - f((l-1)/n)| \le M(t-(l-1)/n)^\alpha$. The integral is therefore bounded by

$$D_n^2 \le \frac{n}{\sigma^2}\sum_{l=1}^n \int_{(l-1)/n}^{l/n} M^2(t-(l-1)/n)^{2\alpha}\,dt = \frac{n}{\sigma^2}\sum_{l=1}^n \int_0^{1/n} t^{2\alpha}\,dt = \frac{M^2}{(2\alpha+1)n^{2\alpha-1}} \overset{n\to\infty}{\to} 0, \tag{3.24}$$

provided that $\alpha > 1/2$. This shows that $\Delta(\mathsf{Expt}_{1,n}, \overline{\mathsf{Expt}_{1,n}}) \le L^1(\mathsf{Expt}_{1,n}, \overline{\mathsf{Expt}_{1,n}}) \to 0$ as $n \to \infty$.

To complete the proof, we note that $\overline{\mathsf{Expt}_{1,n}}$ is actually equivalent to $\mathsf{Expt}_{\square,n}$. This is because one can transform $\overline{\mathsf{Expt}_{1,n}}$ to $\mathsf{Expt}_{\square,n}$ by a sufficient statistic:

$$S(\bar{y}) = (\bar{y}(l/n) - \bar{y}((l-1)/n))_{l=1}^n \in \mathbb{R}^n. \tag{3.25}$$

The distribution of $S(\bar{y})$ is indeed a joint distribution of the multivariate normals, such that the entries are independent. Moreover, the mean for the l-th entry is $f((l-1)/n)$, and the variance of each entries is σ. This verifies that the experiment $\mathsf{Expt}_{\square,n}$ is really the experiment $\overline{\mathsf{Expt}_{1,n}}$ transformed by a sufficient statistics S, and that the two experiments have zero Le Cam's distance. This completes the proof. ■

Let S possesses a regular conditional distribution $Q(s, C)$, where $C \subseteq \mathscr{B}(C^0[0,1])$. Then if we have a decision rule in $\mathsf{Expt}_{\square,n}$, say $\delta_\square(.|y)$ with $y = (y_1, ..., y_n) \in [0,1]^n$, then the decision rule $\delta_\square(.|S(y_1, ..., y_n))$ on $C^0[0,1]$ has a comparable risk with $\mathsf{Expt}_{\square,n}$. If $\delta_\square$ is the Nadaraya-Watson kernel estimator

$$\delta_\square(.|y_1, ..., y_n) = \delta_{\hat{f}(y_1,...,y_n)}(.), \quad \hat{f}(s) = \frac{\sum_{\ell=1}^n K_h(s - l/n)y_\ell}{\sum_{\ell=1}^n K_h(s-l/n)} \tag{3.26}$$

then

$$\delta_\square(.|S(y(t))) = \delta_{\hat{f}(S(y(t)))}(.), \quad \hat{f}(s) = \frac{\sum_{\ell=1}^n K_h(s-l/n)\int_{(l-1)/n}^{l/n} dy(t)}{\sum_{\ell=1}^n K_h(s-l/n)} \tag{3.27}$$

which is a discretised version of the kernel estimator for the Gaussian white noise.

We can also try to go the other way round: if we have a decision rule in $\mathsf{Expt}_{1,n}$, say $\delta_1(.|y(t))$, the the following randomised decision rule

$$Q\delta_1(.|s) = \int_{C^0[0,1]} \delta_1(.|Y)Q(s, dy(t)) \tag{3.28}$$

has a comparable risk with δ_1. However, this estimator is not applicable in daily life, since it involves an integration over the space of $C^0[0,1]$ with respect to a kernel, which is again not easy to be specified.

3.5 Another definition of Le Cam distance

We finally note an equivalent definition of the Le Cam distance:

Proposition 3.15 — Equivalent definition of Le Cam's deficiency and distance ((1), Theorem 3, (10), Theorem 2.7). Consider two experiments $\mathsf{Expt}_0 = (\mathscr{Y}_0, \mathscr{F}_1, \{\mathbb{P}_{0,\theta}\}_{\theta\in\Theta}, \mathscr{A})$ and $\mathsf{Expt}_1 = (\mathscr{Y}_1, \mathscr{F}_1, \{\mathbb{P}_{1,\theta}\}_{\theta\in\Theta}, \mathscr{A})$ with *same parameter space* Θ and decision space $\mathscr{A}$. Then the *Le Cam's*

deficiency of Expt_0 is equal to

$$\Delta_d(\mathsf{Expt}_0, \mathsf{Expt}_1) = \inf_M \sup_{\theta\in\Theta} L^1(M^\vee \mathbb{P}_{0,\theta}, \mathbb{P}_{1,\theta}) \tag{3.29}$$

where $M : \mathscr{Y}_0 \times \mathscr{F}_1 \to [0,1]$ is a stochastic kernel from $(\mathscr{Y}_0, \mathscr{F}_0)$ to $(\mathscr{Y}_1, \mathscr{F}_1)$. (Recall the definition of $M^\vee$ in equation (3.3).)

Proof. (RHS $\geq$ LHS). Assume there is a stochastic kernel M such that $L^1(M^\vee \mathbb{P}_{0,\theta}, \mathbb{P}_{1,\theta}) < \epsilon$. Let's say δ_1 is a decision rule of Expt_1. Consider the decision rule

$$\delta_0(C|y_0) = M\delta_1(C|y_0) = \int_{\mathscr{Y}_1} \delta_1(C|y_1)\, M(y_0, dy_1) \tag{3.30}$$

Choose L such that $\|L\|_\infty \leq 1$. Then we know that

$$\begin{aligned} r_0(\delta_0, \theta) &= \int_{\mathscr{Y}_0}\int_{\mathscr{A}} L(a,\theta)\, M\delta_1(da|y_0)\, \mathbb{P}_{0,\theta}(dy_0) \\ &= \int_{\mathscr{Y}_0}\int_{\mathscr{A}} L(a,\theta) \int_{\mathscr{Y}_1} \delta_1(da|y_1)\, M(y_0, dy_1)\, \mathbb{P}_{0,\theta}(dy_0) \\ &= \int_{\mathscr{Y}_1}\int_{\mathscr{A}} L(a,\theta)\, \delta_1(da|y_1) \int_{\mathscr{Y}_0} M(y_0, dy_1)\, \mathbb{P}_{0,\theta}(dy_0) \\ &= \int_{\mathscr{Y}_1} \delta_1 L(y_1,\theta)\, M^\vee \mathbb{P}_{0,\theta}(dy_0), \end{aligned}$$

with the exchange of integral being justified by the boundedness of L and the contraction property of δ_1 as a stochastic kernel, i.e. $\|\delta_1 L\|_\infty \leq \|L\|_\infty \leq 1$. As a result, it is clear that

$$|r_0(\delta_0,\theta) - r_1(\delta_1,\theta)| \leq L^1(M^\vee \mathbb{P}_{0,\theta}, \mathbb{P}_{1,\theta}) < \epsilon, \tag{3.31}$$

and hence RHS $\geq$ LHS.

(LHS $\geq$ RHS). We only give a sketch. Assume $\Delta_d(\mathsf{Expt}_0, \mathsf{Expt}_1) \leq \epsilon$. Assume, for simplicity, that $\mathscr{A} = \mathscr{Y}_1$. Consider the identity decision rule $\delta_1(.|y_1) = \delta_{y_1}(.)$ in Expt_1, where $\delta_{y_1}(A) := \mathbb{I}_A(y_1)$ (for all $A \in \mathscr{Y}_1$) is the δ-measure on $(\mathscr{Y}_1, \mathscr{A}_1)$. By definition of Δ_d, there is a decision rule $\delta_0(.|y_0)$ in Expt_0 such that for all loss function $L(y_1,\theta)$ with $\|L\|_\infty \leq 1$, one have

$$\sup_\theta |r_L(\delta_0,\theta) - r_L(\delta_1,\theta)| \leq \epsilon. \tag{3.32}$$

Let $L(a,\theta) = \mathbb{I}_B(a)$ for any fixed $B \in \mathscr{F}_1$. Then clearly $\|L\|_\infty \leq 1$, and by (3.5), one have

$$r_{0,L}(\delta_0,\theta) = \int_{\mathscr{A}} \mathbb{I}_B(a)\delta_0^\vee \mathbb{P}_{0,\theta}(dy_1) = \delta_0^\vee \mathbb{P}_{0,\theta}(B). \tag{3.33}$$

In addition,

$$\begin{aligned} r_{1,L}(\delta_1,\theta) &= \int_{\mathscr{Y}}\int_{\mathscr{A}} \mathbb{I}_B(a)\delta_1(da|y)\mathbb{P}_{1,\theta}(dy) \\ &= \int_{\mathscr{Y}} \delta_1(B|y)\mathbb{P}_{1,\theta}(dy) \\ &= \int_{\mathscr{Y}} \mathbb{I}_B(y)\mathbb{P}_{1,\theta}(dy) = \mathbb{P}_{1,\theta}(B). \end{aligned}$$

We therefore see that, for any $B \in \mathscr{F}_1$,

$$\sup_\theta |\delta_0^\vee \mathbb{P}_{0,\theta}(B) - \mathbb{P}_{1,\theta}(B)| \leq \epsilon. \tag{3.34}$$

and therefore $\inf_M \sup_{\theta\in\Theta} L^1(M^\vee \mathbb{P}_{0,\theta}, \mathbb{P}_{1,\theta}) \leq L^1(\delta_0^\vee \mathbb{P}_{0,\theta}, \mathbb{P}_{1,\theta}) \leq \epsilon$. This proves LHS $\geq$ RHS. As a technical note, using a suitable limiting argument one can remove the assumption that $\mathscr{A}$ equals to $\mathscr{Y}_1$ by noting that there is a measurable bijective isomorphism between these two Polish space, and we can incorporate this isomorphism in the loss function. ■

The main message of this proposition is as followed: if we know that $\Delta(\mathsf{Expt}_0, \mathsf{Expt}_1)$ is small, then we can extract a stochastic kernel $M := \delta_0(.|y_0)$ satisfying (3.32), and it serves as a "random map" from δ_1 in Expt_1 to $\delta_0 = M\delta_1$ in Expt_0 which performs almost as good as δ_1. In particular, we can provide an alternative proof to lemma 3.12, noting that

$$\Delta_d(\mathsf{Expt}_0, \mathsf{Expt}_1) \leq L^1(M^\vee \mathbb{P}_{0,\theta}, \mathbb{P}_{1,\theta}) = L^1(\mathbb{P}_{0,\theta}, \mathbb{P}_{1,\theta}), \quad M(y_0, A) = \delta_{y_0}(A). \tag{3.35}$$

Remark 3.16 Under the setting of lemma 3.12, if $S(y)$ is an sufficient estimator of Expt_0 with regular conditional distribution Q, then we have $Q^\vee \mathbb{P}_{0,\theta} = \mathbb{P}_{1,\theta}$. Therefore, the notion of Le Cam distance actually provides a notion of asymptotic sufficiency of a sequence of estimators, as commented by Le Cam in his original paper [1]. It is however note that just having $Q^\vee \mathbb{P}_{0,\theta} = \mathbb{P}_{1,\theta}$ is NOT a sufficient condition to show that S is an sufficient estimator.

The proposition provides a main recipe of proving asymptotic equivalence by constructing appropriate stochastic kernels that satisfies (3.29). Of course, one can also prove asymptotic equivalence by bounding the L^1 or Hellinger distances; but we won't be able to use the stochastic kernel to transform any decision rules from one experiment to the other for practical uses.

3.6 Example: Density Estimation and Multinomial experiment

We apply the above principle to establish the asymptotic equivalence between density estimation and multinomial experiment with a tighter restriction on the smoothness of the underlying density. We consider the parameter space:

$$\Sigma = \Sigma_{\gamma,M,\epsilon} = \left\{ f \in C^1[0,1] \,|\, |f'(x) - f'(y)| \leq M|x-y|^\gamma, \quad f \geq \epsilon \right\}, \quad \gamma \in (0,1], \epsilon > 0; \tag{3.36}$$

and define the following experiments:

$$\mathsf{Expt}_{0,n} = ([0,1]^n, \mathscr{B}([0,1]^n), \left\{ d\mathbb{P}_{0,n,f} = (f\, d\mathsf{Leb})^{\otimes n} \right\}_{f\in\Sigma}), \tag{3.37}$$

$$\mathsf{Expt}_{\mathsf{M},n} = (\mathscr{Y}, 2^{\mathscr{Y}}, \{\mathbb{P}_{1,n,f} \sim \mathsf{Multinomial}(m_n, \theta(f))\}_{f\in\Sigma}), \tag{3.38}$$

where Leb represents the Lebesgue measure on $[0,1]$, $m := m_n$ depends on n, $\mathscr{Y}$ is the lattice $\{z \in [n]^m \mid \sum z_j = n\}$ with $[n] = \{0, ..., n\}$, $2^{\mathscr{Y}}$ is its power set and $\theta(f) \in [0,1]^m$ with

$$\theta_j := \theta_j(f) = \int_{D_j} f(x)\, dx, \quad D_j = [(j-1)/m, j/m), j = 1, 2, ..., m-1. \tag{3.39}$$

Here we follow Carter's arguments in section 8 of [11], paraphrased by Mariucci in section 5 of [10], putting emphasis on how one can use the constructed Markov kernel to construct a "corresponding" decision rule in $\mathsf{Expt}_{\mathsf{M},n}$ from another decision rule from experiment $\mathsf{Expt}_{0,n}$. To start off, recall the probability mass distribution for a multinomial distribution:

$$\mathbb{P}_{1,n,f}(\{z\}) = \mathbb{P}_{1,n,f}(\{z_1, z_2, ..., z_m\}) = \frac{n!}{z_1! z_2! ... z_m!} \theta_1^{z_1} ... \theta_m^{z_m} \tag{3.40}$$

The experiment $\mathsf{Expt}_{\mathsf{M},n}$ is derived from applying the map $S : [0,1]^n \to \mathscr{Y}$ with the entries $[S(y)]_j = \#\{i \in [n] \,|\, y_i \in D_j\}$ and $\#$ denote the counting measure that counts the number of elements in a set, so $\Delta_d(\mathsf{Expt}_{0,n}, \mathsf{Expt}_{\mathsf{M},n}) = 0$. In particular, if S induces a stochastic kernel M, then for any estimators $\delta_{\mathsf{M}}(.|s)$ in $\mathsf{Expt}_{\mathsf{M},n}$, the randomised estimator $M\delta_0(.|y)$ performs as well as δ_{M}. This is not very useful in practice, and one would like to go the other way. Given S is not a sufficient estimator, let us control $\Delta_d(\mathsf{Expt}_{\mathsf{M},n}, \mathsf{Expt}_{0,n})$ by some other careful analysis. The proof is divided into three main steps.

3.6.1 Step 1: Converting a multinomial experiment to a density estimation problem

We introduce the following experiments, parametrised by a probability vector $\theta \in \Delta_n$ with $\Delta_n := \{\theta \in [0,1]^n \mid \sum \theta_i = 1\}$:

$$\mathsf{Expt}^{(2)}_{0,n} = \left([0,1]^n, \mathscr{B}([0,1]^n), \left\{\mathbb{P}^{(2)}_{0,n,\theta} := \left(\mathbb{P}^{(2)}_{0,\theta}\right)^{\otimes n}\right\}_{\theta\in\Delta_n}\right), \tag{3.41}$$

where

$$\mathbb{P}^{(2)}_{0,\theta}(A) = \sum_{i=1}^{m} \theta_j \delta_{y^*_j}(A), \quad y^*_j = \frac{j-1/2}{m}, \quad A \in \mathscr{B}([0,1]) \tag{3.42}$$

is the distribution of a random variable with atoms in the midpoints of intervals D_j, here written as y^*_j. We compare this experiment alongside with the above multinomial experiment (but with $\theta(f)$ replaced by a generic probability vector θ)

$$\mathsf{Expt}_{\mathsf{M},n} = (\mathscr{Y}, 2^{\mathscr{Y}}, \{\mathbb{P}_{1,n,\theta} \sim \mathsf{Multinomial}(m_n, \theta))_{\theta\in\Delta_n}\} \tag{3.43}$$

In this setting, the statistic S as defined above is actually a sufficient statistic. This could be seen by noting that the support of $\mathbb{P}^{(2)}_{0,n,\theta}$ is actually the set $\{y^*_1, y^*_2, ..., y^*_m\}^n$, and has probability mass function

$$\mathbb{P}^{(2)}_{0,n,\theta}(y_1, ..., y_n) = \prod_{j=1}^{m} \theta_j^{\#\{i\in[n]\,|\,y_i\in D_j\}} = \prod_{j=1}^{m} \theta_j^{[S(y)]_j} \tag{3.44}$$

so we can compute the regular conditional distribution given this sufficient statistic, which satisfies

$$Q(z,\{y\}) = \frac{\mathbb{P}_{0,n,\theta}(\{y\}\cap\{S(y)=z\})}{\sum_{x:S(x)=z}\mathbb{P}_{0,n,\theta}(\{x\}\cap\{S(x)=z\})} = \frac{z_1!z_2!...z_m!}{n!}\mathbb{I}_{S(y)=z}(y)$$

and is independent of θ. So S is a sufficient statistic, and that the above experiments are equivalent. Of course, the equivalence is preserved when we restrict θ to probability of the form $\theta(f)$ for $f \in \Sigma$ as defined above. This indicates that if one has a decision rule $\delta(.|y)$ in experiment $\mathsf{Expt}^{(2)}_{0,n}$, then the randomised decision rule

$$Q\delta(.|z) = \int_{\mathscr{Y}} \delta(.|y)Q(z,dy) = \frac{z_1!z_2!...z_m!}{n!} \sum_{\{y\in\{S(y)=z\}\}} \delta(.|y) \tag{3.45}$$

performs as well as the original decision rule $\delta(.|y)$.

Remark 3.17

- Notice that one can prove that S is sufficient by using the Neymann factorisation criterion (see theorem 3.6 of [12]), noting that the probability mass function only depends on θ and $S(y)$. However, here we have computed the actual regular conditional distribution.
- Since we have n large, computing the multinomial coefficient can be extremely inefficient. Therefore, in practice we approximate the sum by using the following Monte Carlo scheme:

$$\frac{z_1!z_2!...z_m!}{n!} \sum_{\{y\in\{S(y)=z\}\}} \delta(.|y) \approx \frac{1}{N}\sum_{k=1}^{N} \delta(.|Y_k) \tag{3.46}$$

  where $Y_1, Y_2, ..., Y_N$ are randomly selected element from the set $\{S(y) = z\}$, and N very large.

3.6.2 Step 2: Convert estimation of a discrete density to estimation of a continuous density by construction of stochastic kernel.

Consider the family of functions $V_1, ..., V_m$, where

$$\begin{aligned}
V_1(x) &= m\mathbb{I}_{[0,y_1^*)}(x) + (m - m^2|x - y_1^*|)\mathbb{I}_{[y_1^*,y_2^*)}(x) \\
V_j(x) &= m^2|x - y_{j-1}^*|\mathbb{I}_{[y_{j-1}^*,y_j^*)}(x) + (m - m^2|x - y_j^*|)\mathbb{I}_{[y_j^*,y_{j+1})}(x), \quad j = 2, ..., m-1 \\
V_m(x) &= m^2|x - y_{m-1}^*|\mathbb{I}_{[y_{m-1}^*,y_m^*)}(x) + m\mathbb{I}_{[y_m^*,1]}(x)
\end{aligned}$$

and consider an approximation of f, denoted as $\hat{f}$:

$$\hat{f}_m(x) = \sum_{j=1}^{m} V_j(x)[\theta(f)]_j \tag{3.47}$$

Consider the experiment

$$\mathsf{Expt}_{0,n}^{(1)} = \left([0,1]^n, \mathscr{B}[0,1]^n, \left\{\mathbb{P}_{0,n,f}^{(1)} := \left(\mathbb{P}_{0,f}^{(1)}\right)^{\otimes n}\right\}_{f\in\Sigma}\right) \tag{3.48}$$

with

$$\mathbb{P}_{0,f}^{(1)}(A) = \int_A \hat{f}_m(x)\,dx \tag{3.49}$$

Define the stochastic kernel from $[0,1]$ to

$$M(y, A) = \sum_{j=1}^{m} \mathbb{I}_{\{y_j^*\}}(y) \int_A V_j(x)\,dx, \tag{3.50}$$

then we know that

$$\begin{aligned}
M^{\vee}\mathbb{P}_{0,f}^{(2)}(A) &= \int_{\{y_1^*,...,y_m^*\}} M(y, A)\,d\mathbb{P}_{0,f}^{(2)} \\
&= \sum_{j=1}^{m} [\theta(f)]_j M(y_j^*, A) \\
&= \sum_{j=1}^{m} [\theta(f)]_j \int_A V_j(x)\,dx = \int_A \hat{f}_m(x)\,dx = \mathbb{P}_{0,f}^{(1)}(A).
\end{aligned}$$

Define $M^{\otimes n}$ be the product stochastic kernel on $\{y_1^*, ..., y_m^*\}^n \times \mathscr{B}[0,1]^n$, such that for each product set $A_1 \times ... \times A_n$ with $A_i \in \mathscr{B}[0,1]$, one have

$$M^{\otimes n}((y_1, ..., y_n), A_1 \times ... \times A_n) = \prod_{i=1}^{n} M(y_i, A_i), \tag{3.51}$$

then

$$\begin{aligned}
(M^{\otimes n})^{\vee}\mathbb{P}_{0,n,f}^{(2)}(A_1 \times ... \times A_n) &= \int_{\{y_1^*,...,y_m^*\}^n} \prod_{i=1}^{n} M(y_i, A_i)\,\left(\mathbb{P}_{0,f}^{(2)}\right)^{\otimes n}(dy) \\
&= \prod_{i=1}^{n} \left(\int_{\{y_1^*,...,y_m^*\}} M(y_i, A_i)\,\mathbb{P}_{0,\theta}^{(2)}(dy_i)\right) \\
&= \prod_{i=1}^{n} \int_{A_i} \hat{f}_m(x)\,dx \\
&= \mathbb{P}_{0,n,f}^{(1)}(A_1 \times ... \times A_n).
\end{aligned}$$

Since product sets generate $\mathscr{B}[0,1]^n$, we know that $(M^{\otimes n})^{\vee}\mathbb{P}^{(2)}_{0,n,f} = \mathbb{P}^{(1)}_{0,n,f}$, and by proposition 3.15 we see that $\Delta_d(\mathsf{Expt}^{(2)}_{0,n}, \mathsf{Expt}^{(1)}_{0,n}) = 0$ In particular, if there is a decision rule in $\mathsf{Expt}^{(1)}_{0,n}$, say $\delta(.|y)$, then the following randomised rule

$$M^{\otimes n}\delta(.|y^*) = \int_{[0,1]^n} \delta(.|y) M^{\otimes n}(y^*, dy) \tag{3.52}$$

behaves as well as the original decision rule. Of course, one can also approximate this decision rule by using a suitable Monte Carlo scheme.

