Finding a Needle in a Haystack: Conditions for Reliable Detection in the Presence of Clutter*

Bruno Jedynak‡and Damianos Karakos§

October 23, 2006

Abstract

We study conditions for the detection of an N-length iid sequence with unknown pmf p_1, among M N-length iid sequences with unknown pmf p_0. We show how the quantity $M2^{-N\,D(p_1||p_0)}$ determines the asymptotic probability of error.

Keywords: reliable detection, probability of error, Sanov's theorem, Kulback-Leibler distance, phase transition.

1 Introduction

Our motivation for this paper has its origins in Geman et. al. (1996), where an algorithm for tracking roads in satellite images was experimentally studied. Below a certain clutter level, the algorithm could track a road acurately, and suddenly, with increased clutter level, tracking would become impossible. This phenomenon was studied theoretically in Yuille et. al.(2000 and 2001) . Using a simplified statistical model, the authors show that, in an

*This research was partially supported by ARO DAAD19/-02-1-0337 and general funds from the Center for Imaging Science at The Johns Hopkins University.

†USTL and Department of Applied Mathematics, Johns Hopkins University

‡Mail should be adressed to: Bruno Jedynak, Clark 302b, Johns Hopkins University, Baltimore, MD, 21286-2686

§Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, 21286-2686

apropriate asymptotic setting, the number of false detections is subject to a phase transition. Our objective in this paper is to generalize these results. First, we demonstrate, in the same setting, that the phase transition phenomenon occurs for the error rate of the maximum likelihood estimator. Second, we consider the situation where the underlying statistical model is unknown; i.e., there is a special object among many others but it is not known *how* it is special (it is an outlier, in some sense). We show that the same phase transition phenomenon occurs in this case as well. Moreover, we propose a target detector that has the same asymptotic performance as the maximum likelihood estmator, had the model been known. Simulations illustrate these results.

Let

$$X = \begin{pmatrix} X_1^1 & X_1^2 & \dots & X_1^N \\ X_2^1 & X_2^2 & \dots & X_2^N \\ \vdots & \vdots & & \vdots \\ X_{M+1}^1 & X_{M+1}^2 & \dots & X_{M+1}^N \end{pmatrix} \tag{1}$$

be a $(M+1) \times N$ matrix made of independent random variables (rvs) taking values in a finite set. We denote by $X_m = (X_m^1, \dots, X_m^N) \in \mathcal{X}^N$ the rvs in line m and by $X_{(m)}$ the ones that are *not* in line m. There is a special line, the *target*, with index t. All the other lines will be called *distractors*. The rvs X_t are identically distributed with point mass function (pmf) p_1. The other ones, $X_{(t)}$, are identically distributed with pmf $p_0 \neq p_1$. The goal is to estimate t, the target, from a single realization of X. If p_0 and p_1 are "close", the target does not differ much from the distractors, a situation akin to "finding a needle in a haystack".

2 Known distributions

Let x be a realization of X. Then, the log-likelihood[1] of x is

$$l(x) = \sum_{n=1}^{N} \log p_1(x_t^n) + \sum_{m=1, m\neq t}^{M+1} \sum_{n=1}^{N} \log p_0(x_m^n) \quad (2)$$

$$= \sum_{n=1}^{N} \log \frac{p_1(x_t^n)}{p_0(x_t^n)} + \sum_{m=1}^{M+1} \sum_{n=1}^{N} \log p_0(x_m^n) \quad (3)$$

The maximum likelihood estimator (mle) for t is then

$$\hat{t}(x) = \arg \max_{1\leq m\leq M+1} \sum_{n=1}^{N} \log \frac{p_1(x_m^n)}{p_0(x_m^n)} \quad (4)$$

We call the *reward* of line m the quantity

$$\frac{1}{N} \sum_{n=1}^{N} \log \frac{p_1(x_m^n)}{p_0(x_m^n)} \quad (5)$$

The mle entails choosing the line with the largest reward. The quantity of interest is the probability that the mle differs from the target:

$$e(M, N) = I\!P(\hat{t}(X) \neq t) \quad (6)$$

which is the probability that a distractor gets a reward which is greater than the reward of the target. If M is fixed , letting $N \to \infty$, and using the law of large numbers, we obtain

$$\frac{1}{N} \sum_{n=1}^{N} \log \frac{p_1(x_t^n)}{p_0(x_t^n)} \to D(p_1, p_0) \text{ and} \quad (7)$$

$$\frac{1}{N} \sum_{n=1}^{N} \log \frac{p_1(x_m^n)}{p_0(x_m^n)} \to -D(p_0, p_1) \text{ for every } m \neq t \quad (8)$$

