Proceedings of Machine Learning Research vol 313:1–38, 2026 37th International Conference on Algorithmic Learning Theory

Nearly Minimax Discrete Distribution Estimation in Kullback-Leibler Divergence with High Probability

Dirk van der Hoeven DIRK@DIRKVANDERHOEVEN.COM
Leiden University, the Netherlands

Julia Olkhovskaia JULIA.OLKHOVSKAYA@GMAIL.COM
TU Delft, the Netherlands

Tim van Erven TIM@TIMVANERVEN.NL
University of Amsterdam, the Netherlands

Editors: Matus Telgarsky and Jonathan Ullman

Abstract

We consider the fundamental problem of estimating a discrete distribution on a domain of size K with high probability in Kullback-Leibler divergence. We provide upper and lower bounds on the minimax estimation rate, which show that the optimal rate is between $\big(K + \ln(K)\ln(1/\delta)\big)/n$ and $\big(K \ln\ln(K) + \ln(K)\ln(1/\delta)\big)/n$ at error probability δ and sample size n, which pins down the rate up to the doubly logarithmic factor $\ln\ln K$ that multiplies K. Our upper bound uses techniques from online learning to construct a novel estimator via online-to-batch conversion. Perhaps surprisingly, the tail behavior of the minimax rate is worse than for the squared total variation and squared Hellinger distance, for which it is $\big(K + \ln(1/\delta)\big)/n$, i.e. without the $\ln K$ multiplying $\ln(1/\delta)$. As a consequence, we cannot obtain a fully tight lower bound from the usual reduction to these smaller distances. Moreover, we show that this lower bound cannot be achieved by the standard lower bound approach based on a reduction to hypothesis testing, and instead we need to introduce a new reduction to what we call weak hypothesis testing. We investigate the source of the gap with other divergences further in refined results, which show that the total variation rate is achievable for Kullback-Leibler divergence after all (in fact by the maximum likelihood estimator) if we rule out outcome probabilities smaller than $O(\ln(K/\delta)/n)$, which is a vanishing set as n increases for fixed K and δ. This explains why minimax Kullback-Leibler estimation is more difficult than asymptotic estimation.

Keywords: Discrete distribution estimation, Kullback-Leibler divergence

1. Introduction

Consider a sample $S = (X_1, \dots, X_n)$ of n independent, identically distributed observations of a random variable X with a finite number of possible values $[K] := \{1, \dots, K\}$. The aim of discrete distribution estimation (Devroye and Lugosi, 2001; Diakonikolas, 2016; Canonne, 2020; Polyanskiy and Wu, 2025) is to approximate the unknown true probability mass function p^* of X by an estimate $\widehat{p}$ based on the sample S, in terms of a given measure of distance or divergence $\mathrm{d}(p^*, \widehat{p})$. In this work we consider the case where d is the Kullback-Leibler (KL) divergence of p^* from $\widehat{p}$:

$$\mathrm{KL}(p^* \| \widehat{p}) = \sum_{i=1}^{K} p^*(i) \ln \frac{p^*(i)}{\widehat{p}(i)},$$

with the understanding that $0 \ln 0/a = 0$ for any $a \geq 0$ and $b \ln(b/0) = \infty$ for $b > 0$. Specifically, we are interested in the minimax rate with high probability: for any given $\delta \in (0, 1)$, what is the smallest bound $r_n^*(\delta)$ on the KL divergence that can the guaranteed by any estimator $\widehat{p}$ uniformly over all possible true distributions p^* with confidence at least $1-\delta$, as n grows large? This question is equivalent to determining the sample complexity in PAC-learning with the log loss $\ell(x,p) = -\ln p(x)$, for which the Kullback-Leibler divergence equals the excess risk:

$$\mathrm{KL}(p^*\|\widehat{p}) = \mathbb{E}[\ell(X,\widehat{p})] - \mathbb{E}[\ell(X,p^*)].$$

This is of special interest, because the log loss is unbounded, so it provides perhaps the simplest setting in which to study unbounded losses.

At first sight, it may seem natural to conjecture that the minimax rate might be of order $\big(K + \ln(1/\delta)\big)/n$. Although this will turn out to be false, we start by reviewing the many known results that point in this direction.

One clue comes from the properties of the maximum likelihood estimator (MLE) $\bar{p}_n(i) = n_i/n$, where n_i is the number of times that outcome i occurs in S. The MLE achieves the minimax rate of order $\sqrt{\frac{K+\ln(1/\delta)}{n}}$ for, amongst others, the total variation distance $V(p^*,\widehat{p}) = \sum_{i=1}^K |p^*(i) - \widehat{p}(i)|$ and the Hellinger distance $H(p^*,\widehat{p}) = \sqrt{\sum_{i=1}^K \big(\sqrt{p^*(i)} - \sqrt{\widehat{p}(i)}\big)}$ (Canonne, 2020).[1] Since

$$\tfrac{1}{4}V(p,q)^2 \leq H(p,q)^2 \leq \mathrm{KL}(p\|q) \qquad \text{for any } p,q \tag{1}$$

(Tsybakov, 2009), it follows that the minimax rate for the KL divergence is at least of order $\frac{K+\ln(1/\delta)}{n}$. Furthermore, Agrawal (2022) shows that this is also the minimax rate for KL with reversed arguments, $\mathrm{KL}(\widehat{p}\|p^*)$, achieved again by the MLE. Finally, as pointed out by Mourtada (2025), if we fix p^* and let n tend to infinity, then the MLE is asymptotically normally distributed around p^*, and consequently a second-order Taylor approximation of the KL divergence implies (by the second-order delta method, van der Vaart, 1998) the following convergence in distribution:

$$n\,\mathrm{KL}(p^*\|\bar{p}_n) \;\rightsquigarrow\; \tfrac{1}{2}Y, \qquad \text{where } Y \sim \chi^2_{M-1}. \tag{2}$$

Here, χ^2_{M-1} denotes a χ^2 distribution with $M-1$ degrees of freedom, and $M \leq K$ is the size of the support of p^*. See Theorem 15 in Appendix A for details. By concentration of Y around its mean (Laurent and Massart, 2000), we have that $Y \leq C(M-1+\ln(1/\delta))$ with probability at least $1-\delta$ for some constant $C > 0$, which leads to

$$\mathrm{KL}(p^*\|\bar{p}_n) \leq C\Big(\frac{M+\ln(1/\delta)}{n}\Big) \quad \text{for all sufficiently large } n, \qquad \text{if } p^* \text{ and } \delta \text{ are fixed.} \tag{3}$$

Unfortunately, it is readily seen that the MLE fails to achieve the minimax rate for our setting of interest: suppose there exists an outcome i that has positive probability $p^*(i) > 0$ but has not been observed in the sample: $n_i = 0$. Then $p^*(i)/\bar{p}_n(i) = \infty$ and consequently $\mathrm{KL}(p^*\|\bar{p}_n) = \infty$. But, for any n, it is possible to take $p^*(i)$ small enough that this happens with probability larger than δ, so the MLE cannot guarantee any finite minimax rate.

A common solution is to smooth the MLE by adding a fake number of observations $\gamma > 0$ to each possible outcome. The resulting add-γ estimator is $p^\gamma(i) = \frac{n_i+\gamma}{n+K\gamma}$. For $\gamma = 1$, this is also called the Laplace estimator, and for $\gamma = 1/2$ it is known as the Krichevsky-Trofimov (KT) estimator (Krichevsky and Trofimov, 1981). Following a sequence of prior results (Cover, 1972;

1. NB There exist multiple conventions in the literature for the scale factors in the definitions of V and H.

Krichevsky and Trofimov, 1981; Braess et al., 2002; Paninski, 2004), Braess and Sauer (2004) show that a variant of the add-γ estimator with data-dependent smoothing achieves the minimax rate *in expectation*:

$$\min_{\widehat{p}} \max_{p^*} \ \mathbb{E}_{p^*}[\mathrm{KL}(p^* \| \widehat{p})] \ = \ \frac{K-1}{2n} + o\Big(\frac{1}{n}\Big),$$

and Catoni (1997) shows that the Laplace estimator achieves this rate up to a factor of 2:

$$\mathbb{E}_{p^*}[\mathrm{KL}(p^* \| p^1)] \leq \frac{K-1}{n} \qquad \text{for all } p^*. \tag{4}$$

See also the proofs by Mourtada and Gaïffas (2022) and, for $K = 2$, by Forster and Warmuth (2002). Kamath et al. (2015) further characterize the minimax rate in expectation for, amongst others, the ℓ_2^2, ℓ_1 and χ^2 divergences.

Although suggestive, since all results mentioned so far are for different settings, the only formal implication for our setting is that $\big(K + \ln(1/\delta)\big)/n$ is a lower bound on the minimax rate. This is not matched by any of the following known upper bounds: First, a sequence of increasingly tight bounds for the Laplace estimator have been improved from order $K \ln(n) \ln(K/\delta)/n$ (Bhattacharyya et al., 2021) to $\frac{K + \sqrt{K} \ln(1/\delta)^3}{n}$ (Han et al., 2021) to the tightest known guarantee by Canonne et al. (2023), who show that

$$\mathrm{KL}(p^* \| p^1) \leq \mathbb{E}[\mathrm{KL}(p^* \| p^1)] + \frac{C\sqrt{K \ln(K/\delta)^5}}{n} \leq \frac{K-1}{n} + \frac{C\sqrt{K \ln(K/\delta)^5}}{n} \tag{5}$$

with probability at least $1 - \delta$, for some absolute constant $C > 0$, where the second inequality holds deterministically by (4). The remaining gap with the lower bound is seen both in the exponent $5/2$ on $\ln(1/\delta)$ and in the factor $\sqrt{K}$ that multiplies it. Canonne et al. (2023) further show that very little further improvement is possible via concentration of $\mathrm{KL}(p^* \| p^1)$ around its mean, because the deviation in (5) cannot be improved to $\frac{CK^\eta \ln(1/\delta)}{n}$ for any $\eta < 1/2$, at least not uniformly for all δ.

An alternative estimator is proposed by Van der Hoeven et al. (2023), who obtain a novel type of regret guarantee for their online learning algorithm, and then convert the algorithm to an estimator using online-to-batch (OTB) conversion (Littlestone, 1989). This estimator guarantees that

$$\mathrm{KL}(p^* \| \widehat{p}) \leq C \frac{K + \ln(n) \ln(1/\delta)}{n}$$

with probability at least $1 - \delta$ for any p^*, where $C > 0$ is an absolute constant. This improves the exponent on $\ln(1/\delta)$ compared to (5), but leaves a multiplicative factor $\ln(n)$ that is not present in the lower bound.

Main Results We provide improved upper and lower bounds for discrete distribution estimation in KL divergence, which show that the minimax rate is sandwiched between

$$C_1 \Big(\frac{K + \ln(K) \ln(1/\delta)}{n} \Big) \ \leq \ r_n^*(\delta) \ \leq \ C_2 \Big(\frac{K \ln\ln(K) + \ln(K) \ln(1/\delta)}{n} \Big) \tag{6}$$

for absolute constants $0 < C_1 \leq C_2$. This characterizes the exact minimax rate up to a doubly logarithmic factor in K.

Our upper bound is obtained by converting an online learning algorithm to an estimator p^{OTB} using online-to-batch conversion with suffix-averaging, i.e. averaging only the predictions on the second half of the data (Rakhlin et al., 2012; Harvey et al., 2019; Aden-Ali et al., 2023). In fact, we obtain a potentially stronger bound: there exists $C > 0$ such that

$$\mathrm{KL}(p^* \| p^{\mathrm{OTB}}) \leq C \frac{K + \widetilde{J} \ln(\ln(\widetilde{J})) + \ln(J) \ln(1/\delta)}{n} \tag{7}$$

with probability at least $1 - \delta$, where $J \leq K$ is the number of outcomes i occurring fewer than $32 \ln(K/\delta)$ times in the first half of the data $X_1, \ldots, X_{n/2}$ and $\widetilde{J} \leq K$ is the number of i for which $p^*(i) < 32 \ln(K/\delta)/n$, except that both J and $\widetilde{J}$ are taken to be at least 3.

In addition, we also analyze an alternative estimator $\widetilde{p}$, which is a variant of the add-γ estimator in which γ is estimated on a hold-out set and is allowed to differ between outcomes i. This estimator satisfies

$$\mathrm{KL}(p^* \| \widetilde{p}) \leq C \Big(1 + \ln\big(\min\{\ln(K/\delta), J\}\big)\Big) \frac{K + \ln(1/\delta)}{n}, \tag{8}$$

with probability at least $1 - \delta$, for some $C > 0$. Although this gives a slightly worse bound than (7) for the worst case that $\widetilde{J} = J = K$ (see Remark 4), it is still within a doubly logarithmic factor of the lower bound, and the estimator is simpler.

Finally, as a side-result, we return to the MLE: as discussed above, it works asymptotically, but may fail for any finite n if there exist outcomes i with $p^*(i) > 0$ too small compared to n. We provide a threshold for when probabilities are 'too small' by showing that, with probability at least $1 - \delta$,

$$\mathrm{KL}(p^* \| \bar{p}_n) \leq C \frac{K + \ln(1/\delta)}{n} \qquad \text{for all } p^* \text{ such that } p^*(i) \geq \frac{32 \ln(K/\delta)}{n} \text{ for all } i, \tag{9}$$

for some absolute constant $C > 0$. Since this is again the asymptotic rate, which beats the lower bound in (6), we also see that these small probabilities cause a difference in the rate. The result in (9) will actually be proved via a bound on the χ^2 divergence, which can be arbitrarily larger than the KL divergence in general, but is within a constant factor with high probability in this case.

Regarding the lower bound: when K/n dominates $\ln(K) \ln(1/\delta)/n$, the lower bound follows directly from the known lower bound for squared variational distance (Canonne, 2020) and (1). (We still include a proof for lack of a reference that explicitly spells out the details.) The interesting part of the lower bound is the case when $\ln(K) \ln(1/\delta)/n$ is dominant. We first obtain this rate from an estimator-dependent lower bound, which makes explicit the trade-off when smoothing probability estimates for outcomes i that have not been observed, i.e. for which $n_i = 0$. This lower bound further implies that the only add-γ estimators that can achieve a rate of $\ln(K) \ln(1/\delta)/n$ must have $\gamma \propto \frac{\ln(1/\delta)}{K}$, whereas any constant γ that does not depend on δ, like in the Laplace and KT estimators, will lead to a rate with order $\ln(1/\delta) \ln \ln(1/\delta)/n$ dependence on δ, which is worse than $\ln(K) \ln(1/\delta)$ in the regime where $\ln(K) \ln(1/\delta)$ dominates K. Although the estimator-dependent bound is informative, it is proved using an ad hoc argument, so we further investigate whether the same lower bound can also be obtained from a general technique. Surprisingly, we find that neither Fano's inequality nor any other approach based on the standard reduction to hypothesis testing (Tsybakov, 2009) can work. Instead, in order to recover the correct rate, we need to introduce a new type of reduction to what we call a *weak hypothesis testing* problem, where the goal is not to identify the true distribution from a finite set of candidates, but merely to identify any incorrect distribution.