3.6.3 Wrapping Up

We can finally bound the L^1 distance between $\mathsf{Expt}_{0,n}$ and $\mathsf{Expt}^{(1)}_{0,n}$. For this, we utilise equation (3.16) and lemma 7.6 to show that

$$\begin{aligned}
(\Delta(\mathsf{Expt}_{0,n}, \mathsf{Expt}^{(1)}_{0,n}))^2 &\leq 4H^2(\mathsf{Expt}_{0,n}, \mathsf{Expt}^{(1)}_{0,n}) \\
&\leq 4n \sup_{f\in\Sigma} H^2(f\, d\mathsf{Leb}, \mathbb{P}_{0,f}) \\
&= 4n \int_0^1 \left(\sqrt{f(x)} - \sqrt{\hat{f}_m(x)}\right)^2 dx \\
&= 4n \int_0^1 \left(\frac{f(x) - \hat{f}_m(x)}{\sqrt{f(x)} + \sqrt{\hat{f}_m(x)}}\right)^2 dx \\
&\leq \frac{n}{\epsilon} \int_0^1 (f(x) - \hat{f}_m(x))^2\, dx
\end{aligned}$$

We can bound the right hand side by using the smoothness of density f. We state without proof that

$$\int_0^1 (f(x) - \hat{f}_m(x))^2\, dx = O(n(m^{-3} + m^{-2-2\gamma})) \tag{3.53}$$

Please see [10] or [11] for detailed argument. As a result, we know that $\Delta(\mathsf{Expt}_{0,n}, \mathsf{Expt}^{(1)}_{0,n}) = O(\sqrt{n}(m^{-3/2} + m^{-1-\gamma}))$. By a suitable choice of m, say $m = n^{(2+\gamma)}$, we have $\Delta(\mathsf{Expt}_{0,n}, \mathsf{Expt}^{(1)}_{0,n}) = o(1)$. We therefore conclude that any decision rule $\delta(.|y)$ in $\mathsf{Expt}_{0,n}$, the randomised rule $Q(M^{\otimes n}\delta)$ can also be applied to $\mathsf{Expt}^{(1)}_{0,n}$ with an extra $o(1)$ risk. A plausible way to simulate the randomised rule using Monte Carlo is as follows:

- We first randomly select elements $Y_1, ..., Y_N$ from the set $\{S(y) = z\}$, where N is much larger than m or n.
- For $\ell = 1, ..., N$, we evaluate $M^{\otimes n}\delta(.|Y_\ell)$ by a suitable Monte Carlo scheme.
- We take average over $\delta(.|Y_\ell)$ to get our new decision rule.

As we can see, the process of randomisation makes the new estimator difficult to be used in daily life.

Remark 3.18 In fact, [10, 11] develops the argument further to prove the asymptotic equivalence of density estimation and Gaussian white noise problems.

4 Local equivalence

The aim of the following two chapters is to prove the following:

Theorem 4.1 — Asymptotic Equivalence between density estimation and Gaussian white noise. Consider the density estimation problem $\mathsf{Expt}_{0,n}$ and the Gaussian white noise experiment $\mathsf{Expt}_{1,n}$ as specified below:

$$\mathsf{Expt}_{0,n} = \left([0,1]^n, \mathscr{B}([0,1]^n), \left\{d\mathbb{P}_{0,n,f} = (f\, d\mathsf{Leb})^{\otimes n}\right\}_{f\in\Sigma}\right), \tag{4.1}$$

$$\mathsf{Expt}_{1,n} = \left(C^0([0,1]), \mathscr{B}(C^0([0,1])), \{\mathbb{P}_{1,n,f}\}_{f\in\Sigma}\right); \tag{4.2}$$

where Leb refers to Lebesgue measure, and Σ is the parameter space

$$\Sigma = \Sigma_{\alpha,M,\epsilon} = \left\{ f \;\middle|\; |f(x)-f(y)| \le M|x-y|^\alpha,\ \int_0^1 f = 1,\ f \ge \epsilon \right\}, \quad \alpha > 1/2,\ M, \epsilon > 0 \tag{4.3}$$

such that the measure $\mathbb{P}_{1,n,f}$ is the distribution characterised by the following Gaussian white noise

$$dy(t) = \sqrt{f(t)}\, dt + \frac{1}{2\sqrt{n}}\, dW_t. \tag{4.4}$$

Then the above experiments are asymptotically equivalent.

We will mainly follow the arguments of [3]. We will first prove the equivalence of the two statements when the parameter space is restricted to some smaller neighborhood around an element in Σ. Specifically, we restrict ourselves to the following neighborhood:

$$\Sigma_n(f_0) = \left\{ f \in \Sigma \;\middle|\; \left\| \frac{f}{f_0} - 1 \right\|_{L^\infty((0,1))} \le \gamma_n := \frac{1}{n^{1/4}\ln n} \right\}, \tag{4.5}$$

where the choice of $\gamma_n = o(1)$ is to be justified. Notice that the CDF of f_0, i.e. F_0, is strictly increasing and hence invertible, we can define the log-likelihood function between densities f and f_0

$$\lambda_{f,f_0}(t) = \ln\left(\frac{f}{f_0}(F_0^{-1}(t))\right), \tag{4.6}$$

such that the integral with respect to Lebesgue measure is minus of the Kullback-Leibler divergence

$$\int_0^1 \lambda_{f,f_0}(t)\, \mathsf{Leb}(dt) = \int_0^1 \lambda_{f,f_0}(x)\, [F_0^* \mathsf{Leb}](dx) = -K(f_0 \| f). \tag{4.7}$$

At the end of the chapter, we will prove the following:

Theorem 4.2 — Local Equivalence. The following sequences of experiments are asymptotically equivalent with each other

$$\mathsf{Expt}_{0,n}(f_0) = ([0,1]^n, \mathscr{B}([0,1]^n), \{d\mathbb{P}_{0,n,f,f_0} = (f\, d\mathsf{Leb})^{\otimes n})_{f\in\Sigma_n(f_0)}\}) \tag{4.8}$$

$$\mathsf{Expt}_{i,n}(f_0) = (C^0([0,1]), \mathscr{B}(C^0([0,1])), \{\mathbb{P}_{i,n,f,f_0}\}_{f\in\Sigma_n(f_0)}), \quad i = 1,2,3 \tag{4.9}$$

where the measures $\{\mathbb{P}_{i,n,f,f_0}\}$ are the distributions characterised by the following Gaussian process:

$$i=1; \quad dy(t) = (\lambda_{f,f_0}(t) + K(f_0\|f))\, dt + \frac{1}{\sqrt{n}}\, dW_t, \tag{4.10}$$

$$i=2; \quad dy(t) = (f(t) - f_0(t))\, dt + \frac{\sqrt{f_0(t)}}{\sqrt{n}}\, dW_t, \tag{4.11}$$

$$i = 3; \quad dy(t) = \left(\sqrt{f(t)} - \sqrt{f_0(t)}\right) dt + \frac{1}{2\sqrt{n}} dW_t. \tag{4.12}$$

Of course, since $f_0(t)$ is a constant in our context, estimating $f(t)$ from $\mathsf{Expt}_{2,n}$ is equivalent to estimating $f(t) - f_0(t)$. Similarly, estimating estimating $\sqrt{f(t)}$ from $\mathsf{Expt}_{3,n}$ is equivalent to estimating $\sqrt{f(t)} - \sqrt{f_0(t)}$. With this, we know that $\mathsf{Expt}_{0,n}$ is actually equivalent to

$$\left(C^0([0,1]), \mathscr{B}(C^0([0,1])), \{\mathbb{P}_{1,n,f,f_0}\}_{f \in \Sigma_n(f_0)}\right)$$

as desired.

We will see that it is easier to prove the equivalence among $\mathsf{Expt}_{i,n}(f_0)$ for $i = 1, 2, 3$, since they share the same space of observables, so that we can safely use the bounds by L^1 and Hellinger distances established in last chapter to complete our proof. Showing the equivalence between $\mathsf{Expt}_{0,n}(f_0)$ and any other experiments, however, requires more careful thought. One would need to rewrite $\mathsf{Expt}_{0,n}(f_0)$ and $\mathsf{Expt}_{i,n}(f_0)$ into equivalent problems $\mathsf{Expt}^*_{0,n}(f_0) = (\mathscr{Y}^*, \mathscr{F}^*, \left\{\mathbb{P}^*_{0,n,f}\right\})_{f \in \Sigma_n(f_0)}$ and $\mathsf{Expt}^*_{i,n}(f_0) = (\mathscr{Y}^*, \mathscr{F}^*, \left\{\mathbb{P}^*_{i,n,f}\right\})_{f \in \Sigma_n(f_0)}$ respectively, so that they have the same space of observables (and action space). It turns out that it is natural to attempt proving equivalence between $\mathsf{Expt}_{0,n}(f_0)$ and $\mathsf{Expt}_{1,n}(f_0)$.

Let us try justify the choice of γ_n by having a sneak peak of what is coming after proving the local equivalence. Given sample $y_1, ..., y_n$, one can construct a sequence of "preliminary estimators" based on a fraction of sample $y_1, ..., y_{N_n}$ with $n \gg N_n \gg 1$, say $\hat{f}_n := \hat{f}_n(y_1, ..., y_{N_n})$, such that

$$\mathbb{P}_{0,n}(f \in \Sigma_n(\hat{f}_n)) = \mathbb{P}_{0,n}\left(\left\| \frac{f}{\hat{f}_n} - 1 \right\|_\infty \leq \gamma_n\right) \overset{n\to\infty}{\to} 1. \tag{4.13}$$

Our choice of N_n will be $N_n \sim n/2$. We can similarly construct another sequence of preliminary estimators $(\check{f}_n)$ for the experiment $\mathsf{Expt}_{1,n}$. Then, one can utilise the equivalence of $\mathsf{Expt}_{0,n}(\hat{f}_n)$ and $\mathsf{Expt}_{1,n}(\check{f}_n)$ for all n to find an estimator for an experiment that has comparable risk with its counterpart in another experiment. We will make this argument precise in the next chapter.

Notice for the sequence $(\hat{f}_n)$ to exist, one requires γ_n not to decay too fast. In fact, since we know that the optimal rate of supremum loss for estimating $f \in \Sigma_{\alpha,M,\epsilon}$ is $(\ln n/n)^{\alpha/(2\alpha+1)}$, which is $o(n^\beta)$ with $\beta > -1/4$ whenever $\alpha > 1/2$, one would naturally choose $\gamma_n = o(n^{-1/4})$. Here we incorporate the factor $1/(\ln n)$ to ensure that the γ_n are small enough to establish local equivalence.

4.1 Setup: Constructing appropriate experiments

To begin, it is shown that $\mathbb{P}_{i,n,f}$ are absolutely continuous with respect to some common measures ν_i, so that we can look at the L^1 or Hellinger distances between them. The proof of the following lemma fills in the technical details in justifying the forms of Radon-Nikodym derivatives $d\mathbb{P}_{i,n,f,f_0}/d\nu_i$.

Lemma 4.3 — Existence of Dominant Measure. In experiment $\mathsf{Expt}_{0,n}(f_0)$, we have $\mathbb{P}_{0,n,f,f_0} \ll \mathbb{P}_{0,n,f_0,f_0}$, and in experiment $\mathsf{Expt}_{1,n}(f_0)$, we have $\mathbb{P}_{1,n,f,f_0} \ll \mathbb{P}_{1,n,f_0,f_0}$.

Proof. Note that if a set A satisfies $\mathbb{P}_{0,n,f_0}(A) = 0$, then we know that

$$0 = \int_A f \, d\mathsf{Leb}^{\otimes n} \geq \epsilon \mathsf{Leb}^{\otimes n}(A) \geq 0 \overset{\epsilon>0}{\Longrightarrow} \mathsf{Leb}^{\otimes n}(A) = 0. \tag{4.14}$$

and that $\mathbb{P}_{0,n,f_0}(A) = 0$, which proves absolute continuity. Indeed, let $A = (a_1, b_1) \times (a_2, b_2) \times ... \times (a_n, b_n)$ be an open hyper-rectangle in $[0,1]^n$. Then

$$\int_A d\mathbb{P}_{0,n,f,f_0} \overset{\text{(Fubini)}}{=} \prod_{i=1}^n \int_{(a_i,b_i)} f(y_i)\,\mathsf{Leb}(dy_i) = \prod_{i=1}^n \int_{(a_i,b_i)} \frac{f(y_i)}{f_0(y_i)} f_0(y_i)\,\mathsf{Leb}(dy_i) = \int_A \prod_{i=1}^n \frac{f(y_i)}{f_0(y_i)}\, d\mathbb{P}_{0,n,f_0}.$$

With the second inequality is justified since f_0 is bounded away from zero; and that first and third inequality justified by Fubini's theorem since both f and f/f_0 are finite. Given the set of open rectangles generates $\mathscr{B}(C^0[0,1])$, we know that Radon-Nikodym derivatives of $\mathbb{P}_{0,n,f,f_0}$ with respect to $\mathbb{P}_{0,n,f_0,f_0}$ exist, and are almost surely equal to

$$\Lambda_{0,n}(f; f_0) = \frac{d\mathbb{P}_{0,n,f,f_0}}{d\mathbb{P}_{0,n,f_0}}(y) = \prod_{i=1}^n \frac{f}{f_0}(y_i) = \exp\left(\sum_{i=1}^n \ln \frac{f}{f_0}(y_i)\right), \quad y = (y_1, ..., y_n). \tag{4.15}$$

The above Radon-Nikodym derivative is also known as the *likelihood process*. We can also obtain the likelihood process in the $\mathsf{Expt}_{1,n}(f_0)$ experiment. Note that $\mathbb{P}_{1,n,f_0,f_0} =: Q_{1,n}$ represents the scaled Brownian motion $W_t/\sqrt{n}$, and that other $\mathbb{P}_{1,n,f,f_0}$ are measures representing Gaussian processes with the same noise term. We therefore know from Girsanov theorem that the two measures are absolutely continuous with

$$\begin{aligned}\Lambda_{1,n}(f; f_0) &:= \frac{d\mathbb{P}_{1,n,f,f_0}}{d\mathbb{P}_{1,n,f_0,f_0}}\left(\frac{1}{\sqrt{n}}\mathsf{w}\right) \\ &= \exp\left(\int_0^1 \sqrt{n}\left(\lambda_{f,f_0}(s) + K(f_0\|f)\right)\, d\mathsf{w}_t - \frac{n}{2}\int_0^1 \left(\lambda_{f,f_0}(s) + (K(f_0\|f)^2)\right)\, ds\right). \end{aligned} \tag{4.16}$$

For further discussions on Girsanov theorem, please refer to chapter 7.3. ■

This indicates that we have the following:

$$\mathsf{Expt}_{0,n}(f_0) = ([0,1]^n, \mathscr{B}([0,1]^n), \{\Lambda_{0,n}(f; f_0)\, d\mathbb{P}_{0,n,f_0}\}_{f\in\Sigma_n(f_0)}), \text{ and} \tag{4.17}$$

$$\mathsf{Expt}_{1,n}(f_0) = (C^0[0,1], \mathscr{B}(C^0[0,1]), \{\Lambda_{1,n}(f; f_0)\, dQ_{1,n}\}_{f\in\Sigma_n(f_0)}). \tag{4.18}$$

For computational convenience, we apply a bijective change of random variable $y \mapsto z$ with entries $z_i = F_0(y_i)$. Clearly the vector z is a sufficient statistic of y, so $\mathsf{Expt}_{0,n}(f_0)$ is equivalent with the following experiments

$$\overline{\mathsf{Expt}}_{0,n}(f_0) = \left([0,1]^n, \mathscr{B}([0,1]^n), \left\{d\overline{\mathbb{P}}_{0,n} := \overline{\Lambda}_{0,n}(f; f_0)(z)\, d\mathsf{Leb}^{\otimes n}\right\}_{f\in\Sigma_n(f_0)}\right), \tag{4.19}$$

where one have

$$[\overline{\Lambda}_{0,n}(f; f_0)](z) = \exp\left(\sum_{i=1}^n \ln \frac{f}{f_0}(F_0^{-1}(z_i))\right) = \exp\left(\sum_{i=1}^n \lambda_{f,f_0}(z_i)\right). \tag{4.20}$$

We will abuse notations and drop the overlines in (4.20). Let us follow [3] closely and further rewrite the likelihood by using empirical measures as defined below:

Definition 4.4 — Empirical Measures and Uniform Processes (see also (7) and (13)). Let Pr be a measure on $([0,1], \mathscr{B}[0,1])$. Assume $\xi_i : (\Omega, \mathscr{F}, \mathbb{P}) \to ([0,1], \mathscr{B}[0,1])$ are independent variable with distribution Pr.

- An unnormalised empirical measure of Pr is a random variable μ_n from $(\Omega, \mathscr{F}, \mathbb{P})$ to space of measure of $([0,1], \mathscr{B}[0,1])$ with

$$[\mu_n(\omega)](A) \overset{d}{=} \sum_{i=1}^n \delta_{\xi_i(\omega)}(A) \text{ in distribution.} \tag{4.21}$$

- A (normalised) empirical measure of Pr is a random variable P_n from $(\Omega, \mathscr{F}, \mathbb{P})$ to space

of measure of $([0,1], \mathscr{B}[0,1])$ with

$$[P_n(\omega)](A) \overset{d}{=} \frac{1}{n}\mu_n(A) \text{ in distribution.} \tag{4.22}$$

where $\overset{d}{=}$ represents equality in distribution, and $\delta_x(A) = \mathbb{I}_A(x)$ is the Dirac measure at x.

- A uniform process of Pr is a random variable U_n mapping elements from $(\Omega, \mathscr{F}, \mathbb{P})$ to measure

$$[U_n(\omega)](A) = \sqrt{n}([P_n(\omega)](A) - \mathsf{Pr}(A)) \text{ in distribution.} \tag{4.23}$$

If not specified, empirical measure below refers to normalised empirical measure.

Empirical measure is an estimator the underlying distribution of a dataset. We highlight the following way to construct an empirical measure as stated in [13]. Consider the space $(\Omega, \mathscr{F}, \mathbb{P}) = ([0,1], \mathscr{B}[0,1], \mathsf{Pr})^{\otimes\infty}$, which exists by Kolmogorov extension theorem. Let $\omega = (\omega_1, \omega_2, ...) \in \Omega$, and define the projection variables $\mathsf{proj}_i(\omega) = \omega_i$; then we know that the following random variable

$$[P_n(\omega)](A) = \frac{1}{n}\sum_{i=1}^{n} \delta_{\mathsf{proj}_i(\omega)}(A) \tag{4.24}$$

is an empirical measure of Pr, defined on the space $(\Omega, \mathscr{F}, \mathbb{P})$ as specified above.