[1] Logarithms are base 2 throughout the paper.

almost surely, where

$$D(p,q) = \sum_x p(x) \log \frac{p(x)}{q(x)} \tag{9}$$

is the Kulback-Leibler distance between p and q. Hence, as long $p \neq q$, $D(p,q) > 0$, and the reward of the target converges to a positive value while the reward of each distractor converges to a negative value which allows us to show that the error of the mle goes to zero. One can even bound $e(M,N)$ from above for any fixed M and N as follows

Theorem 1

$$e(M,N) \leq M \left(\sum_x \sqrt{p_0(x)p_1(x)} \right)^{2N}. \tag{10}$$

Note that

$$0 \leq \sum_x \sqrt{p_0(x)p_1(x)} = 1 - Hellinger(p_0,p_1) \leq 1, \tag{11}$$

where $Hellinger(p_0,p_1)$ is the Hellinger distance between p_0 and p_1. The proof, using classical large deviations techniques, is in Section 6. Note that if the right-hand side of (10) goes to 0 as $M \to \infty$ and $N \to \infty$, the probability that the mle differs from the target goes to 0. This condition, however, is not necessary. As we show below, there is a maximum rate at which M can go to infinity in order for the probability of error to go to zero (if M increases faster, then the probability of error goes to one). A similar result, i.e., that the number of distractors for which the reward is larger than the reward of the target follows a phase transition, was also shown by Yuille et. al.(2000). We present below the same analysis for the convergence of the mle. The phase transition, or in other words, the dependence of the probability of error on the rate at which M goes to infinity, is expressed in the following theorem:

Theorem 2

$$\textit{If } \exists \varepsilon > 0, \textit{ such that } \lim_{M,N\to\infty} M 2^{-N(D(p_1,p_0)-\varepsilon)} = 0 \quad \textit{then } \lim_{M,N\to\infty} e(M,N) = 0, \tag{12}$$

and

$$\textit{If } \exists \varepsilon > 0, \textit{ such that } \lim_{M,N\to\infty} M2^{-N(D(p_1,p_0)+\varepsilon)} = +\infty \quad \textit{then } \lim_{M,N\to\infty} e(M,N) = 1. \quad (13)$$

The intuition is as follows. First, as N goes to infinity, if M remains fixed, the probability of error goes to zero (exponentially fast, following a large deviation phenomenon) since the reward of the target line converges to a positive value, while the reward of the distractors converges to a negative value (as was mentioned earlier). On the other hand, as the number M of distractors increases, when N remains fixed, the probability that there exists a distractor with a reward larger than the reward of the target increases as well. These are two competing phenomena, whose interaction gives rise to the "critical rate" $D(p_1, p_0)$. The detailed proof appears in Section 6.

Note: In order for the limits of functions of M, N to be well-defined as $M, N \to \infty$, we assume that M is, in general, a function of N. Hence, all limits $\lim_{M,N\to\infty}$ should be interpreted as $\lim_{N\to\infty}$, with the proviso that M is increasing according to some function of N. We kept the notation $\lim_{M,N\to\infty}$ for simplicity.

3 Unknown Distributions

We now look at the case where p_0 and p_1 are unknown. It is clear that the error rate of any estimator in this context cannot be lower than the error rate of the mle (with known p_0 and p_1). Hence, (13) holds even when $e(M, N)$ is the error rate of any estimator. Can one build an estimator of the target for which the error rate will satisfy (12) ? The answer is yes as we shall see now.