Concurrent Work In very insightful independent and concurrent work, Mourtada (2025) derives matching lower and upper bounds (up to constants) for the same setting we consider.[2] For estimators that are allowed to depend on δ, he shows that the lower bound in (6) is actually tight, and that it is in fact achieved by the add-γ estimator with $\gamma = 1 \vee \frac{\ln(1/\delta)}{K}$. This is possible by his crucial observation that the standard Chernoff method based on the moment generating function cannot provide tail bounds of the right form, so instead he proceeds by controlling moments directly. His lower bound uses the same construction as our estimator-dependent lower bound. Mourtada further considers estimators that do not depend on δ, where he shows that the Laplace estimator achieves:

$$\mathrm{KL}(p^*\|p^1) \leq C\frac{K + \ln(1/\delta)\ln\ln(1/\delta)}{n}$$

with probability at least $1 - 4\delta$, for some $C > 0$, which matches the rate in our estimator-dependent lower bound. He also shows that this is the minimax rate among the restricted, but large class of all estimators $\widehat{p}$ that a) do not depend on δ and b) achieve $\mathrm{KL}(p^*\|\widehat{p}) \leq \kappa K/n$ with positive probability for all p^*, where $\kappa \geq 1$ can be any absolute constant. Finally, Mourtada also provides results for p^* with sparse support, which are of interest in the regime where K is large compared to n.

Further Related Work A related goal is to minimize the minimax regret in an online learning setting with logarithmic loss, which is known to be $\frac{K-1}{2}\ln(n) + O(1)$ and is achieved by the KT estimator. For $K = 2$, this was shown by Krichevsky and Trofimov (1981), and Xie and Barron (1997) generalize it to general K. For a more comprehensive overview we refer to Shtarkov (1987); Merhav and Feder (1998); Rissanen (1996). Our bounds for p^{OTB} do not directly arise from combining these results with online-to-batch conversion, but instead rely on a new analysis of suffix averaging to avoid the $\ln(n)$ factor that appears in the minimax regret.

Outline The remainder of the paper is organised as follows. In Section 2 we provide detailed statements for our upper bounds as well as the proofs for our main results. In Section 3 we present detailed statements of our lower bounds. Finally, we provide a brief discussion and outlook in Section 4.

Notation For $\delta \in (0, 1]$, the worst-case rate of an estimator $\widehat{p}$ at confidence level $1 - \delta$ is

$$r_n(\widehat{p}, \delta) = \inf\left\{r_n \mid \sup_{p^* \in \triangle^K} \mathbb{P}_{p^*}\left(\mathrm{KL}(p^*\|\widehat{p}) > r_n\right) \leq \delta\right\}, \tag{10}$$

where $\triangle^K = \{p : p(i) \geq 0, \sum_{i=1}^K p(i) = 1\}$. Then $r_n^*(\delta) = \inf_{\widehat{p}} r_n(\widehat{p}, \delta)$ is the minimax rate. We further let $B(n, p)$ denote the binomial distribution with n trials and success probability p, and we write $M(n, p(1), \dots, p(K))$ for the multinomial distribution with n trials and probabilities $p(1), \dots, p(K)$.

2. Upper Bounds

Here we provide detailed statements of our upper bounds. We start with definitions we use throughout this section. For simplicity, we assume that $n/2$ is an integer throughout the section. We will

2. *Note on concurrency:* The results in this paper were developed independently and contemporaneously with those of Mourtada. An earlier version of our work containing these results, available as (van der Hoeven et al., 2025), was submitted to COLT in February 2025, prior to the appearance of Mourtada's April 2025 arXiv preprint, thereby establishing a verifiable submission record. We subsequently improved the presentation and submitted the present version to ALT in October 2025.

consider a version of the add-γ estimator based on the first t data points and where $\gamma_i \geq 0$ is allowed to differ between outcomes i: let $p_{t+1}(i) = \frac{n_i(t)+\gamma_i}{t+\sum_{i=1}^K \gamma_i}$, where $n_i(t) = \sum_{s=1}^t \mathbb{1}[X_s = i]$. Note that, in particular, p_{n+1} is the estimator obtained when using the whole sample. Let $n_i = n_i(n)$, $m_i = n_i(n/2)$, $\bar{p}_t(i) = n_i(t)/t$, and $p^+(i) = \frac{p^*(i)n+\gamma_i}{n+\sum_{i=1}^K \gamma_i}$. Let $\mathcal{J} = \{i : m_i < 32\ln(K/\delta)\}$, $J = \max\{3, |\mathcal{J}|\}$, $\widetilde{\mathcal{J}} = \{i : np^*(i) < 32\ln(K/\delta)\}$, and $\widetilde{J} = \max\{3, |\widetilde{\mathcal{J}}|\}$.

Our first upper bound is for the maximum likelihood estimator:

Theorem 1 *For some $C > 0$, if $|\widetilde{\mathcal{J}}| = 0$, then, with probability at least $1-\delta$, the maximum likelihood estimator guarantees*

$$\tfrac{1}{6}\chi^2(p^*, \bar{p}_n) \leq \tfrac{1}{4}\chi^2(\bar{p}_n, p^*) \leq \mathrm{KL}(p^* \| \bar{p}_n) \leq C\frac{K + \ln(1/\delta)}{n}\,.$$

The proof of Theorem 1 can be found in Appendix C. Theorem 1 tells us that if p^* is in the interior of the simplex (i.e. $|\widetilde{\mathcal{J}}| = 0$), then the empirical mean is a good estimator in the sense that $\frac{1}{6}\chi^2(p^*, \bar{p}_n)$, $\chi^2(\bar{p}_n, p^*)$, and $\mathrm{KL}(p^* \| \bar{p}_n)$ are at most of order $(K + \ln(1/\delta))/n$ with probability at least $1-\delta$. However, the condition $|\widetilde{\mathcal{J}}| = 0$ is not something we observe, which implies that we cannot tell if $\bar{p}_n$ is a good estimator. Furthermore, Theorem 1 does not provide insight into what estimator one should use if $|\widetilde{\mathcal{J}}| > 0$. In fact, if $|\widetilde{\mathcal{J}}| > 0$, it can be seen that the maximum likelihood estimator cannot guarantee that $\mathrm{KL}(p^* \| \bar{p}_n)$ is finite with probability at least $1-\delta$ (Section 3.1).

In the remainder of this section we will provide upper bounds that apply more broadly. We will make use of the add-γ estimator and $p^{\mathrm{OTB}} = \frac{2}{n}\sum_{t=n/2+1}^n p_t$, both with

$$\gamma_i = \begin{cases} 0 & \text{if } \ m_i \geq 32\ln(4K/\delta) \\ \max\{1, \frac{\ln(K/\delta)}{J}\} & \text{otherwise.} \end{cases} \tag{11}$$

Note that we set $\gamma_i = 0$ if m_i is sufficiently large. This allows us to avoid unnecessary bias when possible. We first provide an upper bound for the add-γ estimator:

Theorem 2 *Suppose that $n \geq K$ and $n \geq \ln(K/\delta)$. Then, with probability at least $1-2\delta$, the add-γ estimator with γ_i set according to (11) guarantees that*

$$\mathrm{KL}(p^* \| p_{n+1}) \leq \begin{cases} (2 + \ln(40\min\{J, \ln(K/\delta)\}))\left(\frac{9K+9\ln(1/\delta)}{n}\right) & \text{if } \ |\mathcal{J}| \geq 1 \\ (2+\ln(2))\left(\frac{7K+6\ln(1/\delta)}{n}\right) & \text{if } \ |\mathcal{J}| = 0\,. \end{cases}$$

Our lower bounds in Section 3 suggest that the $\ln(\min\{J, \ln(K)\})K$ term is suboptimal in the worst case. Indeed, in Theorem 3 we show that the following estimator provides a stronger guarantee:

$$p^{\mathrm{OTB}} = \frac{2}{n}\sum_{t=n/2+1}^{n} p_t, \tag{12}$$

where p_t uses γ_i from (11). This estimator is an instance of online-to-batch (OTB) conversion in which we average only over the second half of the data, which is called suffix averaging. As mentioned in Section 1, standard OTB conversions that average over all $t = 1, \ldots, n$, would introduce unnecessary $\ln(n)$ terms, which is why we resort to suffix averaging. For p^{OTB} we obtain the following guarantee:

Theorem 3 *Suppose that $\ln(K/\delta) \leq n$ and that $K \leq n$. Then, with probability at least $1 - 5\delta$, the OTB estimator defined in (12) guarantees that*

$$\mathrm{KL}(p^* \| p^{\mathrm{OTB}}) \leq \frac{1}{n}\Big(128K + 200\ln(800J)\ln(1/\delta) + 8\widetilde{J}\ln\left(24\left(\ln\left(\frac{\widetilde{J}}{\ln(1/\delta)}\right) \vee 1\right)\right)\Big).$$

Ignoring constants, the gap between our lower bounds in Section 3 and the guarantee of p^{OTB} is an additive $\widetilde{J}\ln\left(\ln\left(\frac{\widetilde{J}}{\ln(1/\delta)}\right) \vee 1\right)$. In the worst case, this term is of order $K\ln(\ln(K))$.

Remark 4 *To see that Theorem 3 strictly improves on Theorem 2 in the worst case that $\widetilde{J} = J = K$, consider the following case distinction: If $K \leq \ln(1/\delta)$, then both theorems give a rate of order $\ln(K)\ln(1/\delta)$ and the rates are the same. If $K \geq \ln(1/\delta)$ then the rate for the add-γ estimator in Theorem 3 is of order $K\ln\ln K + \ln(K)\ln(1/\delta)$ and the rate in Theorem 2 is of order $K\ln\big(\ln(K) + \ln(1/\delta)\big)$. We see that the latter always exceeds $K\ln\ln K$, and is also at least of order $K\ln\ln(1/\delta) \gtrsim \ln(K)\ln(1/\delta)$.*

Both Theorem 2 and 3 provide a stronger guarantee as $J = \max\{3, |\mathcal{J}|\}$ decreases. The reason for this behavior is that we set $\gamma_i = 0$ for $i \in \mathcal{J}$, which allows us to avoid unnecessary bias. The reason we can set $\gamma_i = 0$ if $i \in \mathcal{J}$ is because of the following lemma:

Lemma 5 *Suppose that $n \geq 4$. Let $\zeta_i = 3\left(1 + \frac{2(3m_i + 27\ln(4K/\delta))}{\max\{\gamma_i, \frac{1}{2}m_i - 9\ln(4K/\delta)\}}\right)$ and $\xi = 1 + \frac{\sum_{i=1}^K \gamma_i}{n}$. For all $t \in [n/2+1, \dots, n+1]$, $i \in [K]$, and $\delta \in (0,1)$, with probability at least $1 - \delta$,*

$$\frac{p^*(i)}{p_t(i)} \leq \zeta_i \xi, \qquad \frac{p_t(i)}{p^+(i)} \leq \zeta_i, \qquad \frac{p^+(i)}{p_t(i)} \leq 1 + \zeta_i \xi, \qquad \frac{p^*(i)}{p^+(i)} \leq \xi, \qquad \frac{n_i + \gamma_i}{m_i + \gamma_i} \leq 6 + \zeta_i.$$

Lemma 5 allows us to control the density ratio between (a close approximation of) p^* and our estimators. We use Lemma 5 for two purposes. The first purpose is to control the range of the excess loss in order to apply a concentration inequality (a version of Bernstein's inequality, Lemma 21) to relate $\mathrm{KL}(p^* \| p^{\mathrm{OTB}})$ to the regret of an online learning algorithm. Van der Hoeven et al. (2023) use a different OTB conversion but use the same concentration inequality to relate $\mathrm{KL}(p^* \| \widehat{p})$ to a different regret, which ultimately leads to a $O((K + \ln(n)\ln(1/\delta))/n)$ bound. The $\ln(n)$ term in their bound does not come from the online-to-batch conversion but from a weaker control on the range of $\ln\left(\frac{p^*(i)}{p_t(i)}\right)$ than we obtain from Lemma 5. The second purpose of Lemma 5 is to relate different divergences to each other:

Lemma 6 *For any pair $p, q \in \triangle^K$, suppose that $\frac{1}{2} \leq \frac{p(i)}{q(i)} \leq 2$ for all $i \in [K]$. Then*

$$\tfrac{1}{6}\chi^2(p,q) \leq \tfrac{1}{4}\chi^2(q,p) \leq \mathrm{KL}(p\|q) \leq \tfrac{5}{2}H^2(p,q) \leq \tfrac{5}{2}\,\mathrm{KL}(q\|p) \leq \tfrac{5}{2}\chi^2(q,p) \leq 5\chi^2(p,q).$$

The proof of Lemma 6 can be found in Appendix C. Lemma 6 tells us that, if the density ratio between two distributions is bounded by a constant, then several divergence measures are equivalent up to constants. This is useful because for several of these divergences we already have optimal high-probability guarantees. Now, Lemma 5 tells us that, with high probability, as long as either all m_i are sufficiently big or we add sufficient bias in the form of γ_i, the density ratio between (a close approximation of) p^* and the add-γ_i estimators p_t is bounded. Therefore, if we would have set γ_i sufficiently high, we could have used Lemma 6 and prior results for these other diverges to control $\mathrm{KL}(p^*\|p_{n+1})$. However, setting γ_i too high introduces too much bias to obtain minimax rates. Instead, our analysis involves carefully balancing control of the density ratio and the amount of bias that is introduced. In the remainder of this section we prove Theorems 2 and 3.