We can define integrals with respect to the random measure $P_n(\omega)$ by following the standard procedure of first defining on simple functions then extend to any Lebesgue-measurable function by pointwise limit. Observe for indicator functions $g(x) = \mathbb{I}_A(x)$, $A \in \mathscr{B}([0,1])$, we have

$$\int_{[0,1]} g(x)\,[P_n(\omega)](dx) = [\mathbb{P}_n(\omega)](A) \overset{d}{=} \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_A(\xi_i(\omega)) = \frac{1}{n}\sum_{i=1}^{n} g(\xi_i(\omega)) \tag{4.25}$$

We claim that the equality extends for any Lebesgue-measurable functions:

Lemma 4.5

$$\int_{[0,1]} g(x)\,[P_n(\omega)](dx) \overset{d}{=} \frac{1}{n}\sum_{i=1}^{n} g(\xi_i(\omega)) \tag{4.26}$$

Proof. We first extend the definition by linearity so that (4.25) holds for simple functions $g(x) = \sum_{j=1}^{m} \alpha_j \mathbb{I}_{C_j}(x)$ (assuming C_j are disjoint):

$$\begin{aligned}\int_{[0,1]} g(x)\,[P_n(\omega)](dx) &= \sum_{j=1}^{m} \alpha_j [\mathbb{P}_n(\omega)](C_j) \\ &\overset{d}{=} \sum_{j=1}^{m} \left(\frac{\alpha_j}{n} \sum_{i=1}^{n} \mathbb{I}_A(\xi_i(\omega)) \right) \\ &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} \alpha_j \mathbb{I}_{C_j}(\xi_i(\omega)) = \frac{1}{n}\sum_{i=1}^{n} g(\xi_i(\omega))\end{aligned}$$

We can further extend the definition to non-negative Lebesgue-measurable function by taking limits. Note that any non-negative Lebesgue-measurable function $g(x)$ can be approximated by a *pointwise* limit of simple functions $g_i(x)$. By monotone convergence theorem, for *all* ω we have

$$\int_{[0,1]} g(x)\,[P_n(\omega)](dx) \overset{\text{(MCT)},\, d}{=} \lim_{i\to\infty} \left(\frac{1}{n}\sum_{i=1}^{n} g_n(\xi_i(\omega)) \right) = \frac{1}{n}\sum_{i=1}^{n} g(\xi_i(\omega))$$

Finally for general Lebesgue measurable function g, we can decompose it to positive and negative parts $g = g^+ - g^-$, $f^\pm = \max(\pm g, 0)$ and conclude

$$\begin{aligned}\int_{[0,1]} g(x)\,[P_n(\omega)](dx) &= \int_{[0,1]} g^+(x)\,[P_n(\omega)](dx) - \int_{[0,1]} g^-(x)\,[P_n(\omega)](dx)\\ &\overset{d}{=} \frac{1}{n}\sum_{i=1}^{n}(g^+(\xi_i(\omega)) - g^-(\xi_i(\omega))) = \frac{1}{n}\sum_{i=1}^{n} g(\xi_i(\omega))\end{aligned}$$

■

With the above lemma, we can rewrite the likelihood function $\Lambda_{0,n}(f, f_0)$ as

$$[\Lambda_{0,n}(f; f_0)](z) = \exp\left(n\int_0^1 \ln\frac{f}{f_0}(F_0^{-1}(t))\,[P_n(z)](dt)\right) \tag{4.27}$$

where

$$[P_n(z)](A) = \frac{1}{n}\sum_{i=1}^{n}\delta_{z_i}(A) \tag{4.28}$$

is an empirical measure of Leb under $([0,1], \mathscr{B}[0,1], \mathsf{Leb})^{\otimes n}$. If we further define the map $z \mapsto U_n(z)$ with $[U_n(z)](A) := \sqrt{n}([P_n(z)](A) - \mathsf{Leb}(A))$, we know that

$$\begin{aligned}[\Lambda_{0,n}(f; f_0)](z) &= \exp\left(\sqrt{n}\int_0^1 \lambda_{f,f_0}(t)\,[U_n(z)](dt) + n\int_0^1 \lambda_{f,f_0}(t)\,dt\right)\\ &= \exp\left(\sqrt{n}\int_0^1 \lambda_{f,f_0}(t)\,[U_n(z)](dt) - nK(f_0\|f)\right)\end{aligned} \tag{4.29}$$

Abusing notation and consider $\Lambda_{0,n}(f; f_0)$ as a real-valued measurable map on $U_n \in \mathscr{M}_n$, we see that $\mathsf{Expt}_{0,n}(f_0)$ is equivalent to

$$\mathsf{Expt}_{0,n}(f_0) = (\mathscr{M}_n, \mathscr{F}_n, \{\Lambda_{0,n}(f; f_0)\,U_n^*\mathsf{Leb}^{\otimes n}(dz)\}_{f\in\Sigma_n(f_0)}) \tag{4.30}$$

where $\mathscr{M}_n$ is the image of the map U_n, $\mathscr{F}_n$ is a σ-algbera such that the map U_n is Lebesgue-measurable (which we will not specify here), and that $U_n^*\mathsf{Leb}^{\otimes n}$ is the push-forward measure of Lebesgue measure under the map P_n. Note that this push-forward measure is the distribution of a uniform process.

We can also rewrite the likelihood process $\Lambda_{1,n}(f, f_0)$ to similar form as in (4.29). For this, let us recall the definition of a Brownian bridge:

Definition 4.6 — Brownian bridge. A stochastic process $[B(t)](\omega)$ on probability space $(\Omega, \mathscr{F}, \mathbb{P})$ indexed with $t \in [0,1]$ is a Brownian bridge if

1. $B(0) = B(1) = 0$ almost surely,
2. The sample path $[B(t)](\omega)$ is continuous almost surely.
3. For each finite subset $S = \{s_1, ..., s_n\}$ of $[0,1]$, the random vector $\pi_S(U) = (B_{s_1}, ..., B_{s_n})$ has a multivariate normal distribution with zero mean with covariances given by

$$\mathbb{E}[B(s)B(t)] = s - st, \quad \forall s, t \in S, s < t \tag{4.31}$$

Remark 4.7 We can also view a Brownian bridge B as a mapping on $([0,1], \mathscr{B}[0,1]) \otimes (\Omega, \mathscr{F}, \mathbb{P})$, where $[B(t)](.)$ is always a random variable on $(\Omega, \mathscr{F}, \mathbb{P})$ for all t that satisfies the conditions above.

As a key observation, we note that if a stochastic process $[W(t)](\omega)$ is a Brownian motion, then the stochastic process $B(t) = W(t) - tW(1)$ is a Brownian bridge. As a result, we may express the integral:

$$\begin{aligned}\int_0^1 (\lambda_{f,f_0}(s) + K(f_0\|f))\, d\mathsf{w}_t &= \int_0^1 \lambda_{f,f_0}(s)\, d\mathsf{w}_t + K(f_0\|f)\int_0^1 d\mathsf{w}_t \\ &= \int_0^1 \lambda_{f,f_0}(s)\, d\mathsf{w}_t - W_1 \int_0^1 \lambda_{f,f_0}(s)\, dt \\ &= \int_0^1 \lambda_{f,f_0}(s)\, d\mathsf{B}_t\end{aligned}$$

Here we adopt a change of variable $\mathsf{B}(\mathsf{w}) = \mathsf{w}_t - t\mathsf{w}_1$ for sample paths of Brownian motions w_t, which results in Brownian bridges. As a result, we have

$$\Lambda_{1,n}(f; f_0) := \exp\left(\sqrt{n}\int_0^1 \lambda_{f,f_0}(s)\, d\mathsf{B}(\mathsf{w})_t - \frac{n}{2}\int_0^1 \left(\lambda_{f,f_0}(s) + (K(f_0\|f))^2\right)\, ds\right) \tag{4.32}$$

Viewing $\Lambda_{1,n}$ as a function on $\mathsf{B}(\mathsf{w})$, we have

$$\mathsf{Expt}_{1,n}(f_0) = \left(C^0[0,1], \mathscr{B}(C^0[0,1]), \{[\Lambda_{1,n}(f; f_0)](\mathsf{B}(\mathsf{w}))\, dQ_{1,n}\}_{f\in\Sigma_n(f_0)}\right) \tag{4.33}$$

Of course, scaling up from $W/\sqrt{n}$ to W does not affect the experiment, so one can consider $Q_{1,n}$ as the distribution of a Brownian motion. Finally, by noticing that $\mathsf{w} \mapsto \mathsf{B}(\mathsf{w})$ is indeed a sufficient estimator of w, we know that $\mathsf{Expt}_{1,n}(f_0)$ is equivalent to

$$\overline{\mathsf{Expt}}_{1,n}(f_0) = \left(C^0[0,1], \mathscr{B}(C^0[0,1]), \left\{[\Lambda_{1,n}(f; f_0)](\mathsf{B})\, d\overline{Q}_{1,n}\right\}_{f\in\Sigma_n(f_0)}\right) \tag{4.34}$$

where $\overline{Q}_{1,n}$ represents the distribution of a Brownian bridge. We will abuse notation and drop the overlines in the above expression.

The similarity of the expressions in (4.29) and (4.32) provides us a natural route for proving the required asymptotic equivalence by considering some form of convergence of uniform process towards Brownian bridge. Here we state the main non-asymptotic inequality, which allows us to construct suitable intermediate experiments $\mathsf{Expt}^*_{0,n}(f_0)$ and $\mathsf{Expt}^*_{1,n}(f_0)$.

Theorem 4.8 — Komlos-Major-Tusnady (KMT) Inequality for a function ((3), Theorem 2.3). There exists a probability space $(\Omega^*, \mathscr{F}^*, \mathbb{P}^*)$ such that we can construct a sequence of empirical measures of Leb, denoted as $\mathbb{U}_n(\omega)$, and a Brownian bridge $\mathbb{B}$. In particular, for all Lebesgue-measurable functions $g \in L^\infty[0,1]$ such that

$$\|g\|^2_{H_2^{1/2}} = \inf_{h\in(0,1/2)} \frac{1}{h}\int_h^{1-h} (g(x+h) - g(x))^2\, dx < \infty \tag{4.35}$$

Denote

$$\mathbb{U}_n(g) = \int_0^1 g(t)\, \mathbb{U}_n(dt), \quad \mathbb{B}(g) = \int_0^1 g(t)\, d\mathsf{B}_t \tag{4.36}$$

Then we have the non-asymptotic inequality: there exists universal constant C such that

$$\mathbb{P}^*\left[\sqrt{n}|\mathbb{U}_n(g) - \mathbb{B}(g)| \geq C(\|g\|_{L^\infty} + \|g\|_{H_2^{1/2}})(t + \ln n)(\ln n)^{1/2}\right] \leq C\exp(-t) \tag{4.37}$$

This version of the KMT inequality is based on a more generalised version of KMT inequality in theorem 3.5 of [13], and we will use the inequality without proof. From this result, we can consider the following intermediate experiments

$$\mathsf{Expt}^*_{0,n}(f_0) = (\Omega, \mathscr{F}, \{\mathbb{P}^*_{0,n,f}\}_{f\in\Sigma_n(f_0)}) \tag{4.38}$$

$$\mathsf{Expt}^*_{1,n}(f_0) = (\Omega, \mathscr{F}, \{\mathbb{P}^*_{1,n,f}\}_{f\in\Sigma_n(f_0)}) \tag{4.39}$$

where $d\mathbb{P}^*_{0,n,f,f_0} = [\Lambda_{0,n}(f;f_0)](U_n(\omega))\,d\mathbb{P}^*$, and that $d\mathbb{P}^*_{1,n,f} = [\Lambda_{1,n}(f;f_0)](B(\omega))\,d\mathbb{P}^*$. By a change of variable arguments we see that $\mathsf{Expt}^*_{i,n}(f_0)$ is equivalent to $\mathsf{Expt}_{i,n}(f_0)$ for $i = 0, 1$. As a result, we have developed a common probability space for the applications of bounds that control the L^1 or Hellinger distance between two experiments.

4.2 Application for the KMT Inequality, First Attempt

Let us study how can we utilise the KMT inequality by making a few observations for the function λ_{f,f_0}:

Lemma 4.9 We know that, for all sufficiently large n and $f \in \Sigma_n(f_0), f_0 \in \Sigma$, there is a $C > 0$ such that

$$\|\lambda_{f,f_0}\|_{L^\infty} \le C\gamma_n \tag{4.40}$$

and that

$$|\lambda_{f,f_0}(s) - \lambda_{f,f_0}(t)| \le C|s-t|^\alpha \tag{4.41}$$

As a result, $\|\lambda_{f,f_0}\|^2_{H^{1/2}_2} \le C$.

Proof. To show (4.40), we note that γ_n is $o(1)$, so there is a universal constant $\gamma \in (0,1)$ for which $\gamma_n \in [-\gamma, \gamma]$ for sufficiently large n. (For the choice of $\gamma_n = n^{-1/4}(\ln n)^{-1}$ we have $\gamma = 7/10$ for $n \ge 3$.) We have

$$|\ln(1+x)| \le \frac{|\ln(1-\gamma)|}{\gamma}|x| \quad \forall x \in [-\gamma,\gamma] \tag{4.42}$$

Therefore

$$\|\lambda_{f,f_0}\|_{L^\infty} = \sup_{t\in[0,1]}\left|\ln\left(1+\frac{f}{f_0}-1\right)\right| \le \frac{|\ln(1-\gamma)|}{\gamma}\left\|\frac{f}{f_0}-1\right\|_{L^\infty} \le \frac{|\ln(1-\gamma)|}{\gamma}\gamma_n \tag{4.43}$$

To show (4.41), we first note that $F_0^{-1}(t)$ has derivative $1/(f_0(F_0^{-1}(t)))$, which is bounded by $1/\epsilon$. We therefore know that F_0^{-1} is $1/\epsilon$-Lipschitz by mean-value inequality. In addition, note that

$$\frac{d(\ln(1+x))}{dx} = \frac{1}{1+x} \in \left[\frac{1}{1-\gamma}, \frac{1}{1+\gamma}\right], \quad \forall x \in [-\gamma,\gamma] \tag{4.44}$$

therefore we have

$$\begin{aligned}\left|\ln\frac{f}{f_0}(s) - \ln\frac{f}{f_0}(t)\right| &= \left|\ln\left(1+\frac{f}{f_0}(s)-1\right) - \ln\left(1+\frac{f}{f_0}(t)-1\right)\right| \\ &\le \frac{1}{1-\gamma}\left|\frac{f}{f_0}(s) - \frac{f}{f_0}(t)\right| \\ &\le \frac{1}{\epsilon(1-\gamma)}|f(s)-f(t)| \le \frac{M}{\epsilon(1-\gamma)}|s-t|^\alpha\end{aligned}$$

We therefore have

$$|\lambda_{f,f_0}(s) - \lambda_{f,f_0}(t)| \le \frac{M}{\epsilon(1-\gamma)}|F_0^{-1}(s) - F_0^{-1}(t)|^\alpha \le \frac{M}{\epsilon^{1+\alpha}(1-\gamma)}|s-t|^\alpha \tag{4.45}$$

Selecting $C = \max\left(|\ln(1-\gamma)|/\gamma, M/(\epsilon^{1+\alpha}(1-\gamma))\right)$ will prove the first two inequalities. To complete the proof, we note that

$$\frac{1}{h}\int_h^{1-h}(\lambda_{f,f_0}(u+h) - \lambda_{f,f_0}(u))^2\,du \le \int_0^1 Ch^{2\alpha-1} \le C\,(1/2)^{2\alpha-1} < C \tag{4.46}$$

as desired, noting that $2\alpha - 1 > 0$. ∎

We therefore encounter a difficulty of using KMT inequality on function $g = \lambda_{f,f_0} + K(f_0\|f)$, since we only have an $\mathrm{ord}(1)$ bound for $\|g\|_{H_2^{1/2}}$, so the KMT inequality only yields a bound in the form

$$\mathbb{P}\left[\sqrt{n}|\mathbb{U}_n(g) - \mathbb{B}(g)| \geq \tilde{C}(t + \ln n)(\ln n)^{1/2}\right] \leq C\exp(-t) \tag{4.47}$$

and by applying $t = k\ln n$ we have

$$\mathbb{P}\left[\sqrt{n}|\mathbb{U}_n(g) - \mathbb{B}(g)| \geq (k+1)\tilde{C}(\ln n)^{3/2}\right] \leq C/n^k \tag{4.48}$$

which is not sharp enough for our application. We will see how one can improve the analysis by a divide-and-conquer trick.

4.3 Divide and Conquer

We would like to break the experiment $\mathsf{Expt}_{0,n}(f_0)$ and $\mathsf{Expt}_{1,n}(f_0)$ up into k_n smaller experiments. Then, we can obtain an upper bound of the Le Cam distance between $\mathsf{Expt}_{0,n}(f_0)$ and $\mathsf{Expt}_{1,n}(f_0)$ by accumulating the upper bounds of the Le Cam distances among each smaller experiments. In this section we justify the choice of $k_n \sim n^{1/2}(\ln n)^{-3/2}$.

We formalise the idea by first defining the notion of product experiments.

Definition 4.10 — Product Experiment. Consider a finite family of experiments

$$\mathsf{Expt}_i := (\mathscr{Y}_i, \mathscr{F}_i, \{\mathbb{P}_{i,\theta}\}_{\theta\in\Theta}, \mathscr{A}), i = 1, ..., n$$

that shares the same parameter space Θ. Then the product experiments of this family Expt_i is

$$\bigotimes_{i=1}^{n} \mathsf{Expt}_i := \left(\mathscr{Y}, \mathscr{F}, \{\mathbb{P}_\theta\}_{\theta\in\Theta}\right) \tag{4.49}$$

where $(\mathscr{Y}, \mathscr{F}) := \otimes_{i=1}^n(\mathscr{Y}_i, \mathscr{F}_i)$ is the product σ-algebra of the family $(\mathscr{Y}, \mathscr{F}_i)$ generated by Cartesian products of the elements $\mathscr{F}_i$, and that $\mathbb{P}_\theta = \otimes_{i=1}^n \mathbb{P}_{i,\theta}$ is the product measure of $\mathbb{P}_{i,\theta}$.

and note the following lemma for the Hellinger distance between two product experiments

Lemma 4.11 — Hellinger distance. Suppose that $\mathsf{Expt}_{i,j} := \left(\mathscr{Y}_j, \mathscr{F}_j, \left\{\mathbb{P}_{i,\theta}^{(j)}\right\}_{\theta,\Theta}\right)$ for $i = 0, 1$ and $j = 1, 2, ..., k$, so that

- $\mathsf{Expt}_{i,j}$ is defined on the same parameter space Θ,
- when j is fixed, $\mathsf{Expt}_{i,j}$ shares the same sample space $(\mathscr{Y}_j, \mathscr{F}_j)$ for all $i = 0, 1$.

In addition, assume $\mathbb{P}_{1,\theta}^{(j)} \ll \nu_j$ for all i, j. Consider the experiment $\mathsf{Expt}_0 = \otimes_{j=1}^k \mathsf{Expt}_{0,j}$ and similarly for Expt_1, then we have

$$H^2\left(\mathsf{Expt}_0, \mathsf{Expt}_1\right) \leq 2\sum_{i=1}^{k} H^2(\mathsf{Expt}_{0,j}, \mathsf{Expt}_{1,j}) \tag{4.50}$$

This is a natural corollary from lemma 7.6, and please refer to section 7.2 for proof.

We now describe how $\mathsf{Expt}_{0,n}^*(f_0)$ and $\mathsf{Expt}_{1,n}^*(f_0)$ are broken up into products of smaller experiments. For our convenience, we return to the original experiments $\mathsf{Expt}_{0,n}(f_0)$ and $\mathsf{Expt}_{1,n}(f_0)$ (with different sample space). The idea is to partition $[0,1]$ into smaller intervals $(D_j)_{j=1}^{k_n}$, with $D_j = [(j-1)/k_n, j/k_n)$, and approximate $\mathsf{Expt}_{0,n}(f_0)$ as a product of $\mathsf{Expt}_{0,n,j}(f_0)$. Here $\mathsf{Expt}_{0,n,j}(f_0)$

is the experiments $\mathsf{Expt}_{0,n}(f_0)$ with observations truncated over the interval D_j, so that the i-th entry of observations in this new experiment, $\tilde{y}_i$, is equal to the original i-th entry y_i only if $y_i \in D_j$, otherwise we have $\tilde{y}_i$ equals to zero. We can therefore write

$$\mathsf{Expt}_{0,n,j}(f_0) = \left([0,1]^n, \mathscr{B}[0,1]^n, \left\{\mathbb{P}^{(j)}_{0,n,f}\right\}_{f\in\Sigma_n(f_0)}\right) \tag{4.51}$$

where

$$\mathbb{P}^{(j)}_{0,n,f,f_0} = \bigotimes_{i=1}^{n} \mathbb{P}^{(j)}_{0,n,f,f_0,i}, \quad \mathbb{P}^{(j)}_{0,n,f,f_0,i}(B) = (1-p)\delta_0(B) + \int_{B\cap D_j} f \, d\mathsf{Leb}, \quad B \in \mathscr{B}[0,1] \tag{4.52}$$

and that $p = \int_{D_j} f \, d\mathsf{Leb}$. Note that if B does not have any intersection with D_j, then $\mathbb{P}^{(j)}_{0,n,f,f_0,i}(B) = (1-p)\delta_0(B)$, i.e. the contribution of the measure of B is from the atom $\{0\}$ when B contains zero. Moreover if $B \subseteq D_j$ and B does not containing zero, then the probability $\mathbb{P}^{(j)}_{0,n,f,f_0,i}(B)$ is equal to the probability distribution from the non-truncated experiment, $\int_B f \, d\mathsf{Leb}$. The reason behind breaking up the experiment in this particular way is because the log-likelihood λ_{f,f_0} looks much more smoother when we restrict our view to a small neighborhood of $[0,1]$. This may give a much sharper bound of the Hellinger distances between one component $\mathsf{Expt}_{0,n,j}(f_0$ and its counterpart $\mathsf{Expt}_{1,n,j}$, such that the sum of distances is $o(1)$. The local smoothness of the log-likelihood is characterised in lemma 4.13.