A simple way of building an estimator of the target when p_0 and p_1 are unknown is to plug-in estimators of p_0 and p_1 in the previous (mle) estimator (4). Hence, let us define

$$\tilde{t}(x) = \arg \max_{1 \le m \le M+1} \sum_{n=1}^{N} \log \frac{\hat{p}_m(x_m^n)}{\hat{p}_{(m)}(x_m^n)} \tag{14}$$

where $\hat{p}_m$ and $\hat{p}_{(m)}$ are the empirical distributions of the rvs in line m and in all the other lines, respectively. I.e.,

$$\hat{p}_m(x) = \frac{1}{N} \sum_{n=1}^{N} \mathbf{1}\{X_m^n = x\}, \tag{15}$$

and

$$\hat{p}_{(m)}(x) = \frac{1}{MN} \sum_{n=1}^{N} \sum_{j=1, j \neq m}^{M+1} \mathbf{1}\{X_j^n = x\}. \tag{16}$$

Note that

$$\tilde{t}(x) = \arg \max_{1 \le m \le M+1} D(\hat{p}_m, \hat{p}_{(m)}). \tag{17}$$

Hence, $\tilde{t}$ is the line that differs the most (in the Kulback-Leibler sense) from the average distribution of the *other* lines. (The reader may be more familiar with the variant

$$\dot{t}(x) = \arg \max_{1 \le m \le M+1} D(\hat{p}_m, \hat{p}), \tag{18}$$

where $\hat{p}$ is the empirical distribution over all rvs, including line m; both $\dot{t}$ and $\tilde{t}$ are similar in the sense that they pick the sequence which differs the most from the rest.)

It turns out that the error rate of $\tilde{t}$, that is

$$\tilde{e}(M, N) = I\!P(\tilde{t}(X) \neq t), \tag{19}$$

where, as before, t denotes the target, has the same asymptotic behavior as the mle (4) in the case of known distributions.

Theorem 3

$$\textit{If } \exists \varepsilon > 0, \textit{ such that } \lim_{M,N \to \infty} M 2^{-N(D(p_1, p_0) - \varepsilon)} = 0 \quad \textit{then } \lim_{M,N \to \infty} \tilde{e}(M, N) = 0 \tag{20}$$

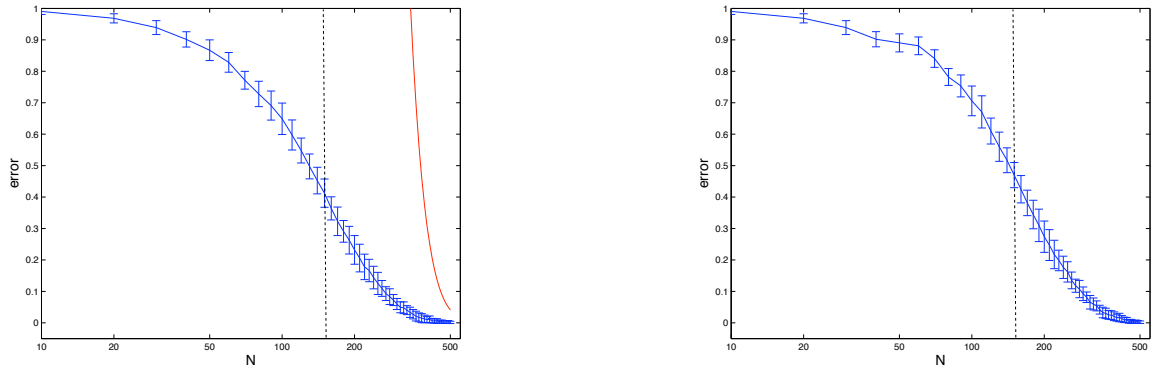

Figure 1: Estimates of the probability of error for various N, for the case $p_0 = (0.9, 0.1)$, $p_1 = (0.8, 0.2)$, $M = 1000$. The two plots correspond to the cases of known and unknown distributions, respectively. The red line represents the upper bound as established by Theorem 1.

and

$$\textit{If } \exists \varepsilon > 0, \textit{ such that } \lim_{M,N\to\infty} M2^{-N(D(p_1,p_0)+\varepsilon)} = +\infty \quad \textit{then } \lim_{M,N\to\infty} \tilde{e}(M,N) = 1. \quad (21)$$

The proof uses the same large deviations techniques as the proof of Theorem 2 but is slightly more complex due to the fact that the rewards are not independent anymore. The proof appears in Section 6.

4 Simulations

We now provide simulations that show Theorems 1, 2 and 3 in action.