2.1. Proof of Theorem 2

By Lemma 5 we have that for all $i \in [K]$, with probability at least $1-\delta$

$$\frac{p^+(i)}{p_{n+1}(i)} \leq \left(1 + \frac{\sum_{i=1}^K \gamma_i}{n}\right)\left(\frac{3}{4} + \frac{9\ln(2K/\delta)}{\max\{n_i, \gamma_i\}}\right) =: \beta_i \,. \tag{13}$$

Let $\beta = \max_i \beta_i$. The inequality in (13) is very useful, as it allows us to control the ratio between $p^+(i)$ and $p_{n+1}(i)$. We use this to relate the KL-divergence to the squared Hellinger distance, which we know how to control from prior work. Specifically, by Lemma 4 of Yang and Barron (1998), on event (13),

$$\mathrm{KL}(p^*\|p_{n+1}) \leq (2 + \ln(\beta))H^2(p^*, p_{n+1})\,,$$

where H^2 is the squared Hellinger distance. By Lemma 19 we have $H^2(p^*, p_{n+1}) - H^2(p^*, \bar{p}_n) \leq \frac{\sum_{i=1}^K \gamma_i}{2n}$. We also have $H^2(p^*, \bar{p}_{n+1}) = H^2(\bar{p}_{n+1}, p^*) \leq \mathrm{KL}(\bar{p}_{n+1}\|p^*)$ (Gibbs and Su, 2002). Combining results from Agrawal (2022) and Paninski (2003) we obtain Lemma 23, which tells us that, with probability at least $1-\delta$, we have

$$\mathrm{KL}(\bar{p}_{n+1}\|p^*) \leq \mathbb{E}[\mathrm{KL}(\bar{p}_{n+1}\|p^*)] + \frac{6K + 6\ln(1/\delta)}{n} \leq \frac{7K + 6\ln(1/\delta)}{n}\,.$$

Thus, with probability at least $1-\delta$,

$$\mathrm{KL}(p^*\|p_{n+1}) \leq (2 + \ln(\beta))\Big(\frac{7K + 6\ln(1/\delta)}{n} + \frac{\sum_{i=1}^K \gamma_i}{2n}\Big)\,. \tag{14}$$

If $J \leq \ln(K/\delta)$, for $i \in \mathcal{J}$ we have $\gamma_i = \frac{\ln(K/\delta)}{J}$, for $i \notin \mathcal{J}$ we have $\gamma_i = 0$, and therefore $\frac{\sum_{i=1}^K \gamma_i}{n} = \frac{\ln(K/\delta)}{n} \leq \frac{K + \ln(1/\delta)}{n}$. As a consequence,

$$\begin{aligned}
\beta &= \max_i \left\{ \left(1 + \frac{\sum_{i=1}^K \gamma_i}{n}\right)\left(\frac{3}{4} + \frac{9\ln(2K/\delta)}{\max\{n_i, \gamma_i\}}\right) \right\} \\
&\leq \max_i \left\{ \left(\frac{6}{4} + \frac{18\ln(2K/\delta)}{\max\{n_i, \gamma_i\}}\right) \right\} \leq 40J = 40\min\{\ln(K/\delta), J\}\,,
\end{aligned}$$

where in the last inequality we used that $\max\{n_i, \gamma_i\} \geq \ln(K/\delta)J^{-1}$ for all i. If on the other hand $J > \ln(K/\delta)$, then for $i \in \mathcal{J}$ we have $\gamma_i = 1$, for $i \notin \mathcal{J}$ we have $\gamma_i = 0$, and therefore $\frac{\sum_{i=1}^K \gamma_i}{n} = \frac{J}{n} \leq \frac{K + \ln(1/\delta)}{n}$. Consequently,

$$\begin{aligned}
\beta &= \max_i \left\{ \left(1 + \frac{\sum_{i=1}^K \gamma_i}{n}\right)\left(\frac{3}{4} + \frac{9\ln(2K/\delta)}{\max\{n_i, \gamma_i\}}\right) \right\} \\
&\leq \left(\frac{6}{4} + \frac{18\ln(2K/\delta)}{\max\{n_i, \gamma_i\}}\right) \leq 40\ln(K/\delta) = 40\min\{\ln(K/\delta), J\}\,,
\end{aligned}$$

where in the last inequality is due to the fact that $\max\{n_i, \gamma_i\} \geq 1$ for all i. Thus, if $|\mathcal{J}| > 0$ then the bounds on β combined with (14) lead to the conclusion that with probability at least $1 - \delta$

$$\mathrm{KL}(p^* \| p_{n+1}) \leq (2 + \ln(40 \min\{\ln(K/\delta), J\}))\Big(\frac{9K + 9\ln(1/\delta)}{n}\Big) .$$

Finally, if $|\mathcal{J}| = 0$, then $\frac{\sum_{i=1}^K \gamma_i}{n} = 0$ and

$$\begin{aligned}\beta &= \max_i \left\{ \left(1 + \frac{\sum_{i=1}^K \gamma_i}{n}\right) \left(\frac{3}{4} + \frac{9\ln(2K/\delta)}{\max\{n_i, \gamma_i\}}\right) \right\} \\ &= \max_i \left\{\frac{3}{4} + \frac{9\ln(2K/\delta)}{n_i}\right\} \leq 2 ,\end{aligned}$$

where the last inequality is due to the fact that $n_i \geq 32\ln(K/\delta)$ for all i. Combined with (14) this leads to the conclusion that with probability at least $1 - \delta$

$$\mathrm{KL}(p^* \| p_{n+1}) \leq (2 + \ln(2))\Big(\frac{7K + 6\ln(1/\delta)}{n}\Big) ,$$

which concludes the proof.

2.2. Proof of Theorem 3

Let $\vartheta = \left(1 + \frac{\sum_{i=1}^K \gamma_i}{n}\right)$. We denote by $\mathcal{Z}$ the event that for all $i \in K$ and $t \in [n/2 + 1, \ldots, n + 1]$

$$\begin{aligned} &\frac{p^*(i)}{p_t(i)} \leq \vartheta \left(\frac{3}{4} + \frac{9\ln(2K/\delta)}{\max\{m_i, \gamma_i\}}\right) && \frac{p_t(i)}{p^+(i)} \leq 3\left(1 + \frac{2\,(3m_i + 27\ln(4K/\delta))}{\max\{\gamma_i, \frac{1}{2}m_i - 9\ln(4K/\delta)\}}\right) \\ &\frac{p^+(i)}{p_t(i)} \leq 1 + \vartheta \left(\frac{3}{4} + \frac{9\ln(4K/\delta)}{\max\{m_i, \gamma_i\}}\right) && \frac{n_i + \gamma_i}{m_i + \gamma_i} \leq 6 + \frac{18\ln(2K/\delta)}{m_i + \gamma_i} . \end{aligned} \tag{15}$$

By Lemma 5 we have that $\mathbb{P}(\mathcal{Z}) \geq 1 - \delta$ and by definition we have that $\frac{p^*(i)}{p^+(i)} \leq 1 + \frac{\sum_{i=1}^K \gamma_i}{n}$. To simplify notation let $\beta_i = \vartheta\left(\frac{3}{4} + \frac{9\ln(2K/\delta)}{\max\{n_i, \gamma_i\}}\right)$ and $\beta = \max_i \beta_i$.

At this point we can use Lemma 20 to relate $\mathrm{KL}(p^* \| p^{\mathrm{OTB}})$ to the regret of an online learning algorithm on the second half of the data, which tells us that on event $\mathcal{Z}$

$$\mathrm{KL}(p^* \| p^{\mathrm{OTB}}) \leq \frac{2}{n}\Big(2\mathcal{R}_T + 2\sum_{i=1}^K \gamma_i + (4 + 2\ln(\beta))\ln(1/\delta)\Big) ,$$

where $\mathcal{R}_T = \sum_{t=n/2+1}^n (-\ln p_t(X_t) - (-\ln p^+(X_t)))$. The result of Lemma 20 is very useful, as we can now rely on standard regret bounds to separate the bound into several parts which are easier to control. The result of Lemma 20 follows from first applying Freedman's inequality. However, a naïve application of Freedman's inequality would lead to a vacuous bound, as $|\ln(p^*(i)/p_t(i))|$ is unbounded. We instead rely on (15) to control said ratio before applying Freedman's inequality. Next, we still need to control the variance term of Freedman's inequality. For this we generalize the proof of Lemma 4 by Yang and Barron (1998) and show that the cumulative variance is bounded by $\mathcal{R}_T$, after which the result of Lemma 20 follows with some computations.

We continue by bounding the regret (Lemma 22):

$$\mathcal{R}_T \leq \sum_{i=1}^K \gamma_i \ln(p_{n/2+1}(i)/p^+(i)) + 2\sum_{i=1}^K \gamma_i + \frac{n}{2}\operatorname{KL}(\bar{p}_{n/2+1}\|p^*) + \sum_{i=1}^K \ln\left(\frac{n_i+\gamma_i}{m_i+\gamma_i}\right) .$$

Note that this is not a standard regret bound for prediction with log loss, as that would lead to a $O(\ln(n))$ term (see e.g. Chapter 9 of (Cesa-Bianchi and Lugosi, 2006)). Instead, we carefully use suffix averaging. When combined with the above, we can see that on event $\mathcal{Z}$

$$\operatorname{KL}(p^*\|p^{\text{OTB}}) \leq \frac{2}{n}\Big(2\sum_{i=1}^K \gamma_i \ln(p_{n/2+1}(i)/p^+(i)) + 6\sum_{i=1}^K \gamma_i + n\operatorname{KL}(\bar{p}_{n/2+1}\|p^*) + 2\sum_{i=1}^K \ln\left(\frac{n_i+\gamma_i}{m_i+\gamma_i}\right) + (4+2\ln(\beta))\ln(1/\delta)\Big) .$$

The term $\operatorname{KL}(\bar{p}_{n/2+1}\|p^*)$ can be controlled with standard results: Lemma 23 tells us that with probability at least $1-\delta$

$$\operatorname{KL}(\bar{p}_{n/2+1}\|p^*) \leq \frac{14K + 12\ln(1/\delta)}{n} .$$

The remaining challenge is to control

$$A = 2\sum_{i=1}^K \gamma_i \ln(p_{n/2+1}(i)/p^+(i)) + 6\sum_{i=1}^K \gamma_i + 2\sum_{i=1}^K \ln\left(\frac{n_i+\gamma_i}{m_i+\gamma_i}\right) + (4+2\ln(\beta))\ln(1/\delta) .$$

Here, the main challenge lies in controlling $2\sum_{i=1}^K \ln\left(\frac{n_i+\gamma_i}{m_i+\gamma_i}\right)$. We cannot naïvely rely on the fact that on event $\mathcal{Z}$ we have that $\frac{n_i+\gamma_i}{m_i+\gamma_i} \leq 6+18\frac{\ln(K/\delta)}{\gamma_i}$, as this could potentially lead to a $K\ln(\ln(1/\delta))$ term in the regret bound. Instead, we will split the analysis in two cases. In the first case $\frac{\ln(K/\delta)}{J} > 1$ and by definition of γ_i we can in fact use the naïve bound, which we do to prove Lemma 24:

$$A \leq 14\ln(K/\delta) + 40K + 4\ln(1/\delta)\ln\left(400J\right) .$$

In the second case $\frac{\ln(K/\delta)}{J} \leq 1$. In this case, we carefully control A in Lemma 25, which tells us that with probability at least $1-3\delta$

$$A \leq 100\ln(1/\delta)\ln(800J) + 4\widetilde{J}\ln\left(24\left(\ln\left(\frac{\widetilde{J}}{\ln(1/\delta)}\right)\vee 1\right)\right) + 20K\ln(100) .$$

All combined, we can see that with probability at least $1-5\delta$

$$\operatorname{KL}(p^*\|p^{\text{OTB}}) \leq \frac{1}{n}\Big(250\ln(1/\delta)\ln(800J) + 8\widetilde{J}\ln\left(24\left(\ln\left(\frac{\widetilde{J}}{\ln(1/\delta)}\right)\vee 1\right)\right) + 200K\Big),$$

where we coarsely bounded all the constants.

3. Lower Bounds

In this section we prove our main lower bound from (6):

Theorem 7 *Let $n > K^2$ and $n > \frac{4}{3}\ln(1/\delta)$ for any $\delta \in (0, \frac{1}{2})$. Then*

$$\inf_{\widehat{p}} \sup_{p^*} \ \mathbb{P}_{p^*}\left(\mathrm{KL}(p^* \| \widehat{p}) \geq C \frac{\max\{K, \ln(K)\ln(1/\delta)\}}{n}\right) > \delta,$$

where $C > 0$ is an absolute constant.

In Section 3.1 we first provide an estimator-dependent lower bound by direct calculations. Then we establish the minimax lower bound from Theorem 7 by breaking it up into two parts: in Section 3.2 we first prove a lower bound of order $\ln(K)\ln(1/\delta)/n$ by building on the intuition developed in the direct calculations. Because standard techniques based on a reduction to hypothesis testing are not precise enough to capture the $\ln(K)$ factor, this requires a novel reduction to what we call a *weak testing* problem (Section 3.2.1). For the second part of the proof of Theorem 7, we provide a lower bound of order K/n in Section 3.3, for which standard techniques based on Fano's inequality suffice after restricting the model to a ball around the uniform distribution. The missing proofs in this section can be found in Appendix D.

3.1. Estimator-dependent Lower Bound

Theorem 8 *Suppose that $n > \frac{4}{3}\ln(1/\delta)$. Denote by $\widehat{p}^0 \in \triangle^K$ the output of an estimator $\widehat{p}$ on the sample $X_1 = \cdots = X_n = 1$. Then, for any $\widehat{p}$, there exists $p^* \in \triangle^K$ such that*

$$\mathbb{P}_{p^*}\left(\mathrm{KL}(p^* \| \widehat{p}) \geq \frac{2\ln(1/\delta)}{3n}\left(\ln\left(\frac{2\ln(1/\delta)(K-1)}{3n\sum_{i=2}^K \widehat{p}^0(i)}\right) - 1\right) + \sum_{i=2}^K \widehat{p}^0(i)\right) > \delta\,.$$

We can see that any estimator that minimizes the lower bound satisfies $\sum_{i=2}^K \widehat{p}^0(i) \propto \frac{\ln(1/\delta)}{n}$, for which the lower bound becomes of order $\frac{\ln(K)\ln(1/\delta)}{n}$, and no estimator that does not depend on δ can match this rate. In particular, if $\widehat{p}$ is an add-γ estimator, so $\widehat{p}(i) = \frac{n_i+\gamma}{n+K\gamma}$, the lower bound specializes to

$$\frac{2\ln(1/\delta)}{3n}\left(\ln\left(\frac{2\ln(1/\delta)(n+\gamma K)}{3n\gamma}\right) - 1\right) + \frac{\gamma}{n+\gamma K},$$

so $\gamma \propto \ln(1/\delta)/K$ will achieve the $\frac{\ln(K)\ln(1/\delta)}{n}$ rate, but the KT ($\gamma = 1/2$) and Laplace ($\gamma = 1$) estimators lead to a rate of $\frac{\ln(\ln(1/\delta))\ln(1/\delta)}{n}$, which is worse in the regime where $\ln(K)\ln(1/\delta)$ dominates K.

3.2. Minimax Lower Bound, Part I: $\ln(K)\ln(1/\delta)/n$

Inspired by the proof of the estimator-dependent lower bound, we will lower bound the minimax rate by restricting the supremum in (10) to a finite set of probability mass functions $p_2, \ldots, p_K \in \triangle^K$ of the form

$$p_j(i) = \begin{cases} 1 - \frac{\alpha}{n} & \text{if } i = 1 \\ \frac{\alpha}{n} & \text{if } i = j \\ 0 & \text{otherwise,} \end{cases} \qquad \text{where we assume that } \alpha := \frac{2}{3}\ln(1/\delta) \leq \tfrac{1}{2}n\,. \tag{16}$$

3.2.1. INSUFFICIENCY OF THE STANDARD REDUCTION TO HYPOTHESIS TESTING

The standard approach to establishing lower bounds on the minimax rate goes via a general reduction to hypothesis testing (Tsybakov, 2009). It is commonly applied to metrics, but may easily be adjusted to KL divergence as shown in the following lemma. Working directly with KL divergence is potentially tighter than first lower-bounding KL by a metric. Since the lemma is not restricted to discrete distributions, we state it in terms of the KL divergence between general distributions, which is $\mathrm{KL}(P\|Q) = \int \ln\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)\mathrm{d}P$ if $P \ll Q$ and equals infinity otherwise.