Note that the truncated experiments also possess a dominant measure, and the following proof fills in the details for computing the likelihood process:

Lemma 4.12 — Existence of Dominant Measure for $\mathsf{Expt}_{0,n,f}$. We have $\mathbb{P}^{(j)}_{0,n,f,f_0} \ll \mathbb{P}^{(j)}_{0,n,f_0}$.

Proof. It suffices to show that $\mathbb{P}^{(j)}_{0,n,f,f_0,i} \ll \mathbb{P}^{(j)}_{0,n,f_0,f_0,i}$ for all $i \in 1, ..., k_n$. Let $B \in \mathscr{B}[0,1]$. Since f, f_0 are bounded below by $\epsilon > 0$, we know that that $\mathbb{P}_{0,n,f_0,f_0,i}(B) = 0 \iff 0 \notin B$ and $B \cap D_j = \varnothing$, and that the necessary and sufficient condition does not depend on the choice of f, so we know that $\mathbb{P}^{(j)}_{0,n,f,f_0,i}$ and $\mathbb{P}^{(j)}_{0,n,f_0,f_0,i}$ are mutually continuous.

Let $\Lambda^{(j)}_{0,n,i}(f;f_0) := \dfrac{d\mathbb{P}_{0,n,f,f_0,i}}{d\mathbb{P}_{0,n,f_0,f_0,i}}$. We divide into two cases:

- If $B \subseteq D_j$, then we have

$$\int_B \Lambda^{(j)}_{0,n,i}(f;f_0)\, d\mathbb{P}_{0,n,f_0,f_0,i} = \int_B d\mathbb{P}_{0,n,f,f_0,i} \overset{\text{Lemma 4.3}}{=} \int_B \frac{f}{f_0} d\mathbb{P}_{0,n,f_0,f_0,i} \tag{4.53}$$

  since the above holds for all $B \subseteq D_j$, we know that $\Lambda^{(j)}_{0,n,i}(f;f_0) = f/f_0$ for almost all $y \in D_j$

- If $B \subseteq D_j^C$, then we have

$$\int_B \Lambda^{(j)}_{0,n,i}(f;f_0)\, d\mathbb{P}_{0,n,f_0,f_0,i} = \int_B d\mathbb{P}_{0,n,f,f_0,i} = (1-p)\delta_0(B) = \frac{1-p}{1-p_0}\int_B d\mathbb{P}_{0,n,f_0,f_0,i} \tag{4.54}$$

  with $p_0 = \int_{D_j} f_0 \, d\mathsf{Leb}$. Since the above holds for all $B \subseteq D_j^C$, we know that $\Lambda^{(j)}_{0,n,i}(f;f_0) = (1-p)/(1-p_0)$ for almost all $y \in D_j$.

This shows that

$$[\Lambda^{(j)}_{0,n,i}(f;f_0)](y) \equiv \frac{f}{f_0}(y)\mathbb{I}_{D_j}(y) + \frac{1-p}{1-p_0}\mathbb{I}_{D_j^C(y)}, \quad \forall i \tag{4.55}$$

almost everywhere in $[0,1]$. Now let $\Lambda^{(j)}_{0,n}(f;f_0) := \dfrac{d\mathbb{P}_{0,n,f,f_0}}{d\mathbb{P}_{0,n,f_0,f_0}}$, then we have

$$[\Lambda^{(j)}_{0,n}(f;f_0)](y) \equiv \exp\left(\sum_{i=1}^{n} \ln\left(\frac{f}{f_0}(y_i)\mathbb{I}_{D_j}(y_i) + \frac{1-p}{1-p_0}\mathbb{I}_{D_j^C}(y_i)\right)\right), \quad \forall i \tag{4.56}$$

$$= \exp\left(\sum_{i=1}^{n}\left(\mathbb{I}_{D_j}(y_i)\ln\frac{f}{f_0}(y_i) + \mathbb{I}_{D_j^C}(y_i)\ln\frac{1-p}{1-p_0}\right)\right) \tag{4.57}$$

■

Of course, one can show that $\mathsf{Expt}_{0,n,j}(f_0)$ is equivalent to the following experiment

$$\overline{\mathsf{Expt}}_{0,n,j}(f_0) = \left([0,1]^n, \mathscr{B}([0,1]^n), \left\{d\overline{\mathbb{P}}^{(j)}_{0,n,f,f_0} := \overline{\Lambda}^{(j)}_{0,n}(f;f_0)(z)\, d\mathsf{Leb}^{\otimes n}\right\}_{f\in\Sigma_n(f_0)}\right), \tag{4.58}$$

with $A_j = F_0(D_j)$ and

$$[\overline{\Lambda}_{0,n}(f;f_0)](z) = \exp\left(\sum_{i=1}^{n}\lambda_{f,f_0,A_j}(z_i)\right) \tag{4.59}$$

$$\lambda_{f,f_0,A_j}(t) = \mathbb{I}_{A_j}(t)\ln\frac{f}{f_0}(F_0^{-1}(t)) + \mathbb{I}_{A_j^C}(t)\ln\frac{1-p}{1-p_0}, \tag{4.60}$$

by a bijective change of random variable $y \mapsto z$ with entries $z_i = F_0(y_i)$.

4.3.1 Properties of λ_{f,f_0,A_j}

At this points, let us characterise the local smoothness of λ_{f,f_0} by looking at λ_{f,f_0,A_j}, which agrees with λ_{f,f_0} on the neighborhood A_j. Let $A_j = [a_{j,1}, a_{j,2})$. Then we know that

$$p_0 = \mathsf{Leb}(A_j) = a_{j,2} - a_{j,1} = \int_{D_n} f_0\, dt \geq \epsilon k_n^{-1} \tag{4.61}$$

Moreover since $f_0 \in \Lambda^\alpha(M)$, we know that $|f_0((j-1)/k_n + t) - f_0((j-1)/k_n)| \leq M|t|^\alpha$. This yields

$$p_0 = \mathsf{Leb}(A_j) \leq \int_{[0,k_n^{-1})} (f_0((j-1)/k_n) + Mt^\alpha)\, dt \leq (1+M)k_n^{-1} \tag{4.62}$$

So there is a universal C_1 and C_2 such that $\mathsf{Leb}(A_j)/k_n \in [C_1, C_2]$. In short, we have $\mathsf{Leb}(A_j) = p_0 = \mathsf{ord}(k_n)$. Therefore, we have the following:

Lemma 4.13 — Properties of λ_{f,f_0,A_j}. There exists a constant $C > 0$ such that

1. $\sup_{t\in A_j} |\lambda_{f,f_0,A_j}| \leq C\gamma_n$
2. $\sup_{t\in A_j^c} |\lambda_{f,f_0,A_j}| \leq Ck_n^{-1}\gamma_n$
3. $\int \lambda^2_{f,f_0,A_j} \leq Ck_n^{-1}\gamma_n^2$
4. $\int -\lambda_{f,f_0,A_j} \leq Ck_n^{-1}\gamma_n^2$
5. For k_n of sufficiently large order, one have $\left\|\lambda_{f,f_0,A_j}\right\|_{H_2^{1/2}} \leq C\gamma_n$

Proof. We follow closely the technical proofs in section 5 of [3], paying attentions to the conditions of k_n for point (5) to hold:

1. Follows from lemma 4.9

2. We control $\ln \dfrac{1-p}{1-p_0}$. Note

$$\left|1 - \frac{1-p}{1-p_0}\right| = \left|\frac{p_0 - p}{1-p_0}\right| \leq \frac{\int_{D_j} |f/f_0 - 1| f_0}{1-p_0} \leq \left(1 - \frac{1}{1-p_0}\right)\gamma_n \lesssim p_0\gamma_n \lesssim \gamma_n k_n^{-1}$$

So by an argument similar to the one in the derivation of inequality (4.42), we see that $\sup_{t \in A_j^c} |\lambda_{f,f_0,A_j}| \lesssim k_n^{-1}\gamma_n$.

3. We break up the integral and note that

$$\int \lambda_{f,f_0,A_j}^2 = \int_{A_j} \lambda_{f,f_0,A_j}^2 + \int_{A_j^c} \lambda_{f,f_0,A_j}^2 \leq Ck_n^{-1}\gamma_n^2 + Ck_n^{-2}\gamma_n^2 \lesssim k_n^{-1}\gamma_n^2 \tag{4.63}$$

4. Once again, we control by noting that there is a $C > 0$ such that

$$\begin{aligned}\int -\lambda_{f,f_0,A_j} &= \int -\ln \exp\left(\lambda_{f,f_0,A_j}\right) \\ &= \int 1 - \exp\left(\lambda_{f,f_0,A_j}\right) + C(\exp\left(\lambda_{f,f_0,A_j}\right) - 1)^2\end{aligned}$$

where the existence of C is guaranteed by points (1), (2) and equation (4.42). Since $\int \exp\left(\lambda_{f,f_0,A_j}\right) = 1$, we have

$$\begin{aligned}\int -\lambda_{f,f_0,A_j} &\lesssim \int \left(\mathbb{I}_{A_j}(t)\frac{f}{f_0}(F_0^{-1}(t)) + \mathbb{I}_{A_j^C}(t)\frac{1-p}{1-p_0} - 1\right)^2 \\ &= \int_{A_j} \left|\frac{f}{f_0}(F_0^{-1}(t)) - 1\right|^2 + \int_{A_j^c} \left|1 - \frac{1-p}{1-p_0}\right|^2 \\ &\leq Ck_n^{-1}\gamma_n^2 + Ck_n^{-2}\gamma_n^2 \lesssim k_n^{-1}\gamma_n^2\end{aligned}$$

with the bound of first term comes from the definition of $f \in \Sigma_n(f_0)$, and the bound of second term comes from the proof of point (2).

5. It suffices to show that, for all $h \in (0, 1/2)$,

$$\frac{1}{h}\int_h^{1-h} (\lambda_{f,f_0,A_j}(t+h) - \lambda_{f,f_0,A_j}(t))^2 \leq C\gamma_n \tag{4.64}$$

uniformly in h. Let $A_{1,j,h} = [a_{j,1} + h, a_{j,2} - h)$ (which can be empty if $h > \mathsf{Leb}(A_j)/2$), $A_{2,j,h} = [a_{j,1} - h, a_{j,2} + h) \cap [h, 1-h]$. We break the above integrals into a few parts:

- If $t \in [h, 1-h] \setminus A_{2,j,h}$ then $[t, t+h] \subseteq A_j^c$, so

$$\frac{1}{h}\int_{[h,1-h]\setminus A_{2,j,h}} (\lambda_{f,f_0,A_j}(t+h) - \lambda_{f,f_0,A_j}(t))^2 = 0 \tag{4.65}$$

- $t \in A_{2,j,h} \setminus A_{1,j,h}$ then either $t \in A_j^c$ and $t + h \in A_j$ or $t \in A_j$ and $t + h \in A_j^c$, so the function λ_{f,f_0,A_j} experiences a jump in the interval $(t, t+h)$. We can, however, use the bound in point (1) and that $\mathsf{Leb}(A_{2,j,h} \setminus A_{1,j,h}) = 4h$ to show that

$$\frac{1}{h}\int_{A_{2,j,h}\setminus A_{1,j,h}} (\lambda_{f,f_0,A_j}(t+h) - \lambda_{f,f_0,A_j}(t))^2 \lesssim \gamma_n^2 \tag{4.66}$$

- $t \in A_{1,j,h}$. This means that the interval $(t, t+h) \subseteq A_{1,j}$, so by lemma 4.9 one have

$$\frac{1}{h}\int_{A_{1,j,h}} (\lambda_{f,f_0,A_j}(t+h) - \lambda_{f,f_0,A_j}(t))^2 \lesssim \frac{1}{h}\int_{A_{1,j,h}} h^{2\alpha} \leq h^{2\alpha-1}k_n^{-1}. \tag{4.67}$$

We have to refine our analysis since the right hand side is currently not of order γ_n. Of course, we can adopt the bound from point (1) to obtain

$$\frac{1}{h}\int_{A_{1,j,h}} (\lambda_{f,f_0,A_j}(t+h) - \lambda_{f,f_0,A_j}(t))^2 \lesssim \gamma_n^2 k_n^{-1}/h, \tag{4.68}$$

but if $h \ll \gamma_n^2 k_n^{-1}$ then the right hand side grows without bound. We therefore split into two cases: $h \leq \gamma_n$ and $h > \gamma_n$. Before doing so, let us combine the above two inequalities, so that we have

$$\frac{1}{h}\int_{A_{1,j,h}} (\lambda_{f,f_0,A_j}(t+h) - \lambda_{f,f_0,A_j}(t))^2 \lesssim \gamma_n^2 k_n^{-1} \min\left(\gamma_n^{-2}h^{2\alpha-1}, h^{-1}\right) \tag{4.69}$$

The ultimate goal is to make $k_n^{-1}\min\left(\gamma_n^{-2}h^{2\alpha-1}, h^{-1}\right) = O(1)$:

- If $h > \gamma_n$ then one have $h^{-1} \leq \gamma_n^{-1}$, so if we requires $k_n^{-1} = O(\gamma_n)$ then the above goal is satisfied.
- If $h \leq \gamma_n$, then by noting that $2\alpha - 1 > 0$ one have $\gamma_n^{-2}h^{2\alpha-1} \leq \gamma_n^{-1}$, so by imposing the same condition as above we reach our goal.

We show that if $k_n^{-1} = O(\gamma_n)$ then we have $\left\|\lambda_{f,f_0,A_j}\right\|_{H_2^{1/2}}^2 \lesssim \gamma_n^2$, completing the proof.

■

4.3.2 Breaking up $\mathsf{Expt}_{1,n}(f_0)$

We break up $\mathsf{Expt}_{1,n}(f_0)$ in a similar fashion. Denote that

$$\mathsf{Expt}_{1,n,j}(f_0) = \left(C^0([0,1]), \mathscr{B}(C^0([0,1])), \left\{\mathbb{P}_{1,n,f,f_0}^{(j)}\right\}_{f\in\Sigma_n(f_0)}\right) \tag{4.70}$$

with $\mathbb{P}_{1,n}^{(j)}$ being the distribution of

$$dy(t) = \mathbb{I}_{A_j}(t)(\lambda_{f,f_0}(t) + K(f_0\|f))\,dt + \frac{1}{\sqrt{n}}\,dW_t \tag{4.71}$$

Note that the decomposition is exact, i.e.

$$\Delta\left(\mathsf{Expt}_{1,n}(f_0), \otimes_{i=1}^{k_n}\mathsf{Expt}_{1,n,j}(f_0)\right) = 0 \tag{4.72}$$

We are one step away from using the KMT inequality to complete the proof. Introduce the intermediate experiment: $\overline{\mathsf{Expt}}_{1,n}(f_0) = \otimes_{i=1}^{k_n}\overline{\mathsf{Expt}}_{1,n,j}(f_0)$, where

$$\overline{\mathsf{Expt}}_{1,n,j}(f_0) = \left(C^0([0,1]), \mathscr{B}(C^0([0,1])), \left\{\overline{\mathbb{P}}_{1,n,f,f_0}^{(j)}\right\}_{f\in\Sigma_n(f_0)}\right) \tag{4.73}$$

and $\overline{\mathbb{P}}_{1,n}^{(j)}$ is the distribution of

$$dy(t) = \left(\lambda_{f,f_0,A_j}(t) + K(f_0\|f, A_j)\right)\,dt + \frac{1}{\sqrt{n}}\,dW_t, \quad K(f_0\|f, A) = -\int \lambda_{f,f_0,A}(t) \tag{4.74}$$

Lemma 4.14 We have, for k_n sufficiently large,

$$H(\mathsf{Expt}_{1,n,j}(f_0), \overline{\mathsf{Expt}}_{1,n,j}(f_0)) \leq Cn^{1/2}\gamma_n\sqrt{k_n^{-2} + k_n^{-1}\gamma_n^2} \tag{4.75}$$

Proof. We have shown that the required squared Hellinger distance is exactly

$$H^2(\mathsf{Expt}_{1,n,j}(f_0), \overline{\mathsf{Expt}}_{1,n,j}(f_0)) = 2\left(1 - \exp\left(-\frac{nD^2}{8}\right)\right) \tag{4.76}$$

where D^2 is defined as

$$D^2 = \int_0^1 \left[\lambda_{f,f_0,A_j}(t) + K(f_0\|f, A_j) - \mathbb{I}_{A_j}(t)\lambda_{f,f_0}(t) - \mathbb{I}_{A_j}(t)K(f_0\|f)\right]^2.$$

See the end of section 7.3 for proof. We are thus required to bound D^2 such that $D^2 \ll 1/n$. Observe that

$$\lambda_{f,f_0,A_j}(t) - \mathbb{I}_{A_j}(t)\lambda_{f,f_0}(t) = \mathbb{I}_{A_j^c} \ln \frac{1-p}{1-p_0}, \tag{4.77}$$

so the integral is upper bounded by (up to a constant)

$$\int_{A_j^c} \left(\ln \frac{1-p}{1-p_0}\right)^2 + [K(f_0\|f)]^2\mathsf{Leb}(A_j) + [K(f_0\|f, A_j)]^2.$$

By point (2) from lemma 4.13, we know that first term is bounded by $k_n^{-2}\gamma_n^2$. The third term is bounded by $k_n^{-2}\gamma_n^4$. We finally make use of the arguments in point (4) to show that the second term is bounded by $k_n^{-1}\gamma_n^4$. We therefore see that the second term have much less contribution than the other terms, and that

$$D^2 \lesssim k_n^{-2}\gamma_n^2 + k_n^{-1}\gamma_n^4 = \gamma_n^2(k_n^{-2} + k_n^{-1}\gamma_n^2) \tag{4.78}$$

If we require $k_n^{-1} = O(\gamma_n)$ as in lemma 4.13, then we see that $D^2 \ll 1$, and so we can Taylor expand the $\exp(.)$ to conclude that

$$H^2(\mathsf{Expt}_{1,n,j}(f_0), \overline{\mathsf{Expt}}_{1,n,j}(f_0)) \lesssim n\gamma_n^2(k_n^{-2} + k_n^{-1}\gamma_n^2) \tag{4.79}$$

■

We are now ready to utilise the KMT inequality in theorem 4.8. Recall the forms of $\overline{\mathsf{Expt}}_{0,n,j}(f_0)$ and $\overline{\mathsf{Expt}}_{1,n,j}(f_0)$. By going through the setups in section 4.1 and utilise the KMT, we may construct intermediate experiments $\mathsf{Expt}^*_{0,n,j}(f_0)$ and $\mathsf{Expt}^*_{1,n,j}(f_0)$ on common measure space $(\Omega, \mathscr{F})$, so that the family of measures for each experiments can be represented as $\left\{\Lambda^{(j)*}_{0,n}(f; f_0)\, d\mathbb{P}^*\right\}_{f\in\Sigma_n(f_0)}$ and $\left\{\Lambda^{(j)*}_{1,n}(f; f_0)\, d\mathbb{P}^*\right\}_{f\in\Sigma_n(f_0)}$, where

$$\Lambda^{(j)*}_{0,n}(f; f_0) = \exp\left(\sqrt{n}\mathbb{U}_n(\lambda_{f,f_0,A_j}) - nK(f_0\|f, A_j)\right)$$
$$\Lambda^{(j)*}_{1,n}(f; f_0) = \exp\left(\sqrt{n}\mathbb{B}(\lambda_{f,f_0,A_j}) - \frac{n}{2}\int_0^1 \left(\lambda_{f,f_0,A_j}(s) + K(f_0\|f, A_j)\right)^2 ds\right)$$

where $\mathbb{U}_n, \mathbb{B}$ are uniform processes and Brownian bridge as specified in the statement of the KMT inequality. Moreover, there exists universal constant C such that

$$\mathbb{P}^*\left[\sqrt{n}|\mathbb{U}_n(g) - \mathbb{B}(g)| \geq C(\|g\|_{L^\infty} + \|g\|_{H_2^{1/2}})(t + \ln n)(\ln n)^{1/2}\right] \leq C\exp(-t) \tag{4.80}$$

with $g = \lambda_{f,f_0,A_j}$. Collecting the properties from lemma 4.13, we see that

$$\mathbb{P}^*\left[\sqrt{n}|\mathbb{U}_n(g) - \mathbb{B}(g)| \geq C\gamma_n(t + \ln n)(\ln n)^{1/2}\right] \leq C\exp(-t). \tag{4.81}$$

Choosing $t = k \ln n$ for some $k > 0$, we see that

$$\mathbb{P}^* \left[\sqrt{n} |\mathbb{U}_n(g) - \mathbb{B}(g)| \geq (k+1) C \gamma_n (\ln n)^{3/2} \right] \leq C/n^k \tag{4.82}$$

Comparing with the crude estimate in our first attempt, we have introduced a factor of γ_n. As we will see in our later discussions, this will provide us a better bound for the Hellinger distances. Before reaching our final steps, let us note the following lemma

Lemma 4.15 For k_n of sufficiently large order, there exists a constant $C > 0$ such that $\mathbb{E}^*[\lambda_{i,n}^{(j)^*}(f; f_0)^2] \leq C$ for $i = 0, 1$, where $\mathbb{E}^*$ represents the expectation with respect to $\mathbb{P}^*$. In addition, if we introduce the intermediate random variable

$$\Lambda_{\bullet,n}^{(j)^*}(f; f_0) = \exp\left(\sqrt{n} \mathbb{B}(\lambda_{f,f_0,A_j}) - nK(f_0 \| f, A_j) \right), \tag{4.83}$$

then we also have $\mathbb{E}^*[\lambda_{\bullet,n}^{(j)^*}(f; f_0)^2] \leq C$.