We generated $M = 1000$ binary sequences with probabilities $p_0 = (0.9, 0.1)$ and $p_1 = (0.8, 0.2)$ for the background and the target, respectively. We varied the number N from 10 to 500, and we observed the probability of error decreasing to zero. We performed the random experiment 100 times for each value of N. The procedure was replicated 20 times in

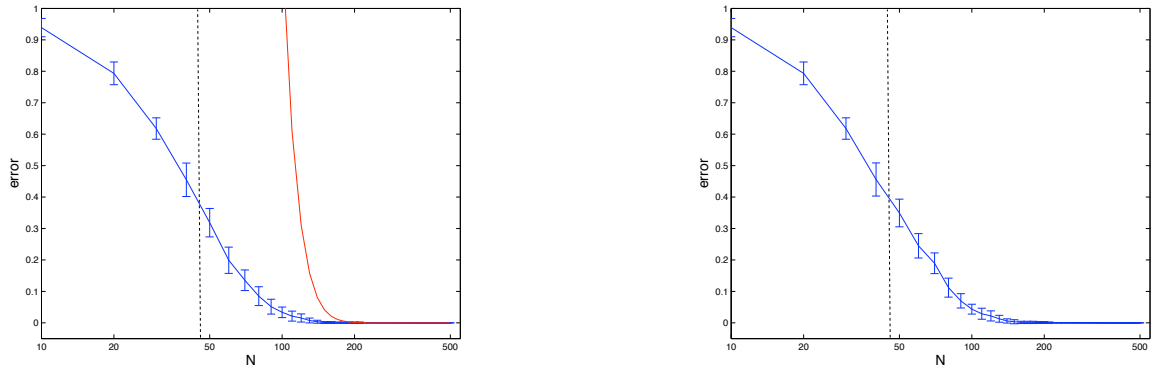

Figure 2: Estimates of the probability of error for various N, for the case $p_0 = (0.9, 0.1)$, $p_1 = (0.7, 0.3)$, $M = 1000$. The two plots correspond to the cases of known and unknown distributions, respectively. The red line represents the upper bound as established by Theorem 1.

order to compute error bars. The plots in Figure 1 show the (estimated) probability of error versus N, for the two maximum likelihood detectors (known and unknown distributions, respectively), along with 1 standard deviation error bars. As expected, the error for the case of unknown distributions is somewhat higher, as there is an additional error due to the inaccuracy in estimating the two distributions. The KL divergence is $D(p_1, p_0) \simeq 0.064$. The dashed line shows the phase transition "boundary", i.e., the value of N such that $M = 2^{ND(p_1,p_0)}$. For $M = 1000$, this value is equal to 155.5. For the known distributions plot, the red line corresponds to the upper bound established by Theorem 1, and it is equal to $1000(0.98)^N$. Similar plots for the case $p_0 = (0.9, 0.1)$ and $p_1 = (0.7, 0.3)$ are shown in Figure 2. As expected, the error curves of Figure 1 are higher than the ones in Figure 2, since the former detection case is "harder" than the latter. The phase transition boundary is depicted in Figure 2 with the dashed line at value $N = 44.9$. The upper bound of Theorem 1 is given by $1000(0.9349)^N$.

5 Conclusions

We have considered a statistical model with $M+1$ sequences of independent random variables, each of length N. All random variables have the same point mass function p_0 except for one sequence, the target, for which the common point mass function is p_1. The error of the maximum likelihood estimator for the target converges to 0 if there exists an $\varepsilon > 0$ such that $M2^{-N(D(p_1,p_0)-\varepsilon)} \to 0$, and it converges to 1 if there exists an $\varepsilon > 0$ such that $M2^{-N(D(p_1,p_0)+\varepsilon)} \to +\infty$. Moreover, when p_0 and p_1 are unknown, we are able to build an estimator of the target with the same performace; this allows us to study the important practical problem of outlier detection. We conjecture that these results can be generalized to the case of ergodic Markov chains, and we plan to report the more general results in a subsequent publication.

6 Proofs

Without loss of generality, we assume that the target line is line number 1.