Lemma 9 *Let $P_1, \ldots, P_M$ be distributions defined on a sample space $\mathcal{X}$, and let P_j^n be the distribution of n independent draws from P_j. Given any $\delta \in [0,1]$, suppose that*

$$\forall j \neq k: \quad \mathrm{KL}\left(P_j \| \frac{P_j + P_k}{2}\right) \geq s_n, \tag{17}$$

$$\inf_{\Psi} \max_{j} P_j^n(\Psi \neq j) > \delta, \tag{18}$$

where the infimum is over all possible hypothesis tests $\Psi : \mathcal{X}^n \to \{1, \ldots, M\}$. Then

$$\inf_{\widehat{P}} \max_{j} P_j^n\Big(\mathrm{KL}(P_j\|\widehat{P}) \geq s_n\Big) > \delta,$$

where the infimum is over all estimators based on n observations.

In our context the sample space is $\mathcal{X} = [K]$ and the conclusion of the lemma gives a lower bound on the minimax rate of $r_n^*(\delta) \geq s_n$. So what is the best lower bound we can hope for using Lemma 9 if we apply it to the distributions in (16)? Since

$$\mathrm{KL}\left(p_j \| \frac{p_j + p_k}{2}\right) = \frac{\alpha}{n} \ln \frac{\alpha/n}{\frac{1}{2}\alpha/n} = \frac{\alpha \ln(2)}{n} \qquad \text{for any } j \neq k,$$

condition (17) requires that $s_n \leq \frac{\alpha \ln(2)}{n} = \frac{2\ln(2)}{3}\frac{\ln(1/\delta)}{n}$, which falls short of the $\frac{\ln(K)\ln(1/\delta)}{n}$ rate that we are trying to show. Consequently, this standard reduction is insufficient to recover the right lower bound.

3.2.2. REDUCTION TO WEAK HYPOTHESIS TESTING

Ordinary hypothesis testing among distributions $P_1, \ldots, P_M$ is hard when there exist two distributions that cannot be properly distinguished. But the distributions corresponding to (16) are indistinguishable in a stronger sense: they all assign probability larger than δ to the same sequence $X_1 = \cdots = X_n = 1$ consisting of only ones, and as a consequence they are all indistinguishable simultaneously.

In order to capture this stronger sense of indistinguishability, we introduce the easier task of *weak hypothesis testing*, where the goal is not to identify the true distribution among $P_1, \ldots, P_M$ but to identify a set of $M - 1$ distributions that contains the true distribution. A weak hypothesis test is defined as a measurable function $\Psi : \mathcal{X}^n \to [M]$, which aims to identify a single distribution P_Ψ that is believed *not* to be the true distribution. In other words, the test Ψ is correct if the set $\{P_j | j \neq \Psi\}$ contains the true distribution, and makes a mistake if $\Psi = j$ when P_j is true.

Lemma 10 *Let $P_1, \ldots, P_M$ be distributions defined on a sample space $\mathcal{X}$, and let P_j^n be the distribution of n independent samples from P_j. Given any $\delta \in [0,1]$, suppose that*

$$\inf_{\Psi} \max_{j} P_j^n(\Psi = j) > \delta, \tag{19}$$

where the infimum is over all possible weak hypothesis tests $\Psi : \mathcal{X}^n \to \{1, \ldots, M\}$. Then

$$\inf_{\widehat{P}} \max_{j} P_j^n\Big(\mathrm{KL}(P_j \| \widehat{P}) \geq s_n\Big) > \delta \qquad \textit{for } \; s_n = \inf_{P} \max_{j} \mathrm{KL}(P_j \| P). \tag{20}$$

Condition (19) is analogous to (18) in that it lower bounds the worst-case probability that the (weak) hypothesis test fails. Applying this lemma to the distributions corresponding to (16), we obtain the desired lower bound:

Theorem 11 *For $K \geq 2$ and any $\delta \in (0,1]$, the minimax rate is at least*

$$r_n^*(\delta) \geq \frac{2\ln(K-1)\ln(1/\delta)}{3n} \qquad \textit{for all } n > \frac{4}{3}\ln(1/\delta).$$

3.3. Minimax Lower Bound, Part II: K/n

We now turn to proving a lower bound of order K/n, which is the standard parametric rate for a model with $K-1$ free parameters. Nevertheless, obtaining the right rate is not entirely straightforward, because it requires identifying a suitable subset $\triangle_0^K$ of $\triangle^K$ for which KL covering numbers and variational distance packing numbers can be shown to line up appropriately.

By Pinsker's inequality $\frac{1}{2}V(p,q)^2 \leq \mathrm{KL}(p\|q)$ for any p and q, so it is sufficient to prove a lower bound on the minimax rate for total variation. Although it is well known that such a lower bound can be obtained using standard techniques (see (Canonne, 2020) or (Polyanskiy and Wu, 2025, Exercise VI.8)), we spell out the details explicitly. In particular, since total variation is a metric, we can build on a result by Yang and Barron (1999), which is itself an elegant application of the standard reduction to hypothesis testing (Tsybakov, 2009) combined with Fano's inequality to lower bound the hypothesis testing error. Specialized to our setting, Yang and Barron's result gives the following:

For any $\triangle_0^K \subset \triangle^K$ and $\epsilon > 0$, let $N(\triangle_0^K, \epsilon, \mathrm{KL})$ be the Kullback-Leibler ϵ^2-entropy of $\triangle_0^K$[3], and let $M(\triangle_0^K, \epsilon, V)$ be the variational distance ϵ-packing entropy of $\triangle_0^K$[4].

Theorem 12 (Yang and Barron, 1999) *For any non-increasing, right-continuous bounds $N(\epsilon) \geq N(\triangle_0^K, \epsilon, \mathrm{KL})$ and $M(\epsilon) \leq M(\triangle_0^K, \epsilon, V)$, let $\epsilon_n, \underline{\epsilon}_n > 0$ be such that $\epsilon_n^2 = \frac{N(\epsilon_n)}{n}$ and $M(\underline{\epsilon}_n) \geq 4n\epsilon_n^2 + 2\ln 2$. Then*

$$\inf_{\widehat{p}} \sup_{p^* \in \triangle_0^K} \mathbb{P}_{p^*}\left(V(p^*, \widehat{p}) \geq \frac{\underline{\epsilon}_n}{2}\right) \geq \frac{1}{2},$$

where the infimum is over all estimators $\widehat{p}$ based on n observations.

3. That is, $\ln m$ for the smallest m for which there exist mass functions $p_1, \ldots, p_m \in \triangle^K$ such that for every $p \in \triangle_0^K$ there exists a j for which $\mathrm{KL}(p\|p_j) \leq \epsilon^2$.
4. That is, $\ln m$ for the largest m such that there exist $p_1, \ldots, p_m \in \triangle_0^K$ with pairwise distance $V(p_j, p_{j'}) > \epsilon$ for all $j \neq j'$.

For the full model $\triangle_0^K = \triangle^K$, Tang (2022) shows that $N(\triangle^K, \epsilon, \text{KL}) \leq \frac{K-1}{2} \ln \frac{800 \ln K}{\epsilon^2}$ for $\epsilon^2 \leq \ln K$, but this does not lead to the right parametric rate $\epsilon_n^2 \approx K/n$. The solution is to restrict the model to a ball of appropriate radius $\alpha \in \left(0, \frac{1}{2K}\right]$ around the uniform distribution with probabilities $u(i) = 1/K$:

$$\triangle_0^K = \{p \in \triangle^K : \|p - u\|_2 \leq \alpha\},$$

where $\|p - u\|_2 = \sqrt{\sum_{i=1}^K (p(i) - u(i))^2}$. This restriction ensures that, for all $p \in \triangle_0^K$,

$$\left(p(i) - \frac{1}{K}\right)^2 \leq \|p - u\|_2^2 \leq \alpha^2 \leq \left(\frac{1}{2K}\right)^2 \qquad \Longrightarrow \qquad p(i) \in \left[\frac{1}{2K}, \frac{3}{2K}\right],$$

so the ratios between any two $p, q \in \triangle_0^K$ are uniformly bounded. We will end up tuning $\alpha \approx \frac{1}{\sqrt{n}}$, so $\triangle_0^K$ is actually shrinking with n. We obtain the following bounds on the covering and packing entropies:

Lemma 13 *For any $\alpha \in \left(0, \frac{1}{2K}\right]$ and $\epsilon > 0$, the covering and packing entropies are bounded by*

$$N(\triangle_0^K, \epsilon, \text{KL}) \leq K \ln\left(\frac{\alpha 2\sqrt{2K}}{\epsilon} + 1\right), \qquad M(\triangle_0^K, \epsilon, V) \geq K \ln \frac{\alpha}{\epsilon} + \frac{K}{2} \ln \frac{K\pi}{8}.$$

Using these in Theorem 12, we obtain the following result:

Theorem 14 *There exists an absolute constant $C > 0$ such that, for any $K \geq 2$ and $\delta \in (0, 1/2)$, the minimax rate is at least*

$$r_n^*(\delta) \geq \frac{CK}{n} \qquad \textit{for all } n \geq \tfrac{\ln 2}{2} K^2.$$

The proof shows that $C = 4 \times 10^{-6}$ will work, but presumably this is very far from tight.

4. Conclusion

We presented nearly minimax rates for discrete distribution estimation in KL divergence with high probability. Our results represent an improvement in the understanding of arguably the simplest excess risk minimization problem with unbounded losses. An intriguing direction for future work is to extend the ideas we have presented here to more involved but related problems with unbounded losses such as prediction of Markov chains (see (Han et al., 2021)) and logistic regression. For that purpose, it is worthwhile to consider the approach of Mourtada (2025), whose estimator is simpler than the OTB estimator. Given that the approaches are technically different, both can be useful for generalizations to more involved settings. Finally, we did not carefully optimize our constants, so there is room for improvement there. Considering the work of Braess and Sauer (2004), there is substantial interest in understanding and obtaining the exact constant factors on the dominant term in the minimax rate.

Acknowledgments

We thank Nikita Zhivotovskiy for bringing (Mourtada, 2025) to our attention. Olkhovskaia and Van Erven were supported by the Netherlands Organization for Scientific Research (NWO) under grant numbers VI.Veni.232.068 and VI.Vidi.192.095, respectively.

References

I. Aden-Ali, Y. Cherapanamjeri, A. Shetty, and N. Zhivotovskiy. Optimal PAC bounds without uniform convergence. *arXiv preprint arXiv:2304.09167*, 2023.

R. Agrawal. Finite-sample concentration of the empirical relative entropy around its mean. *arXiv preprint arXiv:2203.00800*, 2022.

A. Beygelzimer, J. Langford, L. Li, L. Reyzin, and R. Schapire. Contextual bandit algorithms with supervised learning guarantees. In *International Conference on Artificial Intelligence and Statistics*, pages 19–26, 2011.

A. Bhattacharyya, S. Gayen, E. Price, and N. V. Vinodchandran. Near-optimal learning of tree-structured distributions by Chow-Liu. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 147–160, 2021.

D. Braess and T. Sauer. Bernstein polynomials and learning theory. *Journal of Approximation Theory*, 128(2):187–206, 2004.

D. Braess, J. Forster, T. Sauer, and H. U. Simon. How to achieve minimax expected Kullback-Leibler distance from an unknown finite distribution. In N. Cesa-Bianchi, M. Numao, and R. Reischuk, editors, *Algorithmic Learning Theory*, pages 380–394, 2002.

C. L. Canonne. A short note on learning discrete distributions. *arXiv preprint arXiv:2002.11457*, 2020.

C. L. Canonne, Z. Sun, and A. T. Suresh. Concentration bounds for discrete distribution estimation in KL divergence. *arXiv preprint arXiv:2302.06869*, 2023.

O. Catoni. The mixture approach to universal model selection. In *Preprints of École Normale Supérieure*, 1997.

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

T. Cover. Admissibility properties or Gilbert's encoding for unknown source probabilities (corresp.). *IEEE Transactions on Information Theory*, 18(1):216–217, 1972. doi: 10.1109/TIT.1972.1054738.

L. Devroye and G. Lugosi. *Combinatorial methods in density estimation*. Springer Science & Business Media, 2001.

I. Diakonikolas. Learning structured distributions. *Handbook of Big Data*, 267:10–1201, 2016.

J. Forster and M. K. Warmuth. Relative expected instantaneous loss bounds. *Journal of Computer and System Sciences*, 64(1):76–102, 2002.

A. L. Gibbs and F. E. Su. On choosing and bounding probability metrics. *International Statistical Review*, 70(3):419–435, 2002.

Y. Han, S. Jana, and Y. Wu. Optimal prediction of Markov chains with and without spectral gap. *Advances in Neural Information Processing Systems*, 34:11233–11246, 2021.

N. J. A. Harvey, C. Liaw, Y. Plan, and S. Randhawa. Tight analyses for non-smooth stochastic gradient descent. In *Conference on Learning Theory*, pages 1579–1613, 2019.

D. Van der Hoeven, N. Zhivotovskiy, and N. Cesa-Bianchi. High-probability risk bounds via sequential predictors. *arXiv preprint arXiv:2308.07588*, 2023.

S. Kamath, A. Orlitsky, D. Pichapati, and A. T. Suresh. On learning distributions from their samples. In P. Grünwald, E. Hazan, and S. Kale, editors, *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 1066–1100. PMLR, 2015. URL https://proceedings.mlr.press/v40/Kamath15.html.

R. Krichevsky and V. Trofimov. The performance of universal encoding. *IEEE Transactions on Information Theory*, 27(2):199–207, 1981.

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338, 2000. doi: 10.1214/aos/1015957395.

N. Littlestone. From on-line to batch learning. In *Proceedings of the Second Annual Workshop on Computational Learning Theory*, COLT '89, page 269–284, 1989.

A. Maurer and M. Pontil. Empirical Bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009.

N. Merhav and M. Feder. Universal prediction. *IEEE Transactions on Information Theory*, 44(6): 2124–2147, 1998.

J. Mourtada. Estimation of discrete distributions in relative entropy, and the deviations of the missing mass. *arXiv preprint arXiv:2504.21787*, 2025.

J. Mourtada and S. Gaïffas. An improper estimator with optimal excess risk in misspecified density estimation and logistic regression. *Journal of Machine Learning Research*, 23:1–49, 2022.

A. Nemirovski. Lecture notes: Interior-point polynomial time methods for convex programming. 1996.

F. Orabona. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019.

L. Paninski. Estimation of entropy and mutual information. *Neural computation*, 15(6):1191–1253, 2003.

L. Paninski. Variational minimax estimation of discrete distributions under kl loss. In L. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 17. MIT Press, 2004. URL https://proceedings.neurips.cc/paper_files/paper/2004/file/c57168a952f5d46724cf35dfc3d48a7f-Paper.pdf.