Proof. Let us show this for $i = 1$. For simplicity we write $D^2 = n \int_0^1 (\lambda_{f,f_0,A_j}(s) + K(f_0 \| f, A_j))^2$. Notice that the random variable $\sqrt{n}\mathbb{B}(\lambda_{f,f_0,A,})$ follows a Gaussian distribution with zero mean and variance D^2. As a result, one have

$$\begin{aligned}
\mathbb{E}^* \left[\Lambda_{1,n}^{(j)^*}(f; f_0) \right]^2 &= \int_{\mathbb{R}} \exp\left(2\xi - D^2\right) \frac{1}{\sqrt{2\pi D}} \exp\left(-\frac{\xi^2}{2D^2} \right) \\
&= \exp\left(D^2\right) \\
&\leq \exp\left(2n \left(\int_0^1 \lambda_{f,f_0,A_j}^2(s)\, ds + (K(f_0 \| f, A_j)^2) \right) \right)
\end{aligned}$$

By points (3) and (4) of lemma 4.13, we have

$$\mathbb{E}^* \left[\Lambda_{1,n}^{(j)^*}(f; f_0) \right]^2 \leq \exp\left(2n \left(C k_n^{-1} \gamma_n^2 + C k_n^{-2} \gamma_n^4 \right) \right).$$

It is noted that the assumption $k_n^{-1} = O(\gamma_n)$ is not enough for our analysis. One would now $k_n^{-1}\gamma_n^2 = O(n^{-1})$ to bound the expectation, which is equivalent to $k_n^{-1} = O(n^{-1/2}(\ln n)^2)$. Once we have that, then we know that the above expectation is bounded by a uniform constant C.

Similar arguments are also used to bound $\mathbb{E}^*[\lambda_{\bullet,n}^{(j)^*}(f; f_0)^2]$: note that

$$\mathbb{E}^*[\Lambda_{\bullet,n}^{(j)^*}(f; f_0)^2] = \exp\left(D^2 + D^2 - 2n(K(f_0 \| f, A_j)) \right) \leq \exp\left(Cn \left(k_n^{-1}\gamma_n^2 + k_n^{-2}\gamma_n^4 \right) \right) \tag{4.84}$$

noting that $|K(f_0 \| f, A_j)| \leq C k_n^{-1} \gamma_n^2$ by point (4), so with the condition as above we see that the expectation is bounded unfiromly in n.

Let us now show this for $i = 0$. Observe if we look back to the experiment $\overline{\mathsf{Expt}}_{0,n,j}(f_0)$, then we have the

$$\begin{aligned}
\mathbb{E}^* \left[\Lambda_{0,n}^{(j)^*}(f; f_0) \right]^2 &= \mathbb{E}\left(\exp\left(2 \sum_{i=1}^n \lambda_{f,f_0,A}(z_i) \right) \right) \\
&= \left(\int_0^1 \left(\exp(\lambda_{f,f_0,A}) \right)^2 \right)^n \\
&= \left(\int_0^1 \left[(\exp(\lambda_{f,f_0,A}) - 1)^2 + 2\exp(\lambda_{f,f_0,A}) - 1 \right] \right)^n \\
&\leq (1 + C k_n^{-1} \gamma_n^2)^n
\end{aligned}$$

using arguments from point (4). With the condition that $k_n^{-1}\gamma_n^2 = O(n^{-1})$ we see that the expectation is bounded, noticing the important inequality that $(1 + C/n)^n \leq \exp(C)$. ∎

With the above lemma, we are able to show the following

Lemma 4.16 — Application of KMT. We have

$$H^2(\overline{\mathsf{Expt}}_{0,n,j}(f_0), \overline{\mathsf{Expt}}_{1,n,j}(f_0)) \lesssim \gamma_n^2 (\ln n)^3 \tag{4.85}$$

for sufficiently large k_n, independent of j.

Proof. We work with the intermediate experiments with same probability space. Note that

$$\begin{aligned}
&H^2(\overline{\mathsf{Expt}}_{0,n,j}(f_0), \overline{\mathsf{Expt}}_{1,n,j}(f_0)) \\
&= \mathbb{E}\left[\left(\Lambda_{0,n}^{(j)^*}(f;f_0)\right)^{1/2} - \left(\Lambda_{1,n}^{(j)^*}(f;f_0)\right)^{1/2}\right]^2 \\
&\le 2\left(\mathbb{E}\left[\left(\Lambda_{0,n}^{(j)^*}(f;f_0)\right)^{1/2} - \left(\Lambda_{\bullet,n}^{(j)^*}(f;f_0)\right)^{1/2}\right]^2 + \mathbb{E}\left[\left(\Lambda_{\bullet,n}^{(j)^*}(f;f_0)\right)^{1/2} - \left(\Lambda_{1,n}^{(j)^*}(f;f_0)\right)^{1/2}\right]^2\right)
\end{aligned}$$

We show that the second term is bounded by the first term. Note that

$$\mathbb{E}^*\left[\Lambda_{\bullet,n}^{(j)^*}(f;f_0)\right] = \exp\left(D^2/2 - nK(f_0\|f, A_j)\right), \tag{4.86}$$

where D^2 is as defined in the previous lemma, we see that the following holds:

$$\left(\Lambda_{1,n}^{(j)^*}(f;f_0)\right)^{1/2} = \left(\Lambda_{\bullet,n}^{(j)^*}(f;f_0)\right)^{1/2}\left[\mathbb{E}^*\left[\Lambda_{\bullet,n}^{(j)^*}(f;f_0)\right]\right]^{-1/2} \tag{4.87}$$

From the previous lemma, we know that $\left(\Lambda_{0,n}^{(j)^*}\right)^{1/2}$, $\left(\Lambda_{1,n}^{(j)^*}\right)^{1/2}$ and $\left(\Lambda_{\bullet,n}^{(j)^*}\right)^{1/2}$ are all $L^2(\Omega^*, \mathscr{F}^*, \mathbb{P}^*)$ function with $\Lambda_{0,n}^{(j)^*}, \Lambda_{1,n}^{(j)^*}$ lying on the unit ball of L^2 space. Equation (4.87) indicates that the $\left(\Lambda_{1,n}^{(j)^*}\right)^{1/2}$ is the L^2 projection of $\left(\Lambda_{\bullet,n}^{(j)^*}\right)^{1/2}$ on the unit ball, so we know that the squared-L^2 distance between $\left(\Lambda_{1,n}^{(j)^*}\right)^{1/2}$ and $\left(\Lambda_{\bullet,n}^{(j)^*}\right)^{1/2}$ is no greater than that between $\left(\Lambda_{0,n}^{(j)^*}\right)^{1/2}$ and $\left(\Lambda_{\bullet,n}^{(j)^*}\right)^{1/2}$. We conclude that

$$H^2(\overline{\mathsf{Expt}}_{0,n,j}(f_0), \overline{\mathsf{Expt}}_{1,n,j}(f_0)) \lesssim \mathbb{E}^*\left[\left(\Lambda_{0,n}^{(j)^*}(f;f_0)\right)^{1/2} - \left(\Lambda_{\bullet,n}^{(j)^*}(f;f_0)\right)^{1/2}\right]^2. \tag{4.88}$$

We break up the expectation into two parts. Denote B being the event

$$B = \left\{\sqrt{n}|\mathbb{U}_n(\lambda_{f,f_0,A_j}) - \mathbb{B}(\lambda_{f,f_0,A_j})| \le (k+1)C\gamma_n(\ln n)^{3/2}\right\}, \tag{4.89}$$

where $k > 0$ is a constant to be determined, then we know that $\mathbb{P}^*(B^c) \lesssim 1/n^k$ from the KMT inequality. Therefore, one have

$$\begin{aligned}
\mathbb{E}^*\left[\left(\Lambda_{0,n}^{(j)^*}(f;f_0)\right)^{1/2} - \left(\Lambda_{\bullet,n}^{(j)^*}(f;f_0)\right)^{1/2}\right]^2 \mathbb{I}_{B^c} &\lesssim \mathbb{E}\left[\Lambda_{0,n}^{(j)^*}(f;f_0)\right] - \Lambda_{\bullet,n}^{(j)^*}(f;f_0)^2\mathbb{I}_{B^c} \\
&\overset{\text{(Holder)}}{\le} \left(\mathbb{P}^*(B^c)\mathbb{E}^*\left[\left(\Lambda_{0,n}^{(j)^*}(f;f_0)\right)^2 + \left(\Lambda_{\bullet,n}^{(j)^*}(f;f_0)\right)^2\right]\right)^{1/2} \\
&\lesssim n^{-k/2}.
\end{aligned}$$

On the other hand, we know that on the event B, one have

$$\begin{aligned}
\left(\Lambda_{0,n}^{(j)^*}(f;f_0)\right)^{1/2} - \left(\Lambda_{\bullet,n}^{(j)^*}(f;f_0)\right)^{1/2} &= \left(\Lambda_{0,n}^{(j)^*}(f;f_0)\right)^{1/2}\left(1 - \exp\left(\sqrt{n}(\mathbb{B}(\lambda_{f,f_0,A_j}) - \mathbb{U}_n(\lambda_{f,f_0,A_j}))\right)\right) \\
&\le \left(\Lambda_{0,n}^{(j)^*}(f;f_0)\right)^{1/2}\left(1 - \exp\left((k+1)C\gamma_n(\ln n)^{3/2}\right)\right)
\end{aligned}$$

$$\lesssim \left(\Lambda_{0,n}^{(j)^*}(f;f_0)\right)^{1/2}\left((k+1)\gamma_n(\ln n)^{3/2}\right).$$

and therefore

$$\begin{aligned}\mathbb{E}^*\left[\left(\Lambda_{0,n}^{(j)^*}(f;f_0)\right)^{1/2}-\left(\Lambda_{\bullet,n}^{(j)^*}(f;f_0)\right)^{1/2}\right]^2\mathbb{I}_B &\lesssim (k+1)^2\mathbb{E}^*\left[\Lambda_{0,n}^{(j)^*}(f;f_0)\right]\gamma_n^2(\ln n)^3\\ &= (k+1)^2\gamma_n^2(\ln n)^3\\ &= (k+1)^2 n^{-1/2}\ln n.\end{aligned}$$

Selecting any $k>1$ completes the proof. ■

Let us finally collect all the estimates: we know from lemma 4.14 that, whenever $k_n^{-1}=O(\gamma_n)$, then

$$H^2(\mathsf{Expt}_{1,n,j}(f_0),\overline{\mathsf{Expt}}_{1,n,j}(f_0))\lesssim n\gamma_n^2\left(k_n^{-2}+k_n^{-1}\gamma_n^2\right). \tag{4.90}$$

Moreover, if $k_n^{-1}=O(n^{-1}\gamma_n^{-2})=O(n^{-1/2}(\ln n)^2)$, then

$$H^2(\overline{\mathsf{Expt}}_{0,n,j}(f_0),\overline{\mathsf{Expt}}_{1,n,j}(f_0))\lesssim \gamma_n^2(\ln n)^3. \tag{4.91}$$

From lemma therefore have

$$\begin{aligned}&\left(\Delta\left(\bigotimes_{i=1}^{k_n}\mathsf{Expt}_{0,n,j}(f_0),\bigotimes_{i=1}^{k_n}\mathsf{Expt}_{1,n,j}(f_0)\right)\right)^2\\ &=\left(\Delta\left(\bigotimes_{i=1}^{k_n}\overline{\mathsf{Expt}}_{0,n,j}(f_0),\bigotimes_{i=1}^{k_n}\mathsf{Expt}_{1,n,j}(f_0)\right)\right)^2\\ &\lesssim H^2\left(\bigotimes_{i=1}^{k_n}\overline{\mathsf{Expt}}_{0,n,j}(f_0),\bigotimes_{i=1}^{k_n}\mathsf{Expt}_{1,n,j}(f_0)\right)\\ &\lesssim \sum_{i=1}^{k_n}H^2(\overline{\mathsf{Expt}}_{0,n,j}(f_0),\mathsf{Expt}_{1,n,j}(f_0))\\ &\le \sum_{i=1}^{k_n}H^2(\overline{\mathsf{Expt}}_{0,n,j}(f_0),\overline{\mathsf{Expt}}_{1,n,j}(f_0))+H^2(\overline{\mathsf{Expt}}_{0,n,j}(f_0),\overline{\mathsf{Expt}}_{1,n,j}(f_0))\\ &\lesssim k_n\gamma_n^2(\ln n)^3+nk_n^{-1}\gamma_n^2+n\gamma_n^4\\ &= k_n(n^{-1/2}\ln n)+nk_n^{-1}\gamma_n^2+n\gamma_n^4\end{aligned}$$

Note that $n\gamma_n^4=1/\ln n=o(1)$, so this term will not cause us trouble. For the Le Cam distance between two product experiments to be $o(1)$, we also need $k_n=o(n^{1/2}(\ln n)^{-1})$. With the guess of $k_n\sim n^\alpha(\ln n)^{-\beta}$, we need $\alpha=1/2$ and $1<\beta<2$ for k_n to satisfy both decaying conditions for the Le Cam distance. Choosing $\beta=3/2$ yields the optimal bound. To sum up

Proposition 4.17 With the choice of $k_n\sim n^{1/2}(\ln n)^{-\beta}$ with $\beta\in(1,2)$, we see that

$$\left(\Delta\left(\bigotimes_{i=1}^{k_n}\mathsf{Expt}_{0,n,j}(f_0),\bigotimes_{i=1}^{k_n}\mathsf{Expt}_{1,n,j}(f_0)\right)\right)^2=o(1). \tag{4.92}$$

For the choice of $\beta=3/2$, the right hand side is of $O((\ln n)^{-1/2})$.

Stepping back, if we directly apply the KMT inequality by following the steps of 4.16 without using the divide-and-conquer trick, we would arrive at

$$H^2(\mathsf{Expt}_{0,n}^*(f_0),\mathsf{Expt}_{0,n}^*(f_0))\lesssim n^{-k/2}+\exp\left(C(k+1)(\ln n)^{3/2}\right). \tag{4.93}$$

The right hand side diverges as $n\to\infty$, so the bound is meaningless. We therefore need the divide-and-conquer trick to improve our bound.

4.4 Poissonisation

This section is dedicated to proving that

$$\Delta\left(\mathsf{Expt}_{0,n}(f_0), \bigotimes_{i=1}^{k_n} \mathsf{Expt}_{0,n,j}(f_0)\right) \overset{n\to\infty}{\to} 0. \tag{4.94}$$

We begin by recalling the following characterisation of $\mathsf{Expt}_{0,n}(f_0)$:

$$\overline{\mathsf{Expt}}_{0,n}(f_0) = ([0,1]^n, \mathscr{B}([0,1]^n), \{d\overline{\mathbb{P}}_{0,n,f,f_0} := \overline{\Lambda}_{0,n}(f;f_0)(z)\, d\mathsf{Leb}^{\otimes n}\}_{f\in\Sigma_n(f_0)}) \tag{4.95}$$

$$[\overline{\Lambda}_{0,n}(f;f_0)](z) = \exp\left(\sum_{i=1}^n \lambda_{f,f_0}(z_i)\right) = \exp\left(\int_0^1 \lambda_{f,f_0}(t)[\mu_n(z)](dt)\right). \tag{4.96}$$

For computational convenience, we apply a change of random variable $z \mapsto \mu_n(z) := \sum_{i=1}^n \delta_{z_i}$, so that $\mu_n(z)$ is an unnormalised empirical measure of Leb, and that $\mathsf{Expt}_{0,n}(f_0)$ is equivalent with the following experiments

$$\mathsf{Expt}_{0,n}(f_0) = (\mathscr{M}_n, \mathscr{F}_n, \{\Pi_{0,n,f,f_0}\}_{f\in\Sigma_n(f_0)}), \tag{4.97}$$

where $\mathscr{M}_n$, $\mathscr{F}_n$ is as specified in (4.30) (noting that any sets in the σ-algebra generated by the uniform process $U_n := \sqrt{n}(P_n - \mathsf{Leb})$ is also in the σ-algebra generated by μ_n), and that

$$d\Pi_{0,n,f,f_0} = [\Lambda_{0,n}(f;f_0)](\mu)\, \mu_n^* \mathsf{Leb}^{\otimes n}(d\mu), \tag{4.98}$$

$$[\Lambda_{0,n}(f;f_0)](\mu) = \exp\left(\int_0^1 \lambda_{f,f_0}(t)\mu(dt)\right), \quad \mu \in \mathscr{M}_n. \tag{4.99}$$

We enlarge the sample space to $(\mathscr{M}_\infty, \mathscr{F}_\infty)$, the smallest σ-algebra containing $(\mathscr{M}_n, \mathscr{F}_n)$ for all $n \geq 1$. We can clearly construct an experiment on this sample space which is equivalent to the $\mathsf{Expt}_{0,n}(f_0)$, and we will abuse notation and call it $\mathsf{Expt}_{0,n}(f_0)$ as well.

Consider the following experiment

$$\mathsf{Expt}_{*,n}(f_0) = (\mathscr{M}_\infty, \mathscr{F}_\infty, \{\Pi_{*,n,f,f_0}\}_{f\in\Sigma_n(f_0)}), \tag{4.100}$$

where $\Lambda_{0,n}$ is as specified above,

$$d\Pi_{*,n,f,f_0} = [\Lambda_{0,n}(f;f_0)](\mu)\, \mu_\nu^* \mathsf{Leb}^{\otimes\infty}(d\mu); \tag{4.101}$$

and $\mu_\nu^*\mathsf{Leb}^{\otimes\infty}$ the distribution of the following random measure from $([0,1], \mathscr{B}[0,1], \mathsf{Leb})^{\otimes\infty}$:

$$[\mu_\nu(z)](B) = \sum_{i=1}^{\nu(z)} \delta_{z_i}(B), \tag{4.102}$$

with $\nu : ([0,1], \mathscr{B}[0,1]) \to \mathbb{Z}_{\geq 0}$ is a random variable independent of all projection maps $z \mapsto z_i$ that has a Poisson distribution $\mathsf{Po}(n)$. One should note that the $\mu_\nu(z)$ is a Poisson process (with intensity $n \times \mathsf{Leb}$) in the following sense (see [7] for details):

- Let A, B be disjoint sets of $[0,1]$, then the random variables $[\mu_\nu(z)](A)$ and $[\mu_\nu(z)](B)$ are independent.
- $[\mu_\nu(z)](A) \sim \mathsf{Po}(n\mathsf{Leb}(A))$.