Proof of Theorem 1:

$$
\begin{aligned}
e(M,N) &= \mathbb{P}\left(\max_{2\le m\le M+1}\sum_{n=1}^{N}\log\frac{p_1(X_m^n)}{p_0(X_m^n)} > \sum_{n=1}^{N}\log\frac{p_1(X_1^n)}{p_0(X_1^n)}\right) && (22)\\
&\le M\mathbb{P}\left(\sum_{n=1}^{N}\log\frac{p_1(X_2^n)}{p_0(X_2^n)} > \sum_{n=1}^{N}\log\frac{p_1(X_1^n)}{p_0(X_1^n)}\right) && (23)\\
&= M\mathbb{P}\left(\prod_{n=1}^{N}\left(\frac{p_1(X_2^n)p_0(X_1^n)}{p_0(X_2^n)p_1(X_1^n)}\right)^s > 1\right), \text{ for all } s>0 && (24)\\
&\le ME\left[\prod_{n=1}^{N}\left(\frac{p_1(X_2^n)p_0(X_1^n)}{p_0(X_2^n)p_1(X_1^n)}\right)^s\right] && (25)\\
&= M\left[E\left(\frac{p_1(X_2^1)p_0(X_1^1)}{p_0(X_2^1)p_1(X_1^1)}\right)^s\right]^N && (26)
\end{aligned}
$$

where (25) is due to the Markov inequality.

Let us define

$$f(s) \triangleq E\left[\left(\frac{p_1(X_2^1)p_0(X_1^1)}{p_0(X_2^1)p_1(X_1^1)}\right)^s\right] \quad \text{and} \quad g(s) \triangleq \ln f(s) \tag{27}$$

One can check that $f'(1/2) = g'(1/2) = 0$. Moreover, using Hölder's inequality, Grimmett et. al. (1992), it is easy to show that, for any $s, t > 0$ and $0 \leq \alpha \leq 1$,

$$E\left[\left(\frac{p_1(X_2^1)p_0(X_1^1)}{p_0(X_2^1)p_1(X_1^1)}\right)^{\alpha s+(1-\alpha)t}\right] \leq \left(E\left[\left(\frac{p_1(X_2^1)p_0(X_1^1)}{p_0(X_2^1)p_1(X_1^1)}\right)^s\right]\right)^{\alpha} \left(E\left[\left(\frac{p_1(X_2^1)p_0(X_1^1)}{p_0(X_2^1)p_1(X_1^1)}\right)^t\right]\right)^{1-\alpha}. \tag{28}$$

By taking the log on both sides, we deduce that g is a convex function of s. Hence, it achieves its minimum value at $s = 1/2$ (therefore, f achieves its minimum value at $s = 1/2$). This leads to the tightest upper bound in (26), i.e.,

$$e(M, N) \leq M f^N(\frac{1}{2}) = M\left(\sum_x \sqrt{p_0(x)p_1(x)}\right)^{2N}. \tag{29}$$

■

In order to prove Theorems 2 and 3, we start with two technical lemmas that will be useful later on.

Lemma 1 *Let U and R be two rvs, and $y \in \mathbb{R}$. Then, for any $\varepsilon > 0$,*

$$\mathbb{P}(U > y + \varepsilon) - \mathbb{P}(R < -\varepsilon) \leq \mathbb{P}(U + R > y) \leq \mathbb{P}(U > y - \varepsilon) + \mathbb{P}(R > \varepsilon) \tag{30}$$

Proof of Lemma 1:

$$\begin{aligned} \mathbb{P}(U + R > y) &= \mathbb{P}(U + R > y, R \leq \varepsilon) + \mathbb{P}(U + R > y, R > \varepsilon) && (31) \\ &\leq \mathbb{P}(U > y - \varepsilon) + \mathbb{P}(R > \varepsilon) && (32) \end{aligned}$$

and

$$\begin{aligned} I\!P(U+R\leq y) &= I\!P(U+R\leq y, R<-\varepsilon)+I\!P(U+R\leq y, R\geq -\varepsilon) \quad (33)\\ &\leq I\!P(U\leq y+\varepsilon)+I\!P(R<-\varepsilon) \quad (34)\end{aligned}$$

which allows us to obtain the lower bound by computing the complementary event. ■