Y. Polyanskiy and Y. Wu. *Information Theory: From Coding to Learning*. Cambridge University Press, 2025.

A. Rakhlin, O. Shamir, and K. Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the 29th International Conference on Machine Learning*, page 1571–1578, 2012.

J. J. Rissanen. Fisher information and stochastic complexity. *IEEE Transactions on Information Theory*, 42(1):40–47, 1996.

G. G. Roussas. *An introduction to probability and statistical inference*. Elsevier, 2003.

Y. M. Shtarkov. Universal sequential coding of single messages. *Problemy Peredachi Informatsii*, 23(3):3–17, 1987.

J. Tang. *Divergence Covering*. PhD thesis, Massachusetts Institute of Technology, 2022.

A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

D. van der Hoeven, J. Olkhovskaia, and T. van Erven. Nearly minimax discrete distribution estimation in Kullback-Leibler divergence with high probability. *arXiv preprint arXiv:2507.17316v1*, July 2025.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

Q. Xie and A. Barron. Minimax redundancy for the class of memoryless sources. *IEEE Transactions on Information Theory*, 43(2):646–657, 1997.

Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics*, 27(5):1564–1599, 1999. URL http://www.jstor.org/stable/2674082.

Y. Yang and A. R. Barron. An asymptotic property of model selection criteria. *IEEE Transactions on Information Theory*, 44(1):95–116, 1998.

Appendix A. Asymptotic Behavior of the Maximum Likelihood Estimator

We formalize here the claim from the introduction that $n\,\mathrm{KL}(p^*\|\bar{p}_n)$ converges to a χ^2 distribution as $n \to \infty$ with p^* held fixed.

Theorem 15 *Let $\bar{p}_n$ denote the maximum likelihood estimator, and let p^* be arbitrary with support of size $M = |\{i \mid p^*(i) > 0\}|$. The case $M = 1$ is trivial, because then $\mathrm{KL}(p^*\|\bar{p}_n) = 0$ almost surely, so assume $M \geq 2$. Then*

$$2n\,\mathrm{KL}(p^*\|\bar{p}_n) \;\rightsquigarrow\; \chi^2_{M-1}. \tag{21}$$

It follows that

$$\mathrm{KL}(p^*\|\bar{p}_n) \leq \frac{M - 1 + \frac{3}{2}\ln(2/\delta)}{n} \qquad \text{with } p^*\text{-probability at least } 1-\delta \tag{22}$$

for all sufficiently large n.

Proof Our approach will be to apply the second-order delta method (van der Vaart, 1998, Section 3.3), which combines the fact that the MLE is asymptotically Gaussian (by the central limit theorem) with a second-order Taylor expansion of the KL divergence in its second argument to get the desired conclusion.

We assume without loss of generality that $K = M$ (i.e. p^* has full support), because the distribution of $\mathrm{KL}(p^*\|\bar{p}_n)$ does not change if we remove all outcomes i for which $p^*(i) = 0$. Then, because of the simplex constraint, the MLE will be asymptotically Gaussian on a subspace of dimension $K-1$. It is therefore technically convenient to let $\mu, \bar{\mu} \in \mathbb{R}^{K-1}$ denote the vectors with the first $K-1$ coordinates of p^* and $\bar{p}_n$, respectively: $\mu_i = p^*(i)$ and $\bar{\mu}_i = n_i/n$ for $i = 1, \ldots, K-1$. Then we may view $\bar{\mu}$ as

$$\bar{\mu} = \frac{1}{n}\sum_{t=1}^{n} e_{X_t},$$

where e_{X_t} is the standard basis vector in direction X_t if $X_t \neq K$ and the all-zeros vector otherwise. Therefore, by the central limit theorem, the difference between $\bar{\mu}$ and μ converges in distribution to a normal distribution:

$$\sqrt{n}(\bar{\mu} - \mu) \rightsquigarrow \Sigma^{1/2} Z, \tag{23}$$

where $Z \sim \mathcal{N}(0, I)$ is standard normal in $K-1$ dimensions and $\Sigma \in \mathbb{R}^{(K-1)\times(K-1)}$ is the covariance matrix of e_X:

$$\Sigma_{ij} = \mathbb{E}[(\mathbb{1}[X = i] - \mu_i)(\mathbb{1}[X = j] - \mu_j)] = \begin{cases} \mu_i(1-\mu_i) & \text{if } i = j, \\ -\mu_i\mu_j & \text{otherwise.} \end{cases}$$

By the continuous mapping theorem, it follows that

$$n\|\bar{\mu} - \mu\|^2 \rightsquigarrow Z^\top \Sigma Z. \tag{24}$$

Given the one-to-one relation between $\bar{p}_n$ and $\bar{\mu}$, we may consider $\mathrm{KL}(p^*\|\bar{p}_n)$ as a function $\bar{\mu}$:

$$\mathrm{KL}(p^*\|\bar{p}_n) = \phi(\bar{\mu}) := \sum_{i=1}^{K-1} \mu_i \ln\frac{\mu_i}{\bar{\mu}_i} + f_K(\mu)\ln\frac{f_K(\mu)}{f_K(\bar{\mu})},$$

where we have abbreviated $f_K(\mu) = 1 - \sum_{i=1}^{K-1} \mu_i$. Since $\mu_i > 0$ for all i by assumption, ϕ is twice differentiable for all $\bar{\mu}$ close enough to μ, with

$$\nabla\phi(\bar{\mu})_i = \frac{f_K(\mu)}{f_K(\bar{\mu})} - \frac{\mu_i}{\bar{\mu}_i}, \qquad \nabla^2\phi(\bar{\mu})_{ij} = \begin{cases} \frac{f_K(\mu)}{f_K(\bar{\mu})^2} + \frac{\mu_i}{(\bar{\mu}_i)^2} & \text{if } i = j, \\ \frac{f_K(\mu)}{f_K(\bar{\mu})^2} & \text{otherwise.} \end{cases}$$

Let $H = \nabla^2\phi(\mu)$ denote the Hessian of ϕ at $\bar{\mu} = \mu$, i.e. the Fisher information. Then, because $\phi(\mu) = 0$ and $\nabla\phi(\mu) = 0$, a second-order Taylor expansion of ϕ around $\bar{\mu} = \mu$ gives

$$\begin{aligned} \phi(\bar{\mu}) &= \phi(\mu) + (\bar{\mu} - \mu)\nabla\phi(\mu) + \tfrac{1}{2}(\bar{\mu} - \mu)^\top H(\bar{\mu} - \mu) + o_P\big(\|\bar{\mu} - \mu\|^2\big) \\ &= \tfrac{1}{2}(\bar{\mu} - \mu)^\top H(\bar{\mu} - \mu) + o_P\big(\|\bar{\mu} - \mu\|^2\big) \\ 2n\,\mathrm{KL}(p^*\|\bar{p}_n) = 2n\phi(\bar{\mu}) &= n(\bar{\mu} - \mu)^\top H(\bar{\mu} - \mu) + o_P\big(n\|\bar{\mu} - \mu\|^2\big) \rightsquigarrow Z^\top\Sigma^{1/2}H\Sigma^{1/2}Z, \end{aligned}$$

where convergence in the last step holds because both terms converge: first, (23) and the continuous mapping theorem imply that

$$n(\bar{\mu} - \mu)^\top H(\bar{\mu} - \mu) \rightsquigarrow Z^\top\Sigma^{1/2}H\Sigma^{1/2}Z.$$

And, second, since $n\|\bar{\mu} - \mu\|^2$ converges in distribution by (24), it is uniformly tight by Prohorov's theorem (van der Vaart, 1998), which implies that $o_P\big(n\|\bar{\mu} - \mu\|^2\big) = o_P(1)$. The convergence step above therefore follows from Slutsky's theorem, which allows us to combine the convergence of the two terms.

It remains to simplify $\Sigma^{1/2}H\Sigma^{1/2}$, using that the Fisher information matrix H is the inverse of the covariance matrix Σ. Specifically, it can be confirmed that H is the left inverse of Σ by direct calculation:

$$\begin{aligned} (H\Sigma)_{ij} &= \sum_{k=1}^{K-1} \Big(\frac{1}{f_K(\mu)} + \mathbb{1}[k = i]\frac{1}{\mu_i}\Big)\big(\mathbb{1}[k = j]\mu_j - \mu_k\mu_j\big) \\ &= \sum_{k=1}^{K-1} \Big(\mathbb{1}[k = j]\frac{\mu_j}{f_K(\mu)} + \mathbb{1}[k = i = j]\frac{\mu_j}{\mu_i} - \frac{\mu_k\mu_j}{f_K(\mu)} - \mathbb{1}[k = i]\frac{\mu_k\mu_j}{\mu_i}\Big) \\ &= \frac{\mu_j}{f_K(\mu)} + \mathbb{1}[i = j] - \Big(\frac{\mu_j}{f_K(\mu)} - \mu_j\Big) - \mu_j = \mathbb{1}[i = j], \end{aligned}$$

and, since both H and Σ are symmetric, it follows that $H = \Sigma^{-1}$. Being the inverse of a symmetric, positive definite matrix, Σ^{-1} is also positive definite, and therefore

$$\Sigma^{1/2}H\Sigma^{1/2} = \Sigma^{1/2}\Sigma^{-1}\Sigma^{1/2} = \Sigma^{1/2}\Sigma^{-1/2}\Sigma^{-1/2}\Sigma^{1/2} = I.$$

Putting everything together, we have shown that

$$2n\,\mathrm{KL}(p^*\|\bar{p}_n) \rightsquigarrow Z^\top Z,$$

where $Z^\top Z$ has a χ^2_{K-1} distribution. Since we have reduced to the case that $K = M$, this completes the proof of (21).

To obtain (22), it suffices to use concentration of the χ^2 distribution around its mean. In particular, if $Y \sim \chi^2_{M-1}$, then Laurent and Massart (2000) show that

$$\mathbb{P}\Big(Y \geq M - 1 + 2\sqrt{(M-1)\ln(1/\delta)} + 2\ln(1/\delta)\Big) \leq \delta$$
$$\mathbb{P}\Big(Y \geq 2(M-1) + 3\ln(1/\delta)\Big) \leq \delta,$$

where we have used that $\sqrt{ab} \leq \frac{1}{2}a + \frac{1}{2}b$ for $a, b \geq 0$. Since, for all sufficiently large n,

$$\Big|\mathbb{P}\Big(2n\,\mathrm{KL}(p^*\|\bar{p}_n) \geq 2(M-1) + 3\ln(2/\delta)\Big) - \mathbb{P}\Big(Y \geq 2(M-1) + 3\ln(2/\delta)\Big)\Big| \leq \frac{\delta}{2},$$

we conclude that, for all such n,

$$\mathbb{P}\Big(2n\,\mathrm{KL}(p^*\|\bar{p}_n) \geq 2(M-1) + 3\ln(2/\delta)\Big) \leq \mathbb{P}\Big(Y \geq 2(M-1) + 3\ln(2/\delta)\Big) + \frac{\delta}{2} \leq \frac{\delta}{2} + \frac{\delta}{2} = \delta,$$

from which (22) follows. ∎

Appendix B. Auxilliary Results

Lemma 16 *For any distributions p, q, r such that $\frac{r(i)}{q(i)} \leq V$ for all i*

$$\sum_{i=1}^K p(i)\Big(\ln\frac{r(i)}{q(i)}\Big)^2 \leq (2 + \ln V)\sum_{i=1}^K p(i)\ln\frac{r(i)}{q(i)} + \sum_{i=1}^K p(i)\frac{q(i)}{r(i)} - 1.$$

Proof We generalize the proof of Lemma 4 by Yang & Barron, 1998: let $\phi(V) = \frac{\ln V + 1/V - 1}{(\ln V)^2} \geq \frac{1}{2+\ln V}$ be their decreasing function ϕ_1. Then

$$\begin{aligned}
\phi(V)\sum_{i=1}^K p(i)\Big(\ln\frac{r(i)}{q(i)}\Big)^2 &\leq \sum_{i=1}^K p(i)\phi(\frac{r(i)}{q(i)})\Big(\ln\frac{r(i)}{q(i)}\Big)^2 \\
&= \sum_{i=1}^K p\Big(\ln\frac{r(i)}{q(i)} + \frac{q(i)}{r(i)} - 1\Big) \\
&= \sum_{i=1}^K p(i)\ln\frac{r(i)}{q(i)} + \sum_{i=1}^K p(i)\frac{q(i)}{r(i)} - 1.
\end{aligned}$$

∎

Applying this with $p = p^*$ and $r = p^+$ we obtain:

Corollary 17 *For any q such that $\ln\frac{p^+(i)}{q(i)} \leq \alpha$ for all i,*

$$\sum_{i=1}^K p^*(i)\ln\Big(\frac{p^+(i)}{q(i)}\Big)^2 \leq (2+\alpha)\Big(\sum_{i=1}^K p^*(i)\ln\Big(\frac{p^+(i)}{q(i)}\Big) + \frac{\sum_{i=1}^K \gamma_i}{n}\Big).$$

Proof By Lemma 16

$$\begin{aligned}
\sum_{i=1}^{K} p^*(i) \ln\left(\frac{p^+(i)}{q(i)}\right)^2 &\leq (2+\alpha)\Big(\sum_{i=1}^{K} p^*(i) \ln\left(\frac{p^+(i)}{q(i)}\right) + \sum_{i=1}^{K} p^*(i) \frac{q(i)}{p^+(i)} - 1\Big) \\
&= (2+\alpha)\Big(\sum_{i=1}^{K} p^*(i) \ln\left(\frac{p^+(i)}{q(i)}\right) + \sum_{i=1}^{K} q(i) \frac{p^*(i)}{p^+(i)} - 1\Big) \\
&\leq (2+\alpha)\Big(\sum_{i=1}^{K} p^*(i) \ln\left(\frac{p^+(i)}{q(i)}\right) + \frac{n + \sum_{i=1}^{K} \gamma_i}{n} - 1\Big) \\
&= (2+\alpha)\Big(\sum_{i=1}^{K} p^*(i) \ln\left(\frac{p^+(i)}{q(i)}\right) + \frac{\sum_{i=1}^{K} \gamma_i}{n}\Big).
\end{aligned}$$

■

Lemma 18 *If $N_1, \ldots, N_K \sim M(n, p^*(1), \ldots, p^*(K))$ and $\widetilde{N}_i \sim B(n, \min\{1, 3p^*(i)\})$, then for any $R \subset [K]$ such that $\sum_{i \notin R} p^*(i) \geq 1/3$, any $b_1, \ldots, b_K \geq 0$, and any $z > 0$*

$$\mathbb{P}(\sum_{i \in R} \mathbb{1}\{N_i \geq b_i\} \geq z) \leq \mathbb{P}(\sum_{i \in R} \mathbb{1}\{\widetilde{N}_i \geq b_i\} \geq z).$$