We note that $\mathsf{Expt}_{0,n}(f_0)$ and $\mathsf{Expt}_{*,n}(f_0)$ are asymptotically equivalent:

Lemma 4.18 — Asymptotic equivalence of $\mathsf{Expt}_{0,n}(f_0)$ and $\mathsf{Expt}_{*,n}(f_0)$.

$$H^2(\mathsf{Expt}_{0,n}(f_0), \mathsf{Expt}_{*,n}(f_0)) = O(n^{1/2}\gamma_n^2) = O((\ln n)^{-2}). \tag{4.103}$$

Proof. To show this, let us take expectation with respect to $\mathsf{Leb}^{\otimes\infty}$ over $([0,1], \mathscr{B}[0,1])^{\otimes\infty}$:

$$H^2(\mathsf{Expt}_{0,n}(f_0), \mathsf{Expt}_{*,n}(f_0)) = \mathbb{E}\left[\exp\left(\frac{1}{2}\sum_{i=1}^{n}\lambda_{f,f_0}(z_i)\right) - \exp\left(\frac{1}{2}\sum_{i=1}^{\nu}(z)\lambda_{f,f_0}(z_i)\right)\right]^2.$$

Let $\nu_1 = \min(\nu, n)$ and $\nu_2 = \max(\nu, n)$. We consider the conditional expectation given ν, which can be expressed as a measurable function on ν. Since the projection functions $z \mapsto z_i$ are independent, we have the following conditional independence:

$$\begin{aligned}
&\mathbb{E}\left[\left(\exp\left(\frac{1}{2}\sum_{i=1}^{n}\lambda_{f,f_0}(z_i)\right) - \exp\left(\frac{1}{2}\sum_{i=1}^{\nu}\lambda_{f,f_0}(z_i)\right)\right)^2 \,\middle|\, \nu\right] \\
&= \underbrace{\mathbb{E}\left[\exp\left(\sum_{i=1}^{\nu_1}\lambda_{f,f_0}(z_i)\right) \,\middle|\, \nu\right]}_{=1} \mathbb{E}\left[\left(\exp\left(\frac{1}{2}\sum_{i=\nu_1+1}^{\nu_2}\lambda_{f,f_0}(z_i)\right) - 1\right)^2 \,\middle|\, \nu\right].
\end{aligned}$$

We can make use of the arguments in proving lemma 7.6 to show that this quantity is bounded above by (up to a constant)

$$|\nu_2 - \nu_1| \int_0^1 \left[\left(\frac{f^{1/2}}{f_0^{1/2}}(F_0^{-1}(z)) - 1\right)\right]^2 dz \le |\nu - n| \int_0^1 \left[\left(\frac{f}{f_0}(F_0^{-1}(z)) - 1\right)\right]^2 dz \le |\nu - n|\gamma_n^2.$$

Finally, note that

$$\mathbb{E}[|\nu - n|] \le (\mathbb{E}[|\nu - n|^2])^{1/2} = n^{1/2}. \tag{4.104}$$

Therefore we know from law of iterated expectation that

$$H^2(\mathsf{Expt}_{0,n}(f_0), \mathsf{Expt}_{*,n}(f_0)) \le n^{1/2}\gamma_n^2. \tag{4.105}$$

■

The reason of introducing the "Poissonised" experiment $\mathsf{Expt}_{*,n}(f_0)$ is that it enjoys a thinning property (see [7]), which we state without proof: consider a family of the truncated version of $\mu_n(z)$ and $\mu_\nu(z)$:

$$\mu_{n,A_j}(B) = \mu_n(B \cap A_j), \tag{4.106}$$

$$\mu_{\nu,A_j}(B) = \mu_\nu(B \cap A_j), \quad A_j = F_0^{-1}([(j-1)/k_n, j/k_n)), j = 1, ..., k_n. \tag{4.107}$$

The thinning property says that

- μ_{ν,A_j} is also a Poisson process, but with intensity $n\mathsf{Leb}(. \cap A_j)$
- if $i \neq j$, then the random measures μ_{ν,D_i} and μ_{ν,A_j} are independent.

As a result, one can justify the following factorisation:

$$\Delta\left(\mathsf{Expt}_{*,n}(f_0), \bigotimes_{j=1}^{k_n}\mathsf{Expt}_{*,n,j}(f_0)\right) = 0, \tag{4.108}$$

where

$$\mathsf{Expt}_{*,n,j}(f_0) = (\mathscr{M}_\infty, \mathscr{F}_\infty, \{\Pi_{*,n,f,j}\}_{f\in\Sigma_n(f_0)}), \tag{4.109}$$

$$d\Pi_{0,n,f_j} = [\Lambda_{0,n}(f;f_0)](\mu)\,\mu^*_{\nu,A_j}\mathsf{Leb}^{\otimes\infty}(d\mu). \tag{4.110}$$

The proof follows a similar route of the proof of the decomposition of $\mathsf{Expt}_1(f_0)$ in (4.72), so we will not go through the details. Recall the following:

$$\overline{\mathsf{Expt}}_{0,n,j}(f_0) = \left([0,1]^n, \mathscr{B}([0,1]^n), \left\{d\overline{\mathbb{P}}^{(j)}_{0,n,f,f_0} := \overline{\Lambda}^{(j)}_{0,n}(f;f_0)(z)\,d\mathsf{Leb}^{\otimes n}\right\}_{f\in\Sigma_n(f_0)}\right), \tag{4.111}$$

$$[\overline{\Lambda}_{0,n}(f;f_0)](z) = \exp\left(\sum_{i=1}^n \lambda_{f,f_0,A_j}(z_i)\right), \tag{4.112}$$

$$\lambda_{f,f_0,A_j}(t) = \mathbb{I}_{A_j}(t)\ln\frac{f}{f_0}(F_0^{-1}(t)) + \mathbb{I}_{A_j^C}(t)\ln\frac{1-p}{1-p_0}. \tag{4.113}$$

Apply the change of variable we note that $z \mapsto \mu_n(z) := \sum_{i=1}^n \delta_{z_i}$, we have

$$\overline{\mathsf{Expt}}_{0,n,j}(f_0) = \left(\mathscr{M}_n, \mathscr{F}_n, \left\{d\overline{\mathbb{P}}^{(j)}_{0,n,f,f_0} := \overline{\Lambda}^{(j)}_{0,n}(f;f_0)(\mu)\,\mu_n^* d\mathsf{Leb}^{\otimes n}\right\}_{f\in\Sigma_n(f_0)}\right)) \tag{4.114}$$

$$[\overline{\Lambda}_{0,n}(f;f_0)](\mu) = \exp\left(\int_0^1 \left(\mathbb{I}_{A_j}(t)\ln\frac{f}{f_0}(F_0^{-1}(t)) + \mathbb{I}_{A_j^C}(t)\ln\frac{1-p}{1-p_0}\right)d\mu_n\right) \tag{4.115}$$

$$= \exp\left(\int_0^1 \ln\frac{f}{f_0}(F_0^{-1}(t))\mu_{n,A_j}(dt) + \underbrace{\ln\frac{1-p}{1-p_0}\mu_n(A_j^c)}_{=n-\mu_{n,A_j}([0,1])}\right). \tag{4.116}$$

So by another change of variable, the experiment is equivalent to

$$\overline{\mathsf{Expt}}_{0,n,j}(f_0) = \left(\mathscr{M}_n, \mathscr{F}_n, \left\{d\overline{\mathbb{P}}^{(j)}_{0,n,f,f_0} := \overline{\Lambda}^{(j)}_{0,n}(f;f_0)(\mu)\,\mu^*_{n,A_j} d\mathsf{Leb}^{\otimes n}\right\}_{f\in\Sigma_n(f_0)}\right), \tag{4.117}$$

with $[\overline{\Lambda}_{0,n}(f;f_0)](\mu)$ being defined as above. Finally, we state without prove the following result concerning truncated Poisson process:

Lemma 4.19 — Hellinger distance between truncated Poisson processes ((14), Theorem 2).

$$H(\overline{\mathsf{Expt}}_{0,n,j}(f_0), \overline{\mathsf{Expt}}_{*,n,j}(f_0)) \le \sqrt{3}\mathsf{Leb}(A_j). \tag{4.118}$$

We therefore have

$$\Delta\left(\bigotimes_{j=1}^{k_n}\mathsf{Expt}_{0,n,j}(f_0), \bigotimes_{j=1}^{k_n}\mathsf{Expt}_{*,n,j}(f_0)\right) \lesssim k_n\gamma_n^2 = O((\ln n)^{5/2}) = o(1). \tag{4.119}$$

To conclude,

Proposition 4.20 — Local Equivalence.

$$\Delta(\mathsf{Expt}_{0,n}(f_0), \mathsf{Expt}_{1,n}(f_0)) \overset{n\to\infty}{\to} 0. \tag{4.120}$$

4.5 Local equivalence for other experiments

Recall the definitions of the other experiments

$$\mathsf{Expt}_{i,n}(f_0) = (C^0([0,1]), \mathscr{B}(C^0([0,1])), \{\mathbb{P}_{i,n,f,f_0}\}_{f\in\Sigma_n(f_0)}), \quad i=1,2,3, \tag{4.121}$$

where the measures $\{\mathbb{P}_{i,n,f,f_0}\}$ are the distributions characterised by the following Gaussian process:

$$i=1; \quad dy(t) = (\lambda_{f,f_0}(t) + K(f_0\|f))\,dt + \frac{1}{\sqrt{n}}\,dW_t, \tag{4.122}$$

$$i = 2; \quad dy(t) = (f(t) - f_0(t))\, dt + \frac{\sqrt{f_0(t)}}{\sqrt{n}}\, dW_t, \tag{4.123}$$

$$i = 3; \quad dy(t) = \left(\sqrt{f(t)} - \sqrt{f_0(t)}\right)\, dt + \frac{1}{2\sqrt{n}} dW_t. \tag{4.124}$$

Let us note that $\mathbb{P}_{2,n,f,f_0} \ll \mathbb{P}_{2,n,f_0,f_0}$ with likelihood process

$$\Lambda_{2,n}(f; f_0) = \exp\left(\sqrt{n}\int_0^1 \frac{f(t) - f_0(t)}{\sqrt{f_0(t)}} d\mathsf{w}_t - \frac{n}{2}\int_0^1 \frac{(f(t) - f_0(t))^2}{f_0(t)}\, dt\right), \tag{4.125}$$

with respect to $Q_{2,n} := \mathbb{P}_{2,n,f_0,f_0}$, the distribution of the process $dy = \sqrt{f_0/n}\, dW$. Note that by applying a change of variable $t = F_0^{-1}(u)$, then $dt = du/f_0(F_0^{-1}(u))$. Therefore,

$$\int_0^1 \frac{(f(t) - f_0(t))^2}{f_0(t)}\, dt = \int_0^1 \left(\frac{f(F_0^{-1}(u))}{f_0(F_0^{-1}(u))} - 1\right)^2 du.$$

This inspires us to apply a change of time in evaluating the first term in the exponent. We recall the following identity: for all $t \in [0,1]$ and Brownian motion $\mathsf{w}(t) := \mathsf{w}_t$, we have

$$\int_0^t f_0^{1/2}(s)\, d\mathsf{w}_s = \mathsf{w}\left(\int_0^t (f^{1/2}(s))^2\, dt\right) = \mathsf{w}(F_0(t)). \tag{4.126}$$

This is an application of the Dambis-Dubmin-Schwarz theorem (see [15], Theorem 5.13), by noting that the left hand side is a continuous local martingale with quadratic variation $F_0(t)$. As a result, we know that for all continuous function on $[0,1]$,

$$\int_0^1 g(s) f_0^{1/2}(s)\, d\mathsf{w}(s) = \int_0^1 g(s) d\mathsf{w}(F_0(s)). \tag{4.127}$$

As a result, we know that $Q_{2,n}$ is the distribution of the process $dy = dW(F_0(t))/\sqrt{n}$, and that

$$\int_0^1 \frac{f(t) - f_0(t)}{\sqrt{f_0(t)}} d\mathsf{w}_t = \int_0^1 \left(\frac{f(t)}{f_0(t)} - 1\right) d\mathsf{w}(F_0(t)) = \int_0^1 \left(\frac{f(F_0^{-1}(u))}{f_0(F_0^{-1}(u))} - 1\right) d\mathsf{w}_u. \tag{4.128}$$

Finally, we apply a bijective change of variable by mapping a path $\mathsf{w}(t) \mapsto \mathsf{w}(F_0^{-1}(t))$. Then it is clear that the push forward measure of $Q_{2,n}$ by this change of variable is $Q_{1,n}$, the law of scaled Brownian motion $W/\sqrt{n}$. To sum up, $\mathsf{Expt}_{2,n}(f_0)$ is equivalent to the experiment

$$\left(C^0([0,1]), \mathscr{B}(C^0([0,1])), \left\{\tilde{\Lambda}_{2,n}(f; f_0)\, dQ_{1,n}\right\}_{f \in \Sigma_n(f_0)}\right). \tag{4.129}$$

where

$$\tilde{\Lambda}_{2,n}(f; f_0) = \exp\left(\sqrt{n}\int_0^1 \lambda_{2,f,f_0}(t) d\mathsf{w}_t - \frac{n}{2}\int_0^1 \lambda_{2,f,f_0}(t)^2\, dt\right), \tag{4.130}$$

$$\lambda_{2,f,f_0}(t) = \frac{f(F_0^{-1}(t))}{f_0(F_0^{-1}(t))} - 1. \tag{4.131}$$

We will also call it $\mathsf{Expt}_{2,n}(f_0)$. We can also perform similar calculations for $\mathsf{Expt}_{3,n}(f_0)$. Notice that $\mathbb{P}_{3,n,f,f_0} \ll \mathbb{P}_{3,n,f_0,f_0}$ with likelihood process

$$\Lambda_{3,n}(f; f_0) = \exp\left(\sqrt{n}\int_0^1 2\left(\sqrt{f(t)} - \sqrt{f_0(t)}\right) d\mathsf{w}_t - \frac{n}{2}\int_0^1 4(\sqrt{f(t)} - \sqrt{f_0(t)})^2\, dt\right) \tag{4.132}$$

$$= \exp\left(\sqrt{n}\int_0^1 2\left(\frac{\sqrt{f(F_0^{-1}(t))}}{\sqrt{f_0(F_0^{-1}(t))}} - 1\right) d\mathsf{w}(F_0(t)) - \frac{n}{2}\int_0^1 4\left(\frac{\sqrt{f(F_0^{-1}(t))}}{\sqrt{f_0(F_0^{-1}(t))}} - 1\right)^2 dt\right). \tag{4.133}$$

So we can use the above change of time to conclude that $\mathsf{Expt}_{3,n}(f_0)$ is equivalent to

$$\left(C^0([0,1]), \mathscr{B}(C^0([0,1])), \left\{\tilde{\Lambda}_{3,n}(f;f_0)dQ_{1,n}\right\}_{f\in\Sigma_n(f_0)}\right), \tag{4.134}$$

where

$$\tilde{\Lambda}_{3,n}(f;f_0) = \exp\left(\sqrt{n}\int_0^1 \lambda_{3,f,f_0}(t)d\mathsf{w}_t - \frac{n}{2}\int_0^1 \lambda_{3,f,f_0}(t)^2\,dt\right), \tag{4.135}$$

$$\lambda_{3,f,f_0}(t) = 2\left(\sqrt{\frac{f(F_0^{-1}(t))}{f_0(F_0^{-1}(t))}} - 1\right). \tag{4.136}$$

we will again call the experiment $\mathsf{Expt}_{3,n}(f_0)$. Defining $\lambda_{1,f,f_0}(t) := \lambda_{f,f_0}(t) + K(f_0\|f)$. We are now ready to prove that

Lemma 4.21 For $i = 2, 3$

$$H^2(\mathsf{Expt}_{1,n}(f_0), \mathsf{Expt}_{i,n}(f_0)) = o(n^{-1}). \tag{4.137}$$

Proof. We note that $\|\lambda_{i,f,f_0}\|_{L^2}^2$ are uniformly bounded in n for $i = 1, 2, 3$. The case for $i = 1$ is taken care by lemma 4.9. We can also see that for $i = 2$, $\|\lambda_{2,f,f_0}\|_{L^2}^2 \le \gamma_n^2$, and for $i = 3$ we have $\|\lambda_{3,f,f_0}\|_{L^2}^2 \le 4\gamma_n^2$. We therefore have, for $i = 2, 3$,

$$\begin{aligned} H^2(\mathsf{Expt}_{1,n}(f_0), \mathsf{Expt}_{i,n}(f_0)) &\lesssim \mathbb{E}\left[\exp\frac{1}{2}\left|\int_0^1 (\lambda_{1,f,f_0} - \lambda_{i,f,f_0})\,d\mathsf{w}_t\right| - 1\right]^2 \\ &\lesssim \int_0^1 (\lambda_{1,f,f_0} - \lambda_{i,f,f_0})^2 = \|\lambda_{1,f,f_0} - \lambda_{i,f,f_0}\|_{L^2}^2, \end{aligned}$$

provided that $\|\lambda_{1,f,f_0} - \lambda_{i,f,f_0}\|_{L^2}^2$ is small enough. We note that

$$\begin{aligned} \|\lambda_{1,f,f_0} - \lambda_{2,f,f_0}\|_{L^2}^2 &= \int_0^1 \left(\ln\left(\frac{f}{f_0}(F_0^{-1}(t))\right) - \left(\frac{f}{f_0}(F_0^{-1}(t)) - 1\right) + K(f_0\|f)\right)^2 \\ &\lesssim \int_0^1 \left(\left(\frac{f}{f_0}(F_0^{-1}(t)) - 1\right)^2 + K(f_0\|f)\right)^2 \\ &\lesssim \gamma_n^4 + K(f_0\|f)^2. \end{aligned}$$

With the spirit of proving point (4) of lemma 4.13, one can show that

$$\begin{aligned} -\int_0^1 \lambda_{f,f_0}(t)\,dt &= -\int_0^1 \ln\left(1 + \frac{f}{f_0}(F_0^{-1}(t)) - 1\right)\,dt \\ &\le \int_0^1 \underbrace{\left(\frac{f}{f_0}(F_0^{-1}(t)) - 1\right)}_{=0} + C\int_0^1 \left(\frac{f}{f_0}(F_0^{-1}(t)) - 1\right)^2 dt = C\gamma_n^2. \end{aligned}$$

We therefore conclude that

$$\|\lambda_{1,f,f_0} - \lambda_{2,f,f_0}\|_{L^2}^2 \lesssim \gamma_n^4 = O(n^{-1}(\ln n)^{-4}) = o(n^{-1}). \tag{4.138}$$

For $i = 3$, we note that for all $t \in [0,1]$

$$\lambda_{f,f_0} = 2\ln\left(1 + \sqrt{\frac{f}{f_0}(F_0^{-1}(t)) - 1}\right)$$

$$\leq \lambda_{3,f,f_0} + C\left(\sqrt{\frac{f}{f_0}(F_0^{-1}(t))} - 1\right)^2$$
$$\leq \lambda_{3,f,f_0} + C\gamma_n^2.$$

We finally conclude that

$$\|\lambda_{1,f,f_0} - \lambda_{3,f,f_0}\|_{L^2}^2 \lesssim \gamma_n^4 = o(n^{-1}). \tag{4.139}$$

■

So, with the above efforts, we have proven theorem 4.2

5 Global Equivalence

5.1 Proposal Estimator

We begin by recalling the heuristics for establishing asymptotic equivalence of $\text{Expt}_{0,n}$ and $\text{Expt}_{1,n}$ on parameter space $\Sigma := \Sigma_{\alpha,M,\epsilon}$, as proposed in [3]. The first step is to construct "preliminary estimators" based on a fraction of sample $y_1, ..., y_{N_n}$ in Expt_0 with $n \gg N_n \gg 1$, say $\hat{f}_n := \hat{f}_n(y_1, ..., y_{N_n})$, such that

$$\mathbb{P}_{0,n}(f \in \Sigma_n(\hat{f}_n)) = \mathbb{P}_{0,n}\left(\left\|\frac{\hat{f}_n}{f} - 1\right\|_\infty \geq \gamma_n\right) \overset{n\to\infty}{\to} 0. \tag{5.1}$$

To show that, let us recall from the discussions in chapter 2 that there is a sequence of estimators $\tilde{f}_n \in \Sigma$ such that

$$\sup_{f\in\Sigma} \mathbb{P}_{0,n,f}\left(\left\|\tilde{f}_n - f\right\|_\infty \geq C\psi_n\right) \to 0, \tag{5.2}$$

where C is a universal constant and $\psi_n = (\ln n/n)^{\alpha/(2\alpha+1)}$. In particular, setting $n/\ln n \leq N_n \leq n/2$ yields

$$\sup_{f\in\Sigma} \mathbb{P}_{0,n,f}\left(\left\|\tilde{f}_{N_n} - f\right\|_\infty \geq C\psi_{N_n}\right) \to 0, \tag{5.3}$$

with $\psi_{N_n} = O\left((\ln n/n)\right)^{\alpha/(2\alpha+1)} = o(\gamma_n)$, owing to α being strictly greater than $1/2$. This implies that

$$\sup_{f\in\Sigma} \mathbb{P}_{0,n,f}\left(\left\|\tilde{f}_{N_n} - f\right\|_\infty > c\gamma_n\right) \leq \sup_{f\in\Sigma} \mathbb{P}_{0,n,f}\left(\left\|\tilde{f}_{N_n} - f\right\|_\infty > C\psi_n\right) \to 0, \forall c > 0. \tag{5.4}$$

Now since Σ is a bounded and equicontinuous subset of $C^0[0,1]$, it is compact by the Arzela-Ascoli theorem, and one can introduce the $\epsilon\gamma_n/2$-net, $\Sigma_{0,n}$, which is a finite set (with size depending on n) such that if $g \in \Sigma$ then there exists $\tilde{g}_n$ such that $\|g - g_n\|_\infty < \epsilon\gamma_n/2$. Let $\hat{f}_n$ be the closest elements in $\Sigma_{0,n}$ with $\tilde{f}_{N_n}$, then

$$\left\|\hat{f}_n - f\right\|_\infty \leq \left\|\hat{f}_n - \tilde{f}_{N_n}\right\|_\infty + \left\|\hat{f}_{N_n} - f\right\|_\infty \leq \epsilon\gamma_n/2 + \left\|\hat{f}_{N_n} - f\right\|_\infty. \tag{5.5}$$

Therefore if $\left\|f/\hat{f}_n - f\right\|_\infty > \gamma_n$, then we know that $\|f - f_n\|_\infty > \epsilon\gamma_n$. As a result, one have $\left\|\tilde{f}_{N_n} - f\right\|_\infty > \epsilon\gamma_n/2$. Therefore

$$\sup_{f\in\Sigma} \mathbb{P}_{0,n}\left(\left\|\frac{f}{\hat{f}_n} - 1\right\|_\infty \geq \gamma_n\right) \leq \sup_{f\in\Sigma} \mathbb{P}_{0,n,f}\left(\left\|\tilde{f}_{N_n} - f\right\|_\infty > \epsilon\gamma_n/2\right) \overset{n\to\infty}{\to} 0.$$

Moreover there are only finitely many choices for $\hat{f}_n$ (the number of choices, of course, grows as n increases). We may also construct a similar sequence of preliminary estimators for Expt_1. To begin, since if $f \in \Sigma_{\alpha,M,\epsilon}$ then $f^{1/2} \in \Sigma_{\alpha,M\epsilon^{1/2},\epsilon^{1/2}}$, so there exists a sequence $\check{f}_n \in \Sigma$ such that

$$\sup_{f\in\Sigma} \mathbb{P}_{1,n,f}\left(\left\|\tilde{f}_n - f\right\|_\infty \geq C\psi_n\right) \to 0. \tag{5.6}$$

Of course, we note that $\check{f}_{n-N_n}$ also satisfies the above condition, so we may apply a compactness argument similar to above to conclude that there exists sequence of estimators $\check{f}_n \in \Sigma$ such that

$$\sup_{f\in\Sigma} \mathbb{P}_{1,n}\left(\left\|\frac{\check{f}_n}{f} - 1\right\|_\infty \geq \gamma_n\right) \overset{n\to\infty}{\to} 0. \tag{5.7}$$

and that $\check{f}_n$ only takes finitely many values in Σ.