Lemma 2 *Let $(V_1^N,\ldots,V_M^N)$ be a sequence of M independent, identically distributed, discrete random variables. Moreover, assume that the following large deviation property holds for some $z\in I\!R$,*

$$I\!P(V_1^N>z)\doteq 2^{-NI(z)},\ \textit{where } I(z)>0 \textit{ and } a_N\doteq b_N \Leftrightarrow \lim_{N\to+\infty}\frac{1}{N}\log\frac{a_N}{b_N}=0. \quad (35)$$

Then, if

$$\exists\,\varepsilon>0 \quad s.t. \quad \lim_{M,N\to+\infty} M2^{-N(I(z)-\varepsilon)}=0, \quad then \quad \lim_{M,N\to+\infty} I\!P(\max_{1\leq m\leq M} V_m^N>z)=0. \quad (36)$$

Also, if

$$\exists\,\varepsilon>0 \quad s.t. \quad \lim_{M,N\to+\infty} M2^{-N(I(z)+\varepsilon)}=+\infty, \quad then \quad \lim_{M,N\to+\infty} I\!P(\max_{1\leq m\leq M} V_m^N>z)=1. \quad (37)$$

Proof of Lemma 2: Let $\varepsilon>0$ be arbitrarily small. Then, there exists $N_0(\varepsilon)>0$ such that

$$\forall\, N>N_0, \quad \left|\frac{1}{N}\log\left(\frac{I\!P(V_1^N>z)}{2^{-NI(z)}}\right)\right|<\varepsilon. \quad (38)$$

To prove the first part, we start with the following claim:

$$(\exists\,\varepsilon'>0)\ (\forall\, N'>0)\ (\exists\, N>N'):\quad I\!P(\max_{1\leq m\leq M(N)} V_m^N>z)>\varepsilon'.$$

Then, using the union bound, we obtain

$$(\exists\,\varepsilon'>0)\ (\forall\,N'>0)\ (\exists\,N>N'):\quad \sum_{m=1}^{M(N)} I\!P(V_m^N>z)=M(N)I\!P(V_1^N>z)>\varepsilon'. \tag{39}$$

By picking $N'>N_0(\varepsilon)$, (39) becomes

$$(\exists\,\varepsilon'>0)\ (\forall\,N'>N_0(\varepsilon))\ (\exists\,N>N'):\quad M2^{-N(I(z)-\varepsilon)}>\varepsilon'.$$

Hence, $M2^{-N(I(z)-\varepsilon)}$ does not converge to zero for any $\varepsilon>0$, as required.

To prove the second part, we first assume that $N>N_0(\varepsilon)$, as above. Then,

$$\begin{aligned}
I\!P(\max_{1\le m\le M} V_m^N>z) &= 1-I\!P(\max_{1\le m\le M} V_m^N\le z) && (40)\\
&= 1-I\!P^M(V_1^N\le z) && (41)\\
&= 1-2^{M\log(1-I\!P(V_1^N>z))} && (42)\\
&\ge 1-2^{-MI\!P(V_1^N>z)} && (43)\\
&\ge 1-2^{-M2^{-N(I(z)+\varepsilon)}}, && (44)
\end{aligned}$$

where the first inequality is a consequence of the inequality $\log(1-x)\le -x$, and the second inequality arises from (38). Note that (44) is true for any arbitrary $\varepsilon>0$. Hence, if there exists $\varepsilon>0$ such that $M2^{-N(I(z)+\varepsilon)}\to+\infty$, then necessarily $I\!P(\max_{1\le m\le M} V_m^N>z)\to 1$. This concludes the proof of the second part, and the proof of the lemma. ■

We are now ready for

Proof of Theorem 2:

$$\begin{aligned}
e(M,N) &= I\!P\left(\max_{2\le m\le M+1}\frac{1}{N}\sum_{n=1}^{N}\log\frac{p_1(X_m^n)}{p_0(X_m^n)}>\frac{1}{N}\sum_{n=1}^{N}\log\frac{p_1(X_1^n)}{p_0(X_1^n)}\right) && (45)\\
&= I\!P(U_M^N+R_N>D(p_1,p_0)),\ \text{ where} && (46)
\end{aligned}$$

$$U_M^N = \max_{2\le m\le M+1} \frac{1}{N}\sum_{n=1}^{N} \log\frac{p_1(X_m^n)}{p_0(X_m^n)} \quad \text{and} \tag{47}$$

$$R_N = D(p_1,p_0) - \frac{1}{N}\sum_{n=1}^{N} \log\frac{p_1(X_1^n)}{p_0(X_1^n)} \tag{48}$$