Proof For simplicity we only show the proof for $R = [K-1]$ and assume $p(K) \geq \frac{1}{3}$. The general proof follows from the proof for this case. For any i' and $R' = [K-1] \setminus i'$ a basic property of the multinomial distribution is that (see, e.g. (Roussas, 2003, Chapter 4))

$$N_{i'} | \{N_i : i \in R'\} \sim B(n - \sum_{i \in R'} N_i, \frac{p(i')}{\sum_{i \notin R'} p(j)}).$$

Thus, since $\sum_{i \notin R'} p(j) > p(K) \geq \frac{1}{3}$, we have that for any $b \geq 0$

$$\mathbb{P}(N_1 \geq b | N_2, \ldots, N_{K-1}) \leq \mathbb{P}(\widetilde{N}_1 \geq b). \tag{25}$$

Denote by $Y_i = \mathbb{1}\{N_i \geq b_i\}$, by $\widetilde{Y} = \mathbb{1}\{\widetilde{N_i} \geq b_i\}$, and by

$$F = \{x_2, \ldots, x_{K-1} : \sum_{i=2}^{K-1} \mathbb{1}\{x_i > b_i\} = z - 1\}.$$

We have

$$
\begin{aligned}
&\mathbb{P}(\sum_{i=1}^{K-1} Y_i \geq z) \\
&= \mathbb{P}(Y_1 \geq z - \sum_{i=2}^{K-1} Y_i) \\
&= \sum_{j=1}^{K-2} \mathbb{P}(\sum_{i=2}^{K-1} Y_i = j)\, \mathbb{P}(Y_1 \geq z - j | \sum_{i=2}^{K-1} Y_i = j) \\
&\stackrel{a}{=} \mathbb{P}(\sum_{i=2}^{K-1} Y_i > z-1) + \mathbb{P}(\sum_{i=2}^{K-1} Y_i = z-1)\, \mathbb{P}(Y_1 = 1 | \sum_{i=2}^{K-1} Y_i = z-1) \\
&= \mathbb{P}(\sum_{i=2}^{K-1} Y_i > z-1) + \mathbb{P}(Y_1 = 1 \bigcap \left(\sum_{i=2}^{K-1} Y_i = z-1\right)) \\
&= \mathbb{P}(\sum_{i=2}^{K-1} Y_i > z-1) + \sum_{(x_2,\ldots,x_{K-1}) \in F} \mathbb{P}(Y_1 = 1 \bigcap (N_2 = x_2, \ldots, N_{K-1} = x_{K-1})) \\
&= \mathbb{P}(\sum_{i=2}^{K} Y_i > z-1) + \sum_{(x_2,\ldots,x_{K-1}) \in F} \mathbb{P}(N_1 \geq b_1 \bigcap (N_2 = x_2, \ldots, N_{K-1} = x_{K-1})) \\
&\leq \mathbb{P}(\sum_{i=2}^{K} Y_i > z-1) + \sum_{(x_2,\ldots,x_{K-1}) \in F} \mathbb{P}(\widetilde{N}_1 \geq b_1 \bigcap (N_2 = x_2, \ldots, N_{K-1} = x_{K-1})) \\
&= \mathbb{P}(\widetilde{Y}_1 + \sum_{i=2}^{K-1} Y_i \geq z)\,,
\end{aligned}
$$

where a follows from the fact that Y_1 can only take values 0 or 1 and the inequality is due to (25). Repeating the above argument shows that $\mathbb{P}(\sum_{i=1}^{K-1} Y_i \geq z) \leq \mathbb{P}(\sum_{i=1}^{K-1} \widetilde{Y}_i \geq z)$. ■

Appendix C. Additional Proofs for Section 2

Lemma 5 *Suppose that* $n \geq 4$. *Let* $\zeta_i = 3\left(1 + \frac{2(3m_i + 27\ln(4K/\delta))}{\max\{\gamma_i, \frac{1}{2}m_i - 9\ln(4K/\delta)\}}\right)$ *and* $\xi = 1 + \frac{\sum_{i=1}^K \gamma_i}{n}$. *For all* $t \in [n/2+1, \ldots, n+1]$, $i \in [K]$, *and* $\delta \in (0,1)$, *with probability at least* $1-\delta$,

$$
\frac{p^*(i)}{p_t(i)} \leq \zeta_i \xi\,, \qquad \frac{p_t(i)}{p^+(i)} \leq \zeta_i\,, \qquad \frac{p^+(i)}{p_t(i)} \leq 1 + \zeta_i \xi\,, \qquad \frac{p^*(i)}{p^+(i)} \leq \xi\,, \qquad \frac{n_i + \gamma_i}{m_i + \gamma_i} \leq 6 + \zeta_i\,.
$$

Proof We start with

$$
\frac{p^*(i)}{p^+(i)} = \frac{n + \sum_{i=1}^K \gamma_i}{n} \frac{np^*(i)}{np^*(i) + \gamma_i} \leq 1 + \frac{\sum_{i=1}^K \gamma_i}{n}\,.
$$

By the empirical Bernstein inequality (Maurer and Pontil, 2009, Theorem 4) and a union bound we have that for all $i \in [K]$, with probability at least $1 - 2K\delta'$

$$|n_i - p^*(i)n| \leq \sqrt{2n_i \ln(2/\delta)} + 5\ln(2/\delta')$$
$$|m_i - p^*(i)n| \leq \sqrt{2m_i \ln(2/\delta)} + 5\ln(2/\delta')\,. \tag{26}$$

On this event, by the AM-GM inequality, for any fixed $\eta > 0$, we have that

$$(1-\eta)n_i - \left(\frac{2}{\eta} + 5\right)\ln(2/\delta') \leq 2np^*(i) \leq (1+\eta)n_i + \left(\frac{2}{\eta} + 5\right)\ln(2/\delta')$$
$$(1-\eta)m_i - \left(\frac{2}{\eta} + 5\right)\ln(2/\delta') \leq 2np^*(i) \leq (1+\eta)m_i + \left(\frac{2}{\eta} + 5\right)\ln(2/\delta')\,. \tag{27}$$

On the event that (26) holds, we thus have that

$$\begin{aligned}\frac{p^*(i)}{p_t(i)} &\leq \frac{n + \sum_{i=1}^K \gamma_i}{n} \frac{np^*(i)}{m_i + \gamma_i} \\ &\leq \frac{n + \sum_{i=1}^K \gamma_i}{n} \frac{\frac{1}{2}(1+\eta)m_i + \left(\frac{2}{\eta} + 5\right)\ln(2/\delta')}{m_i + \gamma_i} \\ &\leq \left(1 + \frac{\sum_{i=1}^K \gamma_i}{n}\right)\left(\frac{1+\eta}{2} + \frac{\left(\frac{2}{\eta} + 5\right)\ln(2/\delta')}{\max\{m_i, \gamma_i\}}\right)\,.\end{aligned}$$

Likewise, on the event that (26) holds, we have that

$$\begin{aligned}\frac{p^+(i)}{p_t(i)} &\leq \frac{np^*(i) + \gamma_i}{m_i + \gamma_i} \\ &\leq 1 + \frac{np^*(i)}{m_i + \gamma_i} \\ &\leq 1 + \left(1 + \frac{\sum_{i=1}^K \gamma_i}{n}\right)\left(\frac{1+\eta}{2} + \frac{\left(\frac{2}{\eta} + 5\right)\ln(2/\delta')}{\max\{m_i, \gamma_i\}}\right)\,.\end{aligned}$$

And once more, on the event that (26) holds, we have that

$$\begin{aligned}\frac{p_t(i)}{p^+(i)} &= \frac{n + \sum_{i=1}^K \gamma_i}{t - 1 + \sum_{i=1}^K \gamma_i} \frac{n_i(t-1) + \gamma_i}{np^*(i) + \gamma_i} \\ &\leq 3\frac{n_i + \gamma_i}{np^*(i) + \gamma_i} \\ &\leq 3\left(1 + \frac{n_i}{\max\{\gamma_i, (1-\eta)m_i - \left(\frac{2}{\eta} + 5\right)\ln(2/\delta')\}}\right)\end{aligned}$$

Furthermore, we have

$$\begin{aligned}n_i &\leq \frac{1}{1-\eta}\left(2np^*(i) + \left(\frac{2}{\eta} + 5\right)\ln(2/\delta')\right) \\ &\leq \frac{1}{1-\eta}\left(2(1+\eta)m_i + 3\left(\frac{2}{\eta} + 5\right)\ln(2/\delta')\right)\,,\end{aligned} \tag{28}$$

and thus

$$\frac{p_t(i)}{p^+(i)} \leq 3\left(1 + \frac{\frac{1}{1-\eta}\left(2(1+\eta)m_i + 3\left(\frac{2}{\eta}+5\right)\ln(2/\delta')\right)}{\max\{\gamma_i, (1-\eta)m_i - \left(\frac{2}{\eta}+5\right)\ln(2/\delta')\}}\right).$$

Finally, by equation 28 we also have

$$\frac{n_i+\gamma_i}{m_i+\gamma_i} \leq \frac{2(1+\eta)}{1-\eta} + \frac{\left(\frac{2}{\eta}+5\right)\ln(2/\delta')}{(1-\eta)(m_i+\gamma_i)}$$

Setting $\delta' = \frac{\delta}{2K}$ and $\eta = \frac{1}{2}$ completes the proof. ■

Lemma 6 *For any pair $p, q \in \triangle^K$, suppose that $\frac{1}{2} \leq \frac{p(i)}{q(i)} \leq 2$ for all $i \in [K]$. Then*

$$\tfrac{1}{6}\chi^2(p,q) \leq \tfrac{1}{4}\chi^2(q,p) \leq \mathrm{KL}(p\|q) \leq \tfrac{5}{2}H^2(p,q) \leq \tfrac{5}{2}\mathrm{KL}(q\|p) \leq \tfrac{5}{2}\chi^2(q,p) \leq 5\chi^2(p,q).$$

Proof Let R be a self-concordant function, let $r = \sqrt{(x-y)(R''(x))(x-y)}$, and let $\rho(u) = -\ln(1-u) - u$. By (Nemirovski, 1996, 2.4),

$$R(y) \geq R(x) + \langle \nabla R(x), y-x\rangle + \rho(-r) \geq R(x) + \langle \nabla R(x), y-x\rangle + \frac{\min(r, r^2)}{4}.$$

Since $-\ln(x)$ is self-concordant, we have

$$p^*(i)(\ln(p(i)) - \ln(q(i))) \geq p^*(i)\Big(\frac{-1}{p(i)}(q(i)-p(i)) + \frac{1}{4}\min\{r, r^2\}\Big),$$

where

$$r^2 = \frac{(p^*(i) - \bar{p}_n(i))^2}{p^*(i)^2} \leq \frac{1}{p^*(i)^2}\left(\frac{7}{32}\bar{p}_n(i)\right)^2 \leq \frac{1}{p^*(i)^2}\left(\frac{14}{32}p^*(i)\right)^2 \leq \frac{1}{4}.$$

So, we have

$$\begin{aligned}
\mathrm{KL}(p\|q) &= \sum_{i=1}^K p(i)(\ln(p(i)) - \ln(q(i))) \\
&\geq \sum_{i=1}^K p(i)\Big(\frac{-1}{p(i)}(q(i)-p(i)) + \frac{1}{4}\min\{r, r^2\}\Big) \\
&= \sum_{i=1}^K \Big((q(i)-p(i)) + \frac{1}{4}\frac{(p(i)-q(i))^2}{p(i)}\Big) \\
&= \frac{1}{4}\sum_{i=1}^K \frac{(p(i)-q(i))^2}{p(i)} \\
&= \frac{1}{4}\chi^2(q,p) \\
&\geq \frac{1}{6}\sum_{i=1}^K \frac{(p(i)-q(i))^2}{q(i)} \\
&= \frac{1}{6}\chi^2(p,q).
\end{aligned}$$

By Lemma 4 in (Yang and Barron, 1998) we have that

$$\begin{aligned}\mathrm{KL}(p\|q) &\leq (2+\ln(3/2))H^2(p,q) \leq (2+\ln(3/2))\,\mathrm{KL}(q\|p)\\ &\leq (2+\ln(3/2))\chi^2(q,p) \leq (4+2\ln(3/2))\chi^2(p,q)\,,\end{aligned}$$

where $H^2(p,q) \leq \mathrm{KL}(q\|p) \leq \chi^2(q,p)$ can be found in (Gibbs and Su, 2002). Naïvely bounding $\ln(3/2) < 1/2$ and simplifying completes the proof. ■

Theorem 1 *For some $C > 0$, if $|\widetilde{\mathcal{J}}| = 0$, then, with probability at least $1-\delta$, the maximum likelihood estimator guarantees*

$$\tfrac{1}{6}\chi^2(p^*,\bar{p}_n) \leq \tfrac{1}{4}\chi^2(\bar{p}_n,p^*) \leq \mathrm{KL}(p^*\|\bar{p}_n) \leq C\frac{K+\ln(1/\delta)}{n}\,.$$

Proof By Bennet's inequality (Maurer and Pontil, 2009, Theorem 4) and a union bound we have that for all $i \in [K]$, with probability at least $1-\delta$

$$|n_i - p^*(i)n| \leq \sqrt{2np^*(i)\ln(4K/\delta)} + 5\ln(4K/\delta) \leq \frac{7}{32}np^*(i)\,,$$

where the last inequality follows from the assumption that $|\widetilde{\mathcal{J}}| = 0$. Thus, on this event, for all $i \in [K]$

$$\bar{p}_n(i) \in \left[\frac{25}{32}p^*(i), \frac{39}{32}p^*(i))\right] \subset \left[\frac{1}{2}p^*(i), \frac{3}{2}p^*(i)\right]\,.$$

At this point, an application of Lemma 6 allows us to complete the proof of the second part of the statement. For $\mathbb{P}\left(\mathrm{KL}(p^*\|\bar{p}_n) \geq C\frac{K+\ln(1/\delta)}{n}\right) \leq \delta$ we can apply Lemma 6 to obtain $\mathrm{KL}(p^*\|\bar{p}_n) \leq \mathrm{KL}(\bar{p}_n\|p^*)$, after which we can apply Lemma 23 to complete the proof.