5.2 Completing the proof by product experiments

We then formulate the idea of using the local equivalence result on the experiment $\mathsf{Expt}_{i,n}(\hat{f})$, viewed as one of the local experiments $\mathsf{Expt}_{i,n}(f_0)$. Specifically, we write $\mathsf{Expt}_{i,n}(\hat{f})$ in the form

$$\mathsf{Expt}_{i,n}(\hat{f}) = \left(\mathscr{Y} \otimes \mathscr{Y}_i, \mathscr{F} \otimes \mathscr{F}_i, \left\{\hat{\mathbb{P}}_{i,n,f}\right\}_{f\in\Sigma}\right), \tag{5.8}$$

where $(\mathscr{Y},\mathscr{F}) = ([0,1],\mathscr{B}[0,1])^{\otimes N_n}$, $(\mathscr{Y}_0,\mathscr{F}_0) = ([0,1],\mathscr{B}[0,1])^{\otimes n-N_n}$, $(\mathscr{Y}_i,\mathscr{F}_i) = (C^0[0,1],\mathscr{B}(C^0[0,1]))$ for $i = 1,2,3$, and for all $A\times B$ with $A\in\mathscr{F} = \mathscr{B}[0,1]^{\otimes N_n}$ and $B\in\mathscr{F}_i = \mathscr{B}(C^0[0,1])$, $\hat{\mathbb{P}}_{i,n,f}$ satisfies

$$\hat{\mathbb{P}}_{i,n,f}(A\times B) = \int_A P_{i,n-N_n,f,\hat{f}(y)}(B)\ (f\,d\mathsf{Leb})^{\otimes n}, \tag{5.9}$$

where the family of measures $\mathbb{P}_{i,n,f,f_0}$ from the local experiments is as defined from (4.8) to (4.12). These measures $\mathbb{P}_{i,n,f,\hat{f}}$ are now viewed as stochastic kernels. Note that for $i=0$, we have $\mathbb{P}_{0,n,f,\hat{f}} = (f\,d\mathsf{Leb})^{\otimes n}$, so $\hat{\mathbb{P}}_{0,n,f} = \mathbb{P}_{0,n,f}$ and $\mathsf{Expt}_{0,n}(\hat{f}) = \mathsf{Expt}_{0,n}$.

The idea of finding a local experiment defined on the neighborhood $\Sigma_n(\hat{f}_n)$ around Σ_n can be formalised below:

Proposition 5.1 Let $\mathsf{Expt} = (\mathscr{Y},\mathscr{F},(\mathbb{P}_\theta))_{\theta\in\Theta}$ be an experiment with sample space being Polish. Suppose we can formulate local experiments

$$\mathsf{Expt}_i(\phi) = \left(\Omega_i,\mathscr{A}_i,\{Q_{i,\theta,\phi}\}_{\theta\in\Theta(\phi)}\right), i = 1,2 \tag{5.10}$$

where $\Theta(\phi)$ is a collection of subsets of Θ with parameter $\phi\in\Theta$. Suppose further that there is a finite subset $\Theta_0\in\Theta$ and an estimator $\hat{\phi}:(\mathscr{Y},\mathscr{F})\to(\Theta_0,2^{\Theta_0})$, where 2^{Θ_0} is the power set of Θ_0. We formulate the following stochastic kernel from $(\mathscr{Y},\mathscr{F})$ to $(\Omega_i,\mathscr{A}_i)$:

$$Q_{i,\theta}(x,A') = Q_{i,\theta,\hat{\phi}(x)}(A'). \tag{5.11}$$

The measurability of $Q_{i,\theta}(.,A')$ is guaranteed by the fact that $\hat{\phi}$ takes only finitely many values. Formulate the compound experiments

$$\mathsf{Expt}_i(\hat{\phi}) = \left(\tilde{\mathscr{Y}}_i,\tilde{\mathscr{A}}_i,\left\{\tilde{\mathbb{P}}_{i,\theta}\right\}_{\theta\in\Theta}\right), \tag{5.12}$$

where $(\tilde{\mathscr{Y}}_i,\tilde{\mathscr{A}}_i) := (\mathscr{Y},\mathscr{F})\otimes(\Omega_i,\mathscr{A}_i)$ and $\mathbb{P}_{i,\theta}$ is the following measure on the product space $(\tilde{\mathscr{Y}}_i,\tilde{\mathscr{A}}_i)$:

$$\tilde{\mathbb{P}}_{i,\theta}(A\times B) = \int_A Q_{i,\theta}(y,B)\,\mathbb{P}_\theta(dy) \quad A\in\mathscr{F}, B\in\mathscr{A}_i \tag{5.13}$$

If we know that

$$\sup_{\phi\in\Theta}\Delta\left(\mathsf{Expt}_1(\phi),\mathsf{Expt}_2(\phi)\right)\le\epsilon \tag{5.14}$$

and there exists an estimator $\hat{\phi}$ such that

$$\inf_{\theta\in\Theta}\mathbb{P}_\theta\left(\theta\in\Theta(\hat{\phi})\right)\ge 1-\epsilon \tag{5.15}$$

then $\Delta(\mathsf{Expt}_1(\hat{\phi}),\mathsf{Expt}_2(\hat{\phi}))\le 4\epsilon$

Proof. First note that the set $V_\theta = \left\{y\,|\,\theta\in\Theta\left(\hat{\phi}(y)\right)\right\}$ is a measurable set in $(\mathscr{Y},\mathscr{F})$. This is because there are only finitely many possible value of $\hat{\phi}(y)$. Utilising proposition (3.15), for all $\phi\in\Theta$, there

is a stochastic kernel K_ϕ such that

$$\sup_{\theta\in\Theta(\phi)} L^1(K_\phi^\vee Q_{1,\theta,\phi}, Q_{2,\theta,\phi}) \leq 2\epsilon. \tag{5.16}$$

Now consider the stochastic kernel

$$M : \tilde{\mathscr{Y}_1} \times \tilde{A}_2 \to [0,1] \tag{5.17}$$

$$((y,\omega_1), A) \mapsto \int_{\Omega_2} \mathbb{I}_A(y,\omega_2)\, K_{\hat{\phi}(y)}(\omega_1, d\omega_2). \tag{5.18}$$

Note that this is a valid stochastic kernel, in the sense that:

- when (y,ω_1) is fixed, then M is a measure on $\tilde{A}_2$. This can be proved by noting that $K_{\hat{\phi}(y)}(y_0,.)$ is a measure when (y,ω_1) is fixed, and monotone convergence theorem applied on this measure.
- when A is fixed, then M is a measurable function on $\tilde{\mathscr{Y}_1}$. This comes from the fact that $\hat{\phi}$ is a measurable function on $(\mathscr{Y},\mathscr{F})$ which takes finitely many values when n is fixed. Therefore, fixing $\omega_1 \in \Omega_1$ and $B \in \mathscr{Y}_2$, the function $K_.(\omega_1, B)$ can be considered as a measurable function on $(\Theta_0, 2^{\Theta_0})$. Now consider $K_{\hat{\phi}(.)}(., B)$ as a composition of measurable functions to complete the proof.

Now we have, for $A \in \tilde{\mathscr{Y}_2} = \mathscr{Y}_2 \times \mathscr{A}_2$,

$$\begin{aligned}
M^\vee \tilde{\mathbb{P}}_{0,n,f}(A) &= \int_{\mathscr{Y}\times\Omega_1} M((y,\omega_1), A)\, Q_{1,\theta}(y, d\omega_1)\mathbb{P}_\theta(dy) \\
&= \int_{\mathscr{Y}} \int_{\Omega_1} \int_{\Omega_2} \mathbb{I}_A(y,\omega_2)\, K_{\hat{\phi}(y)}(\omega_1, d\omega_2) Q_{1,\theta}(y, d\omega_1)\mathbb{P}_\theta(dy) \\
&= \int_{\mathscr{Y}} \int_{\Omega_2} \mathbb{I}_A(y,\omega_2) \left(\int_{\Omega_1} K_{\hat{\phi}(y)}(\omega_1, d\omega_2) Q_{1,\theta}(y, d\omega_1) \right) \mathbb{P}_\theta(dy) \\
&= \int_{\mathscr{Y}} \int_{\Omega_2} \mathbb{I}_A(y,\omega_2)\, [K_{\hat{\phi}(y)}^\vee Q_{1,\theta}](y, d\omega_2)\mathbb{P}_\theta(dy).
\end{aligned}$$

Therefore we have, for all $A \in \tilde{\mathscr{Y}_2}$,

$$|\tilde{\mathbb{P}}_{2,\theta}(A) - M^\vee \tilde{\mathbb{P}}_{1,\theta}(A)| = \left| \left(\int_{V_\theta} + \int_{V_\theta^c} \right) \int_{\Omega_2} \mathbb{I}_A(y,\omega_2)\, [K_{\hat{\phi}(y)}^\vee Q_{1,\theta} - Q_{2,\theta}](y, d\omega_2)\mathbb{P}_\theta(dy) \right|.$$

We note that

$$\left| \int_{V_\theta} \int_{\Omega_2} \mathbb{I}_A(y,\omega_2)\, [K_{\hat{\phi}(y)}^\vee Q_{1,\theta} - Q_{2,\theta}](y, d\omega_2)\mathbb{P}_\theta(dy) \right| \leq 2\epsilon,$$

and that

$$\left| \int_{V_\theta^c} \int_{\Omega_2} \mathbb{I}_A(y,\omega_2)\, [K_{\hat{\phi}(y)}^\vee Q_{1,\theta} - Q_{2,\theta}](y, d\omega_2)\mathbb{P}_\theta(dy) \right| \leq 2\mathbb{P}_\theta(V_\theta^c) \leq 2\epsilon.$$

Taking supremum over all sets $A \in \mathscr{F} \times \mathscr{A}_2$ completes the proof. ■

Apply the above proposition with the settings as described at the beginning of the section, noting that we can select n such that

$$\sup_{f_0\in\Sigma} \Delta(\mathsf{Expt}_{0,n}(f_0), \mathsf{Expt}_{1,n}(f_0)) \leq \epsilon, \tag{5.19}$$

and the proposal estimator $\hat{\phi} := \hat{f}_n$ satisfying

$$\inf_{f \in \Sigma} \mathbb{P}_{0,n}(f \in \Sigma_n(\hat{f}_n)) \geq 1 - \epsilon. \tag{5.20}$$

Finally, note that $\mathsf{Expt}_{0,n}(\hat{f})$ coincides with Expt_0, we know that

$$\Delta(\mathsf{Expt}_{0,n}, \mathsf{Expt}_{3,n}(\hat{f}_n)) \overset{n\to\infty}{\to} 0. \tag{5.21}$$

Back to the heuristic. We have used the first N_n variables to construct a preliminary estimator $\hat{f}_n$ and utilise local equivalence. We can also use the first $n - N_n$ variables to construct another preliminary estimators $\hat{f}'_n$, then prove that

$$\Delta(\mathsf{Expt}_{0,n}, \mathsf{Expt}_{3,n}(\hat{f}'_n)) \overset{n\to\infty}{\to} 0. \tag{5.22}$$

Finally, we note without proof that

$$\Delta(\mathsf{Expt}_{3,n}(\hat{f}_n) \otimes \mathsf{Expt}_{3,n}(\hat{f}'_n), \mathsf{Expt}_{1,n}) = 0. \tag{5.23}$$

This comes from a sufficiency argument as suggested in [3], which we will omit the details. We therefore establish our desired global equivalence in theorem 4.1.

6 Discussion

6.1 Applicablity of stochastic kernels for constructing new decision rules

We have now completed the discussion of Nussbaum's proof of asymptotic equivalence in [3]. Combining with results in chapter 3, we have also established asymptotic equivalence between density estimation and non-parametric regression. Since we see that the Le Cam distance between these two statistical experiments are $o(1)$, we know from proposition 3.15 that we can extract stochastic kernels $T_{1,n}$ (from $\mathsf{Expt}_{0,n}$ to $\mathsf{Expt}_{1,n}$) and $T_{2,n}$ (from $\mathsf{Expt}_{1,n}$ to $\mathsf{Expt}_{0,n}$). These stochastic kernels serve as recipes to map a decision rule in one experiment to another decision rule in the other experiment so that the risk of original and new decision rules are comparable with respect to a bounded loss function. The kernels are hard to be constructed from Nussbaum's proof, given that such construction would involve manipulation over the abstract probability space $(\Omega^*, \mathscr{F}^*)$ in the KMT inequality.

There are numerous attempts to construct stochastic kernels between the two experiments, including [11, 10]. Indeed, we have covered half of this proof in the example given at the end of chapter 3, and the remaining steps involve techniques we have used in chapter 4. Another seminal paper on this topic is by Brown et al. [9], in which the author proved the asymptotic equivalence by introducing a Poisson process and constructed kernels based on dyadic approximations. Even so, using these maps is extremely hard as this would involve lots of computations when evaluating/approximating the integrals. Moreover, it is often the case that the bounds for Le Cam distance is often too loose (e.g. in Nussbaum's proof, we have a bound of $O((\ln n)^{-1/2})$, which is far worse than the non-parametric rate $o(n^{-1/4})$. Finally, the bounds are only applicable for risks with respect to a uniformly bounded function, which exclude common risks like the L^2 and L^∞ risks. As a result, the new estimators are often not optimal (and perform much worst than an appropriate kernel estimator).

6.2 Necessity of assumptions of asymptotic equivalence

So far, we have only worked with densities with finite supports, α-Hölder with $\alpha > 1/2$ and is bounded away from zero. We see in Nussbaum's proof that these assumptions have played a key role in ensuring that the log-likelihood λ_{f,f_0} is smooth enough for us to utilise the KMT inequality, so it would not be surprising to see that asymptotic equivalence may not hold if we consider a larger parameter space. An example has been constructed by Brown and Zhang in [16] for the case $\alpha = 1/2$, which is motivated by a compound hypothesis testing problem.

As for the condition of the densities being bounded away from zero, Ray and Schmidt-Hieber have proved in [17, 18] that if we allow the parameter space $\Sigma_n := \Sigma_{\alpha,M,\epsilon_n}$ to depend on n, such that $\epsilon_n \to 0$ in a sufficiently slow rate, then we will still have asymptotic equivalence. However, we will lose the asymptotic equivalence once we let $\epsilon_n \to 0$ too quickly, which includes the case of setting $\epsilon_n \equiv 0$. As a result, we cannot utilise any asymptotic results to problems of estimating densities not bounded away from zero, which include densities like Beta distribution that are commonly used in daily life.

Finally, one may wonder if one can drop the condition that the densities have finite supports. In between non-parametric regression and Gaussian white noise experiment, Brown and Low has provided an affirmative result in corollary 4.2 of [2], except we no longer evaluate the sample paths at points $t = 1/n$, but instead $\xi_l = H_n^{-1}(1/n)$ with H_n being a CDF of a well-behaved distribution supported by intervals $[\alpha_n, \beta_n] \overset{n\to\infty}{\uparrow} \mathbb{R}$. However, it is not clear under what condition on parameter space does one have asymptotic equivalence between density estimation and Gaussian white noise, and this is a potential research direction.

6.3 Conclusion and looking forward

To conclude, we have discussed the notion of Le Cam's notion of equivalence and established equivalence among density estimation, Gaussian white noise, and non-parametric regression using arguments from the seminal papers [3, 2]. In particular, we have filled in technical details of proofs, including the justification of the forms of the likelihood process and technicalities of using the divide-and-conquer principle.

We have also seen the role of stochastic kernels in mapping one decision rule in one experiment to another. For the reasons above, potentials to construct new estimators using these mappings is not prosperous. That said, the problem of determining if different statistical experiments are asymptotically equivalent is mathematically elegant by itself. The Le Cam distance also help establish upper bounds of rates of unknown statistical experiments.

7 Appendix

7.1 Conditional Distribution

To begin, we first remind ourselves with the notion of conditional expectation of random variables. For details see chapter 3 of [7].

Definition 7.1 — Conditional expectation. Suppose that $\xi : (\Omega, \mathscr{F}, \mathbb{P}) \to (E, \mathscr{E})$ is an integrable random variable, and that $\mathscr{G}$ is a sub-$\sigma-$algebra of $\mathscr{F}$. A random variable $Y := \mathbb{E}[\xi|\mathscr{G}]$ is a *conditional expectation of ξ given $\mathscr{G}$* if it has these two properties:

- $\mathbb{E}[\xi|\mathscr{G}]$ is $\mathscr{G}$-measurable and integrable.
- For every $G \in \mathscr{G}$,

$$\mathbb{E}[\chi_G \mathbb{E}[\xi|\mathscr{G}]] = \int_G \mathbb{E}[\xi|\mathscr{G}]\mathrm{d}\mathbb{P} = \int_G \xi \mathrm{d}\mathbb{P} = \mathbb{E}[\chi_G \xi] \tag{7.1}$$

Let η be another random variable on Ω, taking values on a potentially different measure space $(E', \mathscr{E}')$. Let $\sigma(\eta)$ be the σ-algebra generated by η, then any conditional expectations of ξ given $\sigma(\eta)$ are also *conditional expectations of ξ given η*, denoted as $\mathbb{E}[\xi|\eta]$.

We note that whenever X is an integrable random variable on $(\Omega, \mathscr{F}, \mathbb{P})$, and that $\mathscr{G}$ is a sub-σ-algebra of $\mathscr{F}$, then such a random variable exists, and is unique almost everywhere. We can then define conditional probability as followed

Definition 7.2 — Conditional probability. Under the same settings as definitions 7.1, the conditional probability of any sets $A \in \mathscr{F}$ given $\mathscr{G}$ is defined as $\mathbb{P}(A \,|\, \mathscr{G}) = \mathbb{E}[\mathbb{I}_A|\mathscr{G}]$. Similarly we have $\mathbb{P}(A \,|\, \eta) = \mathbb{E}[\mathbb{I}_A|\eta]$.

We finally define the notion of (regular) conditional distribution

Definition 7.3 — (Regular) conditional distribution. Under the same setting as in definition 7.1, a (regular) conditional distribution of ξ **given the σ-algebra** $\mathscr{G}$ is a stochastic kernel $Q : \Omega \times \mathscr{E} \to [0, \infty]$ such that for almost all $\omega \in \Omega$, we have

$$\forall B \in \mathscr{E}, \quad Q(\omega, B) = [\mathbb{P}(\xi \in B \,|\, \mathscr{G})](\omega) \tag{7.2}$$

The definition of (regular) conditional distribution of ξ given η is the same as above with $\mathscr{G}$ being equal to $\sigma(\eta)$.