From the law of large numbers, $R_N \to 0$ in probability. Hence, for all $\eta > 0$ and $\alpha > 0$, and for N sufficiently large, using Lemma 1,

$$I\!P(U_M^N > D(p_1,p_0)+\eta) - \alpha \le e(M,N) \le I\!P(U_M^N > D(p_1,p_0)-\eta) + \alpha \tag{49}$$

Let us define

$$V_m^N \triangleq \frac{1}{N}\sum_{n=1}^{N} \log\frac{p_1(X_m^n)}{p_0(X_m^n)} \tag{50}$$

Now, using Sanov's theorem, Dembo at. al.(1998),

$$I\!P(V_2^N \ge D(p_1,p_0)) \doteq 2^{-ND(p_1,p_0)} \tag{51}$$

Indeed,

$$I\!P(V_2^N \ge D(p_1,p_0)) \doteq 2^{-ND(p^*,p_0)} \tag{52}$$

where

$$D(p^*,p_0) = \inf_{p\in\mathcal{C}} D(p,p_0), \text{ with } \mathcal{C} = \{p; E_p \log\frac{p_1}{p_0} \ge D(p_1,p_0)\}, \tag{53}$$

and for $p \in \mathcal{C}$,

$$D(p,p_0) = E_p \log\frac{p}{p_0} = D(p,p_1) + E_p\log\frac{p_1}{p_0} \tag{54}$$

$$\ge D(p,p_1) + D(p_1,p_0) \ge D(p_1,p_0). \tag{55}$$

Now, by continuity of the rate function, there exists $\varepsilon > 0$ such that

$$I\!P(V_2^N > D(p_1,p_0)-\eta) \doteq 2^{-N(D(p_1,p_0)-\varepsilon)} \tag{56}$$

and there exists $\varepsilon' > 0$ such that

$$I\!P(V_2^N > D(p_1, p_0) + \eta) \doteq 2^{-N(D(p_1,p_0)+\varepsilon')} \tag{57}$$

Finally, since

$$U_M^N = \max_{2 \le m \le M+1} V_m^N \tag{58}$$

and the rvs $V_2^N, \ldots, V_{M+1}^N$ are iid, we obtain the required result from Lemma 2. ■

Proof of Theorem 3: We proceed along the same lines as for the proof of Theorem 2.

$$\begin{aligned}
\tilde{e}(M,N) &= I\!P\left(\max_{2\le m\le M+1} \frac{1}{N}\sum_{n=1}^{N} \log\frac{\hat{p}_m(X_m^n)}{\hat{p}_{(m)}(X_m^n)} > \frac{1}{N}\sum_{n=1}^{N}\log\frac{\hat{p}_1(X_1^n)}{\hat{p}_{(1)}(X_1^n)}\right) && (59)\\
&\le I\!P(U_M^N + R_M^N > D(p_1,p_0)), \quad \text{with} && (60)\\
U_M^N &= \max_{2\le m\le M+1} \frac{1}{N}\sum_{n=1}^{N}\log\frac{\hat{p}_m(X_m^n)}{p_0(X_m^n)} && (61)\\
R_M^N &= A_M^N + B^N + C^N && (62)\\
A_M^N &= \max_{2\le m\le M+1} \frac{1}{N}\sum_{n=1}^{N}\log\frac{p_0(X_m^n)}{\hat{p}_{(m)}(X_m^n)} && (63)\\
B^N &= \frac{1}{N}\sum_{n=1}^{N}\log\frac{\hat{p}_{(1)}(X_1^n)}{p_0(X_1^n)} && (64)\\
C^N &= D(p_1,p_0) - \frac{1}{N}\sum_{n=1}^{N}\log\frac{\hat{p}_1(X_1^n)}{p_0(X_1^n)} && (65)
\end{aligned}$$

For all $\eta > 0$, from Lemma 1,

$$\tilde{e}(M,N) \le I\!P(U_M^N > D(p_1,p_0) - \eta) + I\!P(R_M^N > \eta). \tag{66}$$