■

Lemma 19 *We have that*

$$H^2(p^*,p_{n+1}) - H^2(p^*,\bar{p}_n) \leq \frac{\sum_{i=1}^K \gamma_i}{2n}.$$

Proof By definition of the Hellinger distance we have that

$$
\begin{aligned}
H^2(p^*, p_{n+1}) - H^2(p^*, \bar{p}_n) &= \frac{1}{2}\sum_{i=1}^{K} \sqrt{p^*(i)}\big(\sqrt{\bar{p}_n(i)} - \sqrt{p_{n+1}(i)}\big) \\
&= \frac{1}{2}\sum_{i=1}^{K} \sqrt{p^*(i)}\Big(\sqrt{\bar{p}_n(i)} - \sqrt{\bar{p}_n(i)\frac{n}{n+\sum_{i=1}^K \gamma_i} + \frac{\gamma_i}{n+\sum_{i=1}^K \gamma_i}}\Big) \\
&\leq \frac{1}{2}\sum_{i=1}^{K} \sqrt{p^*(i)\bar{p}_n(i)}\Big(1 - \sqrt{\frac{n}{n+\sum_{i=1}^K \gamma_i}}\Big) \\
&= \frac{1}{2}\sum_{i=1}^{K} \sqrt{p^*(i)\bar{p}_n(i)}\frac{n+\sum_{i=1}^K \gamma_i - \sqrt{n+\sum_{i=1}^K \gamma_i}\sqrt{n}}{n+\sum_{i=1}^K \gamma_i} \\
&\leq \frac{\sum_{i=1}^K \gamma_i}{2(n+\sum_{i=1}^K \gamma_i)}\sum_{i=1}^{K} \sqrt{p^*(i)\bar{p}_n(i)} \\
&= \frac{\sum_{i=1}^K \gamma_i}{2(n+\sum_{i=1}^K \gamma_i)}(1 - H^2(p, \bar{p}_n(i))) \\
&\leq \frac{\sum_{i=1}^K \gamma_i}{2n}.
\end{aligned}
$$

■

Lemma 20 *On event $\mathcal{Z}$, with probability at least $1-\delta$ we have that*

$$
\frac{n}{2}\operatorname{KL}(p^*\|p^{\text{OTB}}) \leq 2\mathcal{R}_T + 2\sum_{i=1}^{K}\gamma_i + (4 + 2\ln(\beta))\ln(1/\delta),
$$

where $\mathcal{R}_T = \sum_{t=n/2+1}^{n}(-\ln p_t(X_t) - (-\ln p^+(X_t)))$.

Proof We start with an application of Jensen's inequality

$$
\begin{aligned}
\operatorname{KL}(p^*\|p^{\text{OTB}}) &= \sum_{i=1}^{K} p^*(i)\ln\left(\frac{p^*(i)}{p^{\text{OTB}}(i)}\right) \\
&= \sum_{i=1}^{K} p^*(i)\left(\ln\left(\frac{p^+(i)}{p^{\text{OTB}}(i)}\right) + \ln\left(\frac{p^*(i)}{p^+(i)}\right)\right) \\
&\leq \sum_{i=1}^{K} p^*(i)\left(\ln\left(\frac{p^+(i)}{p^{\text{OTB}}(i)}\right) + \ln\left(1 + \frac{\sum_{i=1}^K\gamma_i}{n}\right)\right) && \text{(By (15))} \\
&\leq \sum_{i=1}^{K} p^*(i)\ln\left(\frac{p^+(i)}{p^{\text{OTB}}(i)}\right) + \frac{\sum_{i=1}^K\gamma_i}{n} && (\ln(1+x)\leq x \text{ for } x > -1) \\
&\leq \frac{2}{n}\sum_{t=n/2+1}^{n}\sum_{i=1}^{K} p^*(i)\ln\left(\frac{p^+(i)}{p_t(i)}\right) + \frac{\sum_{i=1}^K\gamma_i}{n} && \text{(Jensen's inequality)}
\end{aligned}
$$

We need the following concentration inequality for martingales whose proof can be found in (Beygelzimer et al., 2011, Theorem 1).

Lemma 21 (A version of Freedman's inequality) *Let $X_1, \dots, X_T$ be a martingale difference sequence adapted to a filtration $(\mathcal{F}_i)_{i \le T}$. That is, in particular, $\mathbb{E}_{t-1}[X_t] = 0$. Suppose that $|X_t| \le R$ almost surely. Then for any $\delta \in (0, 1), \lambda \in [0, 1/R]$, with probability at least $1 - \delta$, it holds that*

$$\sum_{t=1}^{T} X_t \le \lambda(e-2) \sum_{t=1}^{T} \mathbb{E}_{t-1}[X_t^2] + \frac{\ln(1/\delta)}{\lambda} .$$

By equation 15, we know that

$$|\ln\left(\frac{p^+(i)}{p_t(i)}\right)| \le \ln\left(\left(1 + \frac{\sum_{i=1}^{K} \gamma_i}{n}\right) 3 \left(1 + \frac{2\,(3m_i + 27\ln(4K/\delta))}{\max\{\gamma_i, \frac{1}{2}m_i - 9\ln(4K/\delta)\}}\right)\right) = \ln(\beta_i) \quad (29)$$

Denote $\lambda^+ = \ln(\max_i \beta_i)$. By Lemma 21, with probability at least $1 - \delta$, for any $\lambda \in [0, \frac{1}{\lambda^+}]$

$$\begin{aligned}
&\sum_{t=n/2+1}^{n} \Big(\sum_{i=1}^{K} p^*(i) \ln(\frac{p^+(i)}{p_t(i)}) - \ln(\frac{p^+(X_t)}{p_t(X_t)}) \Big) \\
&\le \lambda \sum_{t=n/2+1}^{n} \sum_{i=1}^{K} p^*(i) (\ln\left(\frac{p^+(i)}{p_t(i)}\right)^2 + \frac{\ln(1/\delta)}{\lambda} , \qquad (\mathbb{E}[(X - \mathbb{E}[X])^2] \le \mathbb{E}[X^2])
\end{aligned}$$

where we used that $\mathbb{E}[(X - \mathbb{E}[X])^2] \le \mathbb{E}[X^2]$. On event (29), by Corollary 17 we have that

$$\begin{aligned}
&\lambda \sum_{t=n/2+1}^{n} \sum_{i=1}^{K} p^*(i) (\ln\left(\frac{p^+(i)}{p_t(i)}\right)^2 \\
&\le \lambda(2 + \ln(\beta)) \sum_{t=n/2+1}^{n} \sum_{i=1}^{K} p^*(i) \ln\left(\frac{p^+(i)}{p_t(i)}\right) + \frac{\lambda}{2}(2 + \ln(\beta)) \sum_{i=1}^{K} \gamma_i .
\end{aligned}$$

Setting $\lambda = \frac{1}{4 + 2\ln(\beta)}$ we can thus conclude that with probability at least $1 - \delta$

$$\begin{aligned}
&\sum_{t=n/2+1}^{n} \sum_{i=1}^{K} p^*(i) \ln\left(\frac{p^+(i)}{p_t(i)}\right) \\
&\le 2 \sum_{t=n/2+1}^{n} \sum_{i=1}^{K} \big(-\ln p_t(X_t) - (-\ln p^+(X_t))\big) \\
&\quad + \sum_{i=1}^{K} \gamma_i + (4 + 2\ln(\beta)) \ln(1/\delta) ,
\end{aligned}$$

and consequently

$$\begin{aligned}
&\mathrm{KL}(p^*\|p^{\mathrm{OTB}})\\
&\leq \frac{2}{n}\Big(2\sum_{t=n/2+1}^{n}\sum_{i=1}^{K}\big(-\ln p_t(X_t) - (-\ln p^+(X_t))\big)\\
&\quad + 2\sum_{i=1}^{K}\gamma_i + (4+2\ln(\beta))\ln(1/\delta)\Big)\,.
\end{aligned}$$

■

Lemma 22 *We have that*

$$\mathcal{R}_T \leq \sum_{i=1}^{K}\gamma_i \ln(p_{n/2+1}(i)/p^+(i)) + 2\sum_{i=1}^{K}\gamma_i + \frac{n}{2}\mathrm{KL}(\bar{p}_{n/2+1}\|p^*) + \sum_{i=1}^{K}\ln\left(\frac{n_i+\gamma_i}{m_i+\gamma_i}\right)\,.$$

Proof Recall that

$$\mathcal{R}_T = \sum_{t=n/2+1}^{n}\big(-\ln p_t(X_t) - (-\ln p^+(X_t))\big)\,.$$

We can write the computation of $p_{t+1}(i)$ as the action of an FTRL algorithm:

$$p_{t+1}(i) = \operatorname*{arg\,min}_{p\in\Delta^K}\sum_{t'=n/2+1}^{t}\sum_{i=1}^{K}(-\mathbb{1}[X_{t'}=i]\ln(p(i))) + R(p)\,,$$

where $R(p)$ is the regularizer and is defined as

$$R(p) = \sum_{i=1}^{K}\left(-\gamma_i\ln(p(i)) + \sum_{t=1}^{n/2} -\mathbb{1}[X_t=i]\ln(p(i))\right)\,.$$

Denote by $\phi_t(p) = \sum_{t'=n/2+1}^{t}\sum_{i=1}^{K} -\mathbb{1}[X_{t'}=i]\ln(p(i)) + R(p)$ the FTRL potential. By Lemma 7.1 in (Orabona, 2019), we have that

$$\begin{aligned}
\mathcal{R}_T &= \phi_T(p_{n+1}) - R(p_{n/2+1})\\
&\quad - \sum_{t=n/2+1}^{n}\sum_{i=1}^{K}(-\ln(p^+(X_t))) + \sum_{t=n/2+1}^{n}(\phi_t(p_t)-\phi_t(p_{t+1}))\\
&= \phi_T(p_{n+1}) - \phi_T(p^+) - R(p_{n/2+1}) + R(p^+)\\
&\quad + \sum_{t=n/2+1}^{n}(\phi_t(p_t)-\phi_t(p_{t+1}))\,.
\end{aligned}$$

Now, since $\phi_T(p_{n+1}) - \phi_T(p^+) \le 0$ and $\phi_{t-1}(p_t) - \phi_{t-1}(p_{t+1}) \le 0$ we have that

$$\begin{aligned}
\mathcal{R}_T &\le R(p^+) - R(p_{n/2+1}) + \sum_{t=n/2+1}^{n} \ln\left(\frac{p_{t+1}(X_t)}{p_t(X_t)}\right) \\
&= \sum_{i=1}^{K} \gamma_i \ln(p_{n/2+1}(i)/p^+(i)) + \sum_{t=1}^{n/2} \ln(p_{n/2+1}(X_t)/p^+(X_t)) \\
&\quad + \sum_{t=n/2+1}^{n} \sum_{i=1}^{K} \mathbb{1}[X_t = i] \ln\Big(\frac{(n_{t,i} + \gamma_i)(t - 1 + \sum_{i=1}^K \gamma_i)}{(n_{t-1,i} + \gamma_i)(t + \sum_{i=1}^K \gamma_i)}\Big) \\
&\le \sum_{i=1}^{K} \gamma_i \ln(p_{n/2+1}(i)/p^+(i)) + \sum_{t=1}^{n/2} \ln(p_{n/2+1}(X_t)/p^+(X_t)) \\
&\quad + \sum_{t=n/2+1}^{n} \sum_{i=1}^{K} \mathbb{1}[X_t = i] \ln\Big(\frac{n_{t,i} + \gamma_i}{n_{t-1,i} + \gamma_i}\Big) && \Big(\tfrac{t-1+\sum_{i=1}^K \gamma_i}{t+\sum_{i=1}^K \gamma_i} \le 1\Big) \\
&= \sum_{i=1}^{K} \gamma_i \ln(p_{n/2+1}(i)/p^+(i)) + \sum_{t=1}^{n/2} \ln(p_{n/2+1}(X_t)/p^+(X_t)) + \sum_{i=1}^{K} \ln\left(\frac{n_i + \gamma_i}{m_i + \gamma_i}\right),
\end{aligned}$$

where the last equality follows from a telescoping sum. At this point, we can use that $\sum_{t=1}^{n/2} \mathbb{1}[X_t = i] = \frac{n}{2}\bar{p}_{n/2+1}(i)$ to see that

$$\begin{aligned}
&\sum_{t=1}^{n/2} \ln(p_{n/2+1}(X_t)/p^+(X_t)) \\
&= \sum_{i=1}^{K} \sum_{t=1}^{n/2} \mathbb{1}[X_t = i] \left(\ln\left(\frac{p_{n/2+1}(i)}{\bar{p}_{n/2}(i)}\right) + \ln\left(\frac{\bar{p}_{n/2}(i)}{p^*(i)}\right) + \ln\left(\frac{p^*(i)}{p^+(i)}\right)\right) \\
&= \sum_{i=1}^{K} \sum_{t=1}^{n/2} \mathbb{1}[X_t = i] \left(\ln\left(\frac{p_{n/2+1}(i)}{\bar{p}_{n/2}(i)}\right) + \ln\left(\frac{p^*(i)}{p^+(i)}\right)\right) + \frac{n}{2}\mathrm{KL}(\bar{p}_{n/2+1} \| p^*) \\
&\le \sum_{i=1}^{K} \sum_{t=1}^{n/2} \mathbb{1}[X_t = i] \left(\ln\left(\frac{p_{n/2+1}(i)}{\bar{p}_{n/2}(i)}\right) + \frac{\gamma_i}{n}\right) + \frac{n}{2}\mathrm{KL}(\bar{p}_{n/2+1} \| p^*) \\
&\qquad\qquad \Big(\tfrac{p^*(i)}{p^+(i)} \le 1 + \tfrac{\sum_{i=1}^K \gamma_i}{n} \text{ and } \ln(1+x) \le x \text{ for } x > 0\Big) \\
&= \sum_{i=1}^{K} \sum_{t=1}^{n/2} \mathbb{1}[X_t = i] \left(\ln\left(\frac{(m_i + \gamma_i)n/2}{m_i(n/2 + \sum_{i=1}^K \gamma_i)}\right) + \frac{\sum_{i=1}^K \gamma_i}{n}\right) + \frac{n}{2}\mathrm{KL}(\bar{p}_{n/2+1} \| p^*) \\
&\qquad\qquad \text{(By definition)} \\
&\le \sum_{i=1}^{K} \sum_{t=1}^{n/2} \mathbb{1}[X_t = i] \left(\frac{\sum_{i=1}^K \gamma_i}{m_i} + \frac{\sum_{i=1}^K \gamma_i}{n}\right) + \frac{n}{2}\mathrm{KL}(\bar{p}_{n/2+1} \| p^*) \quad \Big(\text{Since } \tfrac{n/2}{n/2+\sum_{i=1}^K \gamma_i} \le 1\Big) \\
&\le 2\sum_{i=1}^{K} \gamma_i + \frac{n}{2}\mathrm{KL}(\bar{p}_{n/2+1} \| p^*),
\end{aligned}$$

which combined with the above completes the proof. ∎

Lemma 23 *With probability at least* $1-\delta$

$$\mathrm{KL}(\bar{p}_{n+1}\|p^*) \leq \mathbb{E}[\mathrm{KL}(\bar{p}_{n+1}\|p^*)] + \frac{6K + 6\ln(1/\delta)}{n} \leq \frac{7K + 6\ln(1/\delta)}{n}.$$

Proof The first inequality can be found as Corollary 1.7 in (Agrawal, 2022). The second inequality uses that $\mathbb{E}[\mathrm{KL}(\bar{p}_{n+1}\|p^*)] \leq \frac{K-1}{n}$ (Paninski, 2003, Section 4). ∎