Regular condition exists whenever $(E, \mathscr{E})$ is a Polish space.

Remark 7.4 Let Q be a regular conditional distribution of ξ given η, then for almost all $\omega \in \Omega$, and for all $B \in \mathscr{E}$, we have

$$Q(\omega, B) = [\mathbb{P}(\xi \in B \,|\, \eta)](\omega) \tag{7.3}$$

When Q is restricted to the set of ω such that the above equality is satisfied, then $Q(., B)$ is η measurable, and hence $Q(\omega, B) = \tilde{Q}(\eta(\omega), B)$ for some function $\tilde{Q} : (E', \mathscr{E}') \times \mathscr{E} \to [0, \infty]$, such that $\tilde{Q}(., B)$ is $\mathscr{E}$ measurable for all $B \in \mathscr{E}$. We will also refer to this function $\tilde{Q}$ as a regular conditional distribution of ξ on η.

7.2 Distance between measures

Consider a measure space $(\Omega, \mathscr{F})$, and let $\mathbb{P}, \mathbb{Q}$ are two measures on $(\Omega, \mathscr{F})$ which is absolutely continuous with respect to another measure ν with density p and q respectively. We consider the following notions of distance between these two measures

- The total variation distance, or the L^1 distance:

$$L^1(\mathbb{P},\mathbb{Q}) = \sup_{A\in\mathscr{F}} |\mathbb{P}(A)-\mathbb{Q}(A)| = \mathbb{E}_\nu|p-q| = \int |p(x)-q(x)|\nu(dx) \tag{7.4}$$

- The Hellinger distance:

$$(H(\mathbb{P},\mathbb{Q}))^2 = \mathbb{E}_\nu\left|\sqrt{p}-\sqrt{q}\right|^2 = 2\left(1-\mathbb{E}_\nu\left(\sqrt{pq}\right)\right) \tag{7.5}$$

We then have the following basic inequality

Lemma 7.5 — Bound of L^1 distance by the Hellinger distance.

$$L^1(\mathbb{P},\mathbb{Q}) \le 2H(\mathbb{P},\mathbb{Q}) \tag{7.6}$$

Proof.

$$\begin{aligned} L^1(\mathbb{P},\mathbb{Q}) &= \int \left(\sqrt{p}-\sqrt{q}\right)\left(\sqrt{p}+\sqrt{q}\right)\,d\nu \\ &\overset{\text{(Hölder)}}{\le} H(\mathbb{P},\mathbb{Q})\sqrt{\int \left(\sqrt{p}+\sqrt{q}\right)^2\,d\nu} \\ &\overset{\text{(Young)}}{\le} H(\mathbb{P},\mathbb{Q})\sqrt{2\int (p+q)\,d\nu} = 2H(\mathbb{P},\mathbb{Q}) \\ &= H(\mathbb{P},\mathbb{Q})\sqrt{2(1+1)} = 2H(\mathbb{P},\mathbb{Q}) \end{aligned}$$

notice the Young inequality $2ab \le a^2+b^2$ and that the Radon-Nikodym derivatives all integrate to 1. ■

We also make note of the bounds of Hellinger distance between product measures:

Lemma 7.6 — Hellinger distance between product measures (see (19)). Suppose that $\mathbb{P}_i$ and $\mathbb{Q}_i$ are probability measures on a measurable space $(\Omega_i,\mathscr{F}_i)$ for $i=1,...,k$ dominated by ν_i. Then

$$\left(H\left(\bigotimes_{i=1}^k \mathbb{P}_i, \bigotimes_{i=1}^k \mathbb{Q}_i\right)\right)^2 \le \sum_{i=1}^k (H(\mathbb{P}_i,\mathbb{Q}_i))^2 \tag{7.7}$$

Proof. We only prove that $(H(\mathbb{P}_1\otimes\mathbb{P}_2,\mathbb{Q}_1\otimes\mathbb{Q}_2))^2 \le 2\left((H(\mathbb{P}_1,\mathbb{Q}_1))^2+(H(\mathbb{P}_2,\mathbb{Q}_2))^2\right)$, the more general case can be extended by induction. Notice that

$$\begin{aligned} 1-\frac{1}{2}(H(\mathbb{P}_1\otimes\mathbb{P}_2,\mathbb{Q}_1\otimes\mathbb{Q}_2))^2 &= \mathbb{E}_\nu\left(\sqrt{\frac{d(\mathbb{P}_1\otimes\mathbb{P}_2)}{d(\nu_1\otimes\nu_2)}\frac{d(\mathbb{Q}_1\otimes\mathbb{Q}_2)}{d(\nu_1\otimes\nu_2)}}\right) \\ &= \int\int \sqrt{\frac{d\mathbb{P}_1}{d\nu_1}\frac{d\mathbb{P}_2}{d\nu_2}\frac{d\mathbb{Q}_1}{d\nu_1}\frac{d\mathbb{Q}_2}{d\nu_2}}\,d\nu_1 d\nu_2 \\ &= \left(\int \sqrt{p_1q_1}\,d\nu_1\right)\left(\int \sqrt{p_2q_2}\,d\nu_2\right) \\ &= \left(1-\frac{1}{2}(H(\mathbb{P}_1,\mathbb{Q}_1))^2\right)\left(1-\frac{1}{2}(H(\mathbb{P}_2,\mathbb{Q}_2))^2\right) \\ &= 1-\frac{1}{2}\left[(H(\mathbb{P}_1,\mathbb{Q}_1))^2+(H(\mathbb{P}_2,\mathbb{Q}_2))^2\right]+\frac{1}{4}(H(\mathbb{P}_1,\mathbb{Q}_1))^2(H(\mathbb{P}_2,\mathbb{Q}_2))^2 \end{aligned}$$

which completes the proof. ■

We consider this instructive example.

Example 7.7 — L^1 and Hellinger distance between two Gaussian sequences with same covariances, see Chapter 3 of (5). Consider the measures on $\mathbb{R}^n$ (where n can be both finite or infinite):

$$\mathbb{P}_\theta(A) = \int_A \left(\prod_{i=1}^n \frac{1}{\epsilon\rho_i\sqrt{2\pi}} \exp\left(-\frac{(y_i-\theta_i)^2}{2\epsilon^2\rho_i^2} \right) dy_i \right), \quad \theta = (\theta_1, \theta_2, ...) \tag{7.8}$$

which is the joint distribution of the following Gaussian sequence

$$y_i = \theta_i + \epsilon\rho_i z_i, \quad z_i \overset{\text{iid}}{\sim} \mathsf{N}(0,1) \tag{7.9}$$

Computing the L^1 distance between $\mathbb{P}_\theta$ and $\mathbb{P}_0$. Firstly, note that the measures are mutually absolutely continuous with

$$\frac{d\mathbb{P}_\theta}{d\mathbb{P}_0} = \exp\left(\zeta(y) - \frac{|h|^2}{2} \right) \tag{7.10}$$

where h is the vector $(h_1/(\epsilon\rho_1), h_2/(\epsilon\rho_2), ...)$, and ζ is the random variable $\zeta(y) = \sum_{i=1}^n (y_i\theta_i)/(\epsilon^2\rho_i^2)$. Note that under measure $\mathbb{P}_0$, the distribution of the random variable $\zeta^*\mathbb{P}_0$ follows a single-variable Gaussian distribution with zero mean and variance

$$\sum_{i=1}^n \frac{\theta_i^2}{\epsilon^4\rho_i^4}\epsilon^2\rho_i^2 = \sum_{i=1}^n \frac{\theta_i^2}{\epsilon^2\rho_i^2} = |h|^2$$

Therefore, by change of variable formula, we know that

$$\begin{aligned}
L^1(\mathbb{P}_\theta, \mathbb{P}_0) &= \int_{\mathbb{R}^n} \left| \frac{d\mathbb{P}_\theta}{d\mathbb{P}_0}(y) - 1 \right| \mathbb{P}_0(dy) \\
&= \int_{\mathbb{R}^n} \left| \exp\left(\zeta(y) - \frac{|h|^2}{2} \right) - 1 \right| \mathbb{P}_0(dy) \\
&= \int_{\mathbb{R}} \left| \exp\left(\zeta - \frac{|h|^2}{2} \right) - 1 \right| \zeta^*\mathbb{P}_0(d\zeta) \\
&= \int_{\mathbb{R}} \left| \exp\left(\zeta - \frac{|h|^2}{2} \right) - 1 \right| \frac{1}{|h|\sqrt{2\pi}} \exp\left(-\frac{\zeta^2}{2|h|^2} \right) d\zeta \\
&= 2\int_{\{\zeta \geq |h|^2/2\}} \frac{1}{|h|\sqrt{2\pi}} \left(\exp\left(-\frac{(\zeta - |h|^2)^2}{2|h|^2} \right) - \exp\left(-\frac{\zeta^2}{2|h|^2} \right) \right) d\zeta \\
&\overset{z=\zeta/|h|}{=} 2\left(\int_{\{z \geq |h|/2\}} \frac{1}{\sqrt{2\pi}} \left(\exp\left(-\frac{(z-|h|)^2}{2} \right) - \exp\left(-\frac{z^2}{2} \right) \right) dz \right) \\
&= 2\left(\int_{\{z \geq -|h|/2\}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2} \right) dz - \int_{\{z \geq |h|/2\}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2} \right) dz \right)
\end{aligned}$$

Recall the notation $\tilde{\Phi}(z) = \int_z^\infty \exp(-t^2/2)/\sqrt{2\pi}\, dt$, we have

$$L^1(\mathbb{P}_\theta, \mathbb{P}_0) = 2(1 - 2\tilde{\Phi}(|h|/2)) \tag{7.11}$$

where $|h|$ is as specified above. We also note that, in general,

$$L^1(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) = 2\left(1 - 2\tilde{\Phi}\left(\frac{1}{2}\sqrt{\sum_{i=1}^n \frac{(\theta_i - \theta_i')^2}{\epsilon^2\rho_i^2}} \right) \right) \tag{7.12}$$

Computing the Hellinger distance between $\mathbb{P}_\theta$ and $\mathbb{P}_0$. The computation is much simpler, and we have

$$\begin{aligned}
H^2(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) &= 1 - \int \prod_{i=1}^n \sqrt{\left(\frac{1}{\sqrt{2\pi}\epsilon\rho_i} \exp\left(-\frac{(y_i-\theta_i)^2}{2\epsilon^2\rho_i^2}\right)\right)\left(\frac{1}{\sqrt{2\pi}\epsilon\rho_i} \exp\left(-\frac{(y_i-\theta_i')^2}{2\epsilon^2\rho_i^2}\right)\right)} \\
&= 1 - \int \prod_{i=1}^n \left(\frac{1}{\sqrt{2\pi}\epsilon\rho_i} \exp\left(-\frac{2y_i^2 - 2(\theta_i+\theta_i')y_i + \theta_i^2 + \theta_i'^2}{4\epsilon^2\rho_i^2}\right)\right) \\
&= 1 - \int \prod_{i=1}^n \left(\frac{1}{\sqrt{2\pi}\epsilon\rho_i} \exp\left(-\frac{(y_i-(\theta_i+\theta_i')/2)^2 + \theta_i^2/2 + \theta_i'^2/2 - (\theta_i+\theta_i')^2/4}{2\epsilon^2\rho_i^2}\right)\right) \\
&= 1 - \exp\left(-\frac{1}{8}\sum_{i=1}^n \frac{(\theta_i-\theta_i')^2}{\epsilon^2\rho_i^2}\right)
\end{aligned}$$

■

We can use similar procedure to compute the L^1 distance and the Hellinger distance between two Gaussian white noise. The main obstacle is to look at the likelihood ratio, which would be discussed in the next section.

7.3 Girsanov Theorem

Let $\mathbb{P}$ be a probability measures on $\mathscr{B}(C^0[0,1])$, the Borel σ-algebra of $C^0[0,1]$. Consider Brownian motion X_t as a random variable on $\mathscr{B}(C^0[0,1])$, and let $f(t) \in L^2[0,1]$. Define measure $\mathbb{Q}$ on $\mathscr{B}(C^0[0,1])$ such that it is absolutely continuous with $\mathbb{P}$, and that

$$\frac{d\mathbb{Q}}{d\mathbb{P}} = \exp\left(\int_0^1 f(s)\, dW_s - \frac{1}{2}\int_0^1 (f(s))^2\, ds\right) \tag{7.13}$$

Then by Girsanov Theorem ([15], theorem 5.22-23), X_t has the same distribution as the stochastic integral under $\mathbb{Q}$

$$X_t \overset{d}{=} W_t + \int_0^t f(s)\, ds \tag{7.14}$$

where W_t is any Brownian motion under $\mathbb{Q}$. In other words, X_t is a (weak) solution (under $\mathbb{Q}$) to the Gaussian white noise

$$dX_t = f(t)\, dt + dW_t \tag{7.15}$$

The exponential martingale in (7.17) hence gives the likelihood ratio of the Gaussian white noise (7.15) with respect to Brownian motion. For convenience, we note that by scaling:

Theorem 7.8 — Girsanov Theorem for scaled Gaussian white noise. Let $\mathbb{P}_0$ be the distribution of $X_t = \sigma n^{-1/2} W_t$ and $\mathbb{P}_f$ be the distribution of the Gaussian white process

$$dX_t = f(t)\, dt + \sigma n^{-1/2} dW_t \tag{7.16}$$

Then we have

$$\frac{d\mathbb{P}_f}{d\mathbb{P}_0} = \exp\left(\frac{\sqrt{n}}{\sigma}\int_0^1 f(s)\, dW_s - \frac{n}{2\sigma^2}\int_0^1 (f(s))^2\, ds\right) \tag{7.17}$$

From this, we can calculate the total variation distance between $\mathbb{P}_f$ and $\mathbb{P}_0$:

Example 7.9 — L^1 and Hellinger distance between two Gaussian white noise ((20), Theorem 1;

(5), equation 3.61). We have

$$\begin{aligned} L^1(\mathbb{P}_f, \mathbb{P}_0) &= \int_{C^0[0,1]} \left| \frac{d\mathbb{P}_\theta}{d\mathbb{P}_0}(\mathsf{w}) - 1 \right| \mathbb{P}_0(d\mathsf{w}) \\ &= \int_{C^0[0,1]} \left| \exp\left(\frac{\sqrt{n}}{\sigma} \int_0^1 f(s)\, d\mathsf{w}(s) - \frac{n}{2\sigma^2} \int_0^1 (f(s))^2\, ds \right) - 1 \right| \mathbb{P}_0(d\mathsf{w}) \end{aligned}$$

Similar as before, we define $D^2 = \int_0^1 \frac{n}{\sigma^2}(f(s))^2\, ds$ and consider the random variable

$$\xi(\mathsf{w}) = \int_0^1 \frac{\sqrt{n}}{\sigma} f(t)\, d\mathsf{w}(t) \tag{7.18}$$

Under $\mathbb{P}_0$, this random variable follows a Gaussian distribution with zero mean. By Ito's isometry lemma, we have

$$\mathbb{E}_{\mathbb{P}_0}[\xi(\mathsf{w})^2] = \int_0^1 \frac{n}{\sigma^2}(f(t))^2\, dt = D^2 \tag{7.19}$$

We therefore have

$$\begin{aligned} L^1(\mathbb{P}_f, \mathbb{P}_0) &= \int_{\mathbb{R}} \left| \exp\left(\xi - \frac{D^2}{2} \right) - 1 \right| \xi^* \mathbb{P}_0(d\xi) \\ &= \int_{\mathbb{R}} \left| \exp\left(\xi - \frac{D^2}{2} \right) - 1 \right| \frac{1}{D\sqrt{2\pi}} \exp\left(-\frac{\zeta^2}{2D^2} \right) d\xi \\ &= 2(1 - 2\tilde{\Phi}(D/2)) \end{aligned}$$

More generally, we have

$$L^1(\mathbb{P}_f, \mathbb{P}_g) = 2\left(1 - 2\tilde{\Phi}\left(\frac{\sqrt{n}}{2\sigma} \sqrt{\int_0^1 (f(t) - g(t))^2\, dt} \right) \right) \tag{7.20}$$

As for Hellinger distance, we have

$$\begin{aligned} H^2(\mathbb{P}_f, \mathbb{P}_0) &= 1 - \int_{C^0[0,1]} \sqrt{\exp\left(\frac{\sqrt{n}}{\sigma} \int_0^1 f(s)\, d\mathsf{w}(s) - \frac{n}{2\sigma^2} \int_0^1 (f(s))^2\, ds \right)} \mathbb{P}_0(d\mathsf{w}) \\ &= 1 - \int_{C^0[0,1]} \exp\left(\frac{\xi(\mathsf{w})}{2} - \frac{D^2}{4} \right) \mathbb{P}_0(d\mathsf{w}) \\ &= 1 - \int_{\mathbb{R}} \frac{1}{D\sqrt{2\pi}} \exp\left(\frac{\xi}{2} - \frac{D^2}{4} - \frac{\xi^2}{2D^2} \right) d\xi \\ &= 1 - \exp\left(-\frac{D^2}{8} \right) \end{aligned}$$

And in general one have

$$H^2(\mathbb{P}_f, \mathbb{P}_g) = 1 - \exp\left(-\frac{n}{8\sigma^2} \int_0^1 (f(s) - g(s))^2\, ds \right) \tag{7.21}$$

■

References

[1] Cam LL. Sufficiency and Approximate Sufficiency. The Annals of mathematical statistics. 1964;35(4):1419-55.

[2] Brown LD, Low MG. Asymptotic equivalence of nonparametric regression and white noise. The Annals of Statistics. 1996;24(6):2384 2398. Available from: https://doi.org/10.1214/aos/1032181159.

[3] Nussbaum M. Asymptotic Equivalence of Density Estimation and Gaussian White Noise. The Annals of statistics. 1996;24(6):2399-430.

[4] Tsybakov AB. Introduction to Nonparametric Estimation. 1st ed. Springer Series in Statistics. New York, NY: Springer New York; 2009.

[5] Johnstone I. Gaussian Estimation: Sequence and Wavelet Models (Draft); 2019.

[6] Ibragimov IA, Khas'minskii RZ. Estimation of distribution density. Journal of Soviet Mathematics. 1983;21(1):40-57.

[7] Kallenberg O. Foundations of Modern Probability. 3rd ed. Probability Theory and Stochastic Modelling, 99. Cham: Springer International Publishing; 2021.

[8] Durrett R. Probability : theory and examples. Fifth edition. ed. Cambridge series in statistical and probabilistic mathematics ; 49. Cambridge: Cambridge University Press; 2019.

[9] Brown LD, Carter AV, Low MG, Zhang CH. Equivalence theory for density estimation, Poisson processes and Gaussian white noise with drift. The Annals of Statistics. 2004;32(5):2074 2097. Available from: https://doi.org/10.1214/009053604000000012.

[10] Mariucci E. Asymptotic equivalence for density estimation and Guassian white noise: an extension. Annales de l'ISUP. 2016;60(1-2):23-34. 11 pages. Available from: https://hal.archives-ouvertes.fr/hal-01132442.

[11] Carter AV. Deficiency Distance between Multinomial and Multivariate Normal Experiments. The Annals of statistics. 2002;30(3):708-30.

[12] Keener RW. Theoretical Statistics Topics for a Core Course. 1st ed. Springer Texts in Statistics. New York, NY: Springer New York; 2010.

[13] Koltchinskii VI. Komlos-Major-Tusnady approximation for the general empirical process and Haar expansions of classes of functions. Journal of theoretical probability. 1994;7(1):73-118.

[14] Falk M, Reiss RD. Poisson approximation of empirical processes. Statistics & probability letters. 1992;14(1):39-48.

[15] Le Gall JF. Brownian Motion, Martingales, and Stochastic Calculus. 1st ed. Graduate Texts in Mathematics, 274. Cham: Springer International Publishing; 2016.

[16] Brown LD, Zhang CH. Asymptotic Nonequivalence of Nonparametric Experiments When the Smoothness Index is 1/2. The Annals of Statistics. 1998;26(1):279-87. Available from: http://www.jstor.org/stable/119987.

[17] Ray K, Schmidt-Hieber J. The Le Cam distance between density estimation, Poisson processes and Gaussian white noise. Mathematical Statistics and Learning. 2018-Sep;1.

[18] Ray K, Schmidt-Hieber J. Asymptotic nonequivalence of density estimation and Gaussian white noise for small densities. Annales de l'Institut Henri Poincaré, Probabilités et Statistiques. 2019;55(4):2195 2208. Available from: https://doi.org/10.1214/18-AIHP946.

[19] Zolotarev VM. Probability Metrics. Theory of probability and its applications. 1984;28(2):278-302.

[20] Barsov SS, Ulyanov VV. Estimates of the proximity of Gaussian measures. Soviet Math Dokl. 1987;34(3):462-6.