Let

$$V_m^N = \frac{1}{N}\sum_{n=1}^{N}\log\frac{\hat{p}_m(X_m^n)}{p_0(X_m^n)}, \quad 2 \le m \le M+1. \tag{67}$$

Using Sanov's theorem,

$$I\!P(V_2^N \geq D(p_1, p_0)) \doteq 2^{-ND(p_1,p_0)}. \tag{68}$$

Indeed,

$$I\!P(V_2^N \geq D(p_1, p_0)) \doteq 2^{-ND(p^*,p_0)}, \tag{69}$$

where

$$D(p^*, p_0) = \inf_{p \in \mathcal{C}} D(p, p_0), \text{ with } \mathcal{C} = \{p; E_p \log \frac{p}{p_0} \geq D(p_1, p_0)\}. \tag{70}$$

And for $p \in \mathcal{C}$,

$$D(p, p_0) = E_p \log \frac{p}{p_0} \geq D(p_1, p_0). \tag{71}$$

Now, by continuity of the rate function, there exists $\varepsilon > 0$ such that

$$I\!P(V_2^N > D(p_1, p_0) - \eta) \doteq 2^{-N(D(p_1,p_0)-\varepsilon)}. \tag{72}$$

To show that (66) approaches zero as $M, N \to \infty$ with $M2^{N(D(p_1,p_0)-\varepsilon)} \to 0$, it suffices to prove that $R_M^N \to 0$, since the term $I\!P(U_M^N > D(p_1, p_0) - \eta)$ of (66) goes to zero by virtue of (72), Lemma 2, and the fact that

$$U_M^N = \max_{2 \leq m \leq M+1} V_m^N, \tag{73}$$

and the rvs $V_2^N, \ldots, V_{M+1}^N$ are iid.

Using the law of large numbers, $C^N \to 0$ in probability. Also,

$$\begin{aligned} I\!P(A_M^N > \eta) &\leq M I\!P\left(\frac{1}{N}\sum_{n=1}^{N} \log \frac{p_0(X_2^n)}{\hat{p}_{(2)}(X_2^n)} > \eta\right) &(74)\\ &\leq M I\!P\left(\max_{1 \leq n \leq N} \log \frac{p_0(X_2^n)}{\hat{p}_{(2)}(X_2^n)} > \eta\right) &(75)\\ &\leq MN I\!P\left(\log \frac{p_0(X_2^1)}{\hat{p}_{(2)}(X_2^1)} > \eta\right) &(76) \end{aligned}$$

$$\leq \quad MN \max_x \mathbb{P}\left(\log \frac{p_0(x)}{\hat{p}_{(2)}(x)} > \eta\right) \tag{77}$$

$$= \quad MN \max_x \mathbb{P}(\hat{p}_{(2)}(x) < 2^{-\eta} p_0(x)) \tag{78}$$

$$\leq \quad MN \max_x 2^{-N(M-1)I(x,\eta)} = MN 2^{-N(M-1)J(\eta)} \tag{79}$$

where $I(x,\eta) > 0$ is a rate function, and $J(\eta) = \min_x I(x,\eta)$. The last inequality comes from the fact that $\hat{p}_{(2)}(x) \to p_0(x)$ in probability. A similar argument shows that $B^N \to 0$ in probability as well.

Note that the result would still hold if we replaced $\hat{p}_{(m)}$ with $\hat{p}$, i.e., with the empirical distribution over the full data. ■

References

[1] Amir Dembo and Ofer Zeitouni. *Large Deviations Techniques and Applications*. Springer Verlag, 1998.

[2] D. Geman and B. Jedynak. An active testing model for tracking roads from satellite images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(1):1–14, January 1996.

[3] G.R. Grimmett and D. R. Stirzaker. *Probability and Random Processes*. Oxford Science Publications, 1992.

[4] Alan L. Yuille and James M. Coughlan. Fundamental limits of bayesian inference: Order parameters and phase transitions for road tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(2):160–173, February 2000.

[5] Alan L. Yuille, James M. Coughlan, Yingnian Wu, and Song Chun Zhu. Order parameters for detecting target curves in images: When does high level knowledge help? *International Journal of Computer Vision*, 41:9–33, 2001.