Lemma 24 *Suppose that* $\frac{\ln(K/\delta)}{J} > 1$*. Then on event* $\mathcal{Z}$ *we have that*

$$\begin{aligned}
&\sum_{i=1}^K \gamma_i \big(\ln(p_{n/2+1}(i)/p^+(i)) + 3\big) + \sum_{i=1}^K \ln\left(\frac{n_i+\gamma_i}{m_i+\gamma_i}\right) + (2+\ln(\beta))\ln(1/\delta) \\
&\leq 7\ln(K/\delta) + 20K + 2\ln(1/\delta)\ln(400J).
\end{aligned}$$

Proof On event $\mathcal{Z}$ we have that

$$\begin{aligned}
&\sum_{i=1}^K \gamma_i \ln(p_{n/2+1}(i)/p^+(i)) + \sum_{i=1}^K \ln\left(\frac{n_i+\gamma_i}{m_i+\gamma_i}\right) \\
&\leq \sum_{i=1}^K \gamma_i \ln\left(3 + \frac{6\,(3m_i + 27\ln(4K/\delta))}{\max\{\gamma_i, \frac{1}{2}m_i - 9\ln(4K/\delta)\}}\right) + \sum_{i=1}^K \ln\left(6 + \frac{18\ln(K/\delta)}{m_i+\gamma_i}\right). \\
&\leq 2K\ln(100) + \ln(K/\delta) + J\ln(J) \\
&\leq 2K\ln(100) + \ln(K/\delta) + \ln(K/\delta)\ln(J) \\
&\leq 2K\ln(100) + \ln(1/\delta) + \ln(1/\delta)\ln(J) + 2\ln(K)^2 \\
&\leq 20K + \ln(1/\delta) + \ln(1/\delta)\ln(400J)
\end{aligned}$$

where we used that $J < \ln(K/\delta)$ and $\ln(K)^2 \leq K$. By definition of γ_i we can also bound

$$\sum_{i=1}^K \gamma_i = \ln(K/\delta).$$

Finally, since $J < \ln(K/\delta)$ we can also see that $\beta \leq 400J$. ∎

Lemma 25 *Suppose that* $\frac{\ln(K/\delta)}{J} \leq 1$*. Then on event* $\mathcal{Z}$ *we have that with probability at least* $1-3\delta$

$$\begin{aligned}
&\sum_{i=1}^K \gamma_i \big(\ln(p_{n/2+1}(i)/p^+(i)) + 3\big) + \sum_{i=1}^K \ln\left(\frac{n_i+\gamma_i}{m_i+\gamma_i}\right) + (2+\ln(\beta))\ln(1/\delta) \\
&\leq 50\ln(1/\delta)\ln(800J) + 2\widetilde{J}\ln\left(24\left(\ln\left(\frac{\widetilde{J}}{\ln(1/\delta)}\right)\vee 1\right)\right) + 10K\ln(100).
\end{aligned}$$

Proof Since $\frac{\ln(K/\delta)}{J} \leq 1$ and we have $\gamma_i = 1$ or $\gamma_i = 0$ and by definition $\sum_{i=1}^K \gamma_i = J$. Furthermore, we can see that

$$\beta \leq 800 \ln(K/\delta) \leq 800J \,,$$

Recall that $\widetilde{\mathcal{J}} = \{i : np^*(i) < 32\ln(K/\delta)\}$ and $\widetilde{J} = \max\{3, |\widetilde{\mathcal{J}}|\}$. On event $\mathcal{Z}$ we have that $\frac{n_i+\gamma_i}{np^*(i)+\gamma_i} \leq \beta_i$ for all i. For $i \notin \mathcal{J}$ we have that $\frac{n_i}{m_i} \leq \beta_i \leq 3 + \frac{6(3m_i+27\ln(4K/\delta))}{\max\{\gamma_i, \frac{1}{2}m_i - 9\ln(4K/\delta)\}} \leq 100$ and thus

$$\begin{aligned}
&\sum_{i=1}^K \gamma_i \ln(p_{n/2+1}/p^+(i)) + \sum_{i=1}^K \ln\left(\frac{n_i+\gamma_i}{m_i+\gamma_i}\right) \\
&\leq \sum_{i\in\mathcal{J}} \ln\left(\frac{n_i+1}{np^*(i)+1}\right) + \sum_{i\in[K]\setminus\mathcal{J}} \ln\left(\frac{n_i}{m_i}\right) \\
&\leq \sum_{i\in\mathcal{J}} \ln\left(\frac{n_i+1}{np^*(i)+1}\right) + K\ln(100) \,.
\end{aligned}$$

By a union bound and Bennet's inequality, for all $i \in [K]$, with probability at least $1-\delta$, we have that

$$|n_i - np^*(i)| \leq \sqrt{2np^*(i)\ln(K/\delta)} + \ln(K/\delta)/3 \,.$$

On this event, for $i \notin \widetilde{\mathcal{J}}$, we have that

$$|n_i - np^*(i)| \leq 2np^*(i) \,,$$

which implies that $n_i \leq 3np^*(i)$. On the same event, for $i \in \widetilde{\mathcal{J}}$ we have that $n_i \leq 9\ln(K/\delta)$. Therefore, we have that

$$\begin{aligned}
&\sum_{i\in\mathcal{J}} \ln\left(\frac{n_i+1}{np^*(i)+1}\right) + K\ln(100) \\
&\leq \sum_{i\in\widetilde{\mathcal{J}}} \ln\left(\frac{n_i+1}{np^*(i)+1}\right) + 2K\ln(100) \,,
\end{aligned}$$

where we used that for $i \notin \mathcal{J}$ $n_i \geq 32\ln(K/\delta)$. Let $p^*_{\min} = \min_{i\in[K]} p^*(i)$. Suppose that $\sum_{i\in[K]\setminus\widetilde{\mathcal{J}}} p^*(i) > \frac{1}{3}$. Let $\widetilde{X}_i \sim B(n, \min\{1, 3p^*(i)\})$. Let $x \geq 1$ and $z \geq 2\widetilde{J}\exp(-\frac{1}{4}x(3np^*_{\min}+1))$. By Lemma 18 we have that

$$\mathbb{P}(\sum_{i\in\widetilde{\mathcal{J}}} \mathbb{1}\{n_i > 3np^*(i) + x(3np^*(i)+1)\} \geq z) \leq \mathbb{P}(\sum_{i\in\widetilde{\mathcal{J}}} \mathbb{1}\{\widetilde{X}_i > 3np^*(i) + x(3np^*(i)+1)\} \geq z)$$

By Bernstein's inequality, we have that

$$\begin{aligned}
\mathbb{P}(\widetilde{X}_i > 3np^*(i) + x(3np^*(i)+1)) &\leq \exp\left(-\frac{\frac{1}{2}(x(3np^*(i)+1))^2}{3np^*(i) + \frac{1}{3}x(3np^*(i)+1)}\right) \\
&\leq \exp\left(-\tfrac{1}{4}x(3np_{\min}+1)\right)
\end{aligned}$$

Let $Z \sim B(\widetilde{J}, \exp(-\frac{1}{4}x(3np_{\min}+1)))$. By Bernstein's inequality, we have that

$$\begin{aligned}
&\mathbb{P}(\sum_{i \in \widetilde{\mathcal{J}}} \mathbb{1}\{\widetilde{X}_i > 3np^*(i) + x(3np^*(i)+1)\} \geq z) \\
&= \mathbb{P}(Z \geq z) \\
&= \mathbb{P}(Z - \mathbb{E}[Z] \geq z - \mathbb{E}[Z]) \\
&\leq \mathbb{P}(Z - \mathbb{E}[Z] \geq \tfrac{1}{2}z) \\
&\leq \exp\left(-\frac{1}{8}\frac{z^2}{\widetilde{J}\exp(-\frac{1}{4}x(3np_{\min}+1)) + 1/6z}\right) \\
&\leq \exp\left(-\frac{z}{24}\right)
\end{aligned}$$

Let $\mathcal{Z} = \{i \in \widetilde{\mathcal{J}} : n_i > 3np^*(i) + x(3np^*(i)+1)\}$. With probability at least $1 - \exp(-z/24)$ we have that $|\mathcal{Z}| \leq z$ and thus

$$\begin{aligned}
\leq \sum_{i \in \widetilde{\mathcal{J}}} \ln\left(\frac{n_i+1}{np^*(i)+1}\right) + 2K\ln(100) &\leq \sum_{i \in \mathcal{Z}} \ln\left(\frac{n_i+1}{np^*(i)+1}\right) + (\widetilde{J} - |\mathcal{Z}|)\ln(6x) + 2K\ln(100) \\
&\leq z\ln(400(1+\ln(K/\delta))) + (\widetilde{J} - |\mathcal{Z}|)\ln(6x) + K\ln(100)\,.
\end{aligned}$$

Setting $z = \max\{24\ln(1/\delta), 2\widetilde{J}\exp(-\frac{x}{4}(3np^*_{\min}+1))\}$ and $x = 4\left(\ln\left(\frac{\widetilde{J}}{\ln(1/\delta)}\right) \vee 1\right)$ we find that with probability at least $1 - 2\delta$

$$\begin{aligned}
&\sum_{i=1}^K \gamma_i \ln(p_{n/2+1}/p^+(i)) + \sum_{i=1}^K \ln\left(\frac{n_i+\gamma_i}{m_i+\gamma_i}\right) \\
&\leq 24\ln(1/\delta)\ln(400(1+\ln(K/\delta))) + \widetilde{J}\ln\left(24\ln\left(\ln\left(\frac{\widetilde{J}}{\ln(1/\delta)}\right) \vee 1\right)\right) + K\ln(100)\,.
\end{aligned}$$

Now, suppose that $\sum_{i \in [K] \setminus \widetilde{\mathcal{J}}} p^*(i) < \frac{1}{3}$. In that case we can always construct a two sets $R_1 \subset \widetilde{\mathcal{J}}$ and $R_2 \subset \widetilde{\mathcal{J}}$ such that $R_1 \bigcap R_2 = \widetilde{\mathcal{J}}$, $\sum_{i \notin R_1} p^*(i) \geq \frac{1}{3}$ and $\sum_{i \notin R_2} p^*(i) \geq \frac{1}{3}$. With these two sets we can repeat the analysis for the first case, where we apply Lemma 18 to R_1 and R_2 separately, except now we need a union bound to combine the analyses. Ultimately, we find that, with probability at least $1 - 3\delta$

$$\begin{aligned}
&\sum_{i=1}^K \gamma_i \ln(p_{n/2+1}/p^+(i)) + \sum_{i=1}^K \ln\left(\frac{n_i+\gamma_i}{m_i+\gamma_i}\right) \\
&\leq 48\ln(1/\delta)\ln(400(1+\ln(K/\delta))) + 2\widetilde{J}\ln\left(24\left(\ln\left(\frac{\widetilde{J}}{\ln(1/\delta)}\right) \vee 1\right)\right) + 4K\ln(100)\,.
\end{aligned}$$

At this point we can use $\ln(K/\delta) \leq J$ to complete the proof. ∎

Appendix D. Additional Proofs for Section 3

Theorem 8 *Suppose that $n > \frac{4}{3}\ln(1/\delta)$. Denote by $\widehat{p}^0 \in \triangle^K$ the output of an estimator $\widehat{p}$ on the sample $X_1 = \cdots = X_n = 1$. Then, for any $\widehat{p}$, there exists $p^* \in \triangle^K$ such that*

$$\mathbb{P}_{p^*}\left(\mathrm{KL}(p^*\|\widehat{p}) \geq \frac{2\ln(1/\delta)}{3n}\left(\ln\left(\frac{2\ln(1/\delta)(K-1)}{3n\sum_{i=2}^K \widehat{p}^0(i)}\right) - 1\right) + \sum_{i=2}^K \widehat{p}^0(i)\right) > \delta\,.$$

Proof We construct a distribution p^* that is hard for $\widehat{p}$ as follows: for $i' = \arg\min_{i\in\{2,\dots,K\}} \widehat{p}^0(i)$ let

$$p^*(i) = \begin{cases} 1 - \frac{\alpha}{n} & \text{if } i = 1 \\ \frac{\alpha}{n} & \text{if } i = i' \\ 0 & \text{otherwise,} \end{cases}$$

where $0 < \alpha \leq n/2$ will be chosen later. For this choice of p^*, using that $\ln(1-x) \geq -x - x^2$ for $0 < x < \frac{1}{2}$, we have that

$$\mathbb{P}(n_1 = 1) = \left(1 - \frac{\alpha}{n}\right)^n = \exp\left(n\ln\left(1 - \frac{\alpha}{n}\right)\right) \geq \exp\left(-\alpha - \frac{\alpha^2}{n}\right) > \exp\left(-\frac{3\alpha}{2}\right),$$

where the last inequality follows from $n > 2\alpha$, and the event $n_1 = n$ is equivalent to $X_1 = \cdots = X_n = 1$. By choosing $\alpha = \frac{2}{3}\ln(1/\delta)$, we find that

$$\mathbb{P}(n_1 = n) > \delta.$$

On the event that $n_1 = n$, the KL divergence of p^* from $\widehat{p}$ is

$$\begin{aligned}\mathrm{KL}(p^*\|\widehat{p}) &= \frac{\alpha}{n}\ln\left(\frac{\alpha}{n\widehat{p}^0(i')}\right) + \left(1 - \frac{\alpha}{n}\right)\ln\left(\frac{1 - \frac{\alpha}{n}}{1 - \sum_{i=2}^K \widehat{p}^0(i)}\right) \\ &\geq \frac{\alpha}{n}\ln\left(\frac{\alpha}{n\widehat{p}^0(i')}\right) + \sum_{i=2}^K \widehat{p}^0(i) - \frac{\alpha}{n}\,,\end{aligned}$$

where the last inequality follows by $\ln(x) \geq 1 - 1/x$ for $x > 0$. By the definition of i', we get that $\widehat{p}^0(i') \leq \sum_{i=2}^K \frac{\widehat{p}^0(i)}{(K-1)}$, which gives us

$$\begin{aligned}\mathrm{KL}(p^*\|\widehat{p}) &\geq \frac{\alpha}{n}\ln\left(\frac{\alpha}{n\widehat{p}^0(i')}\right) + \sum_{i=2}^K \widehat{p}^0(i) - \frac{\alpha}{n} \\ &\geq \frac{\alpha}{n}\ln\left(\frac{\alpha(K-1)}{n\sum_{i=2}^K \widehat{p}^0(i)}\right) + \sum_{i=2}^K \widehat{p}^0(i) - \frac{\alpha}{n}.\end{aligned}$$

Recalling that $\alpha = \frac{2}{3}\ln(1/\delta)$, we obtain the following lower bound:

$$\mathrm{KL}(p^*\|\widehat{p}) \geq \frac{2\ln(1/\delta)}{3n}\left(\ln\left(\frac{2\ln(1/\delta)(K-1)}{3n\sum_{i=2}^K \widehat{p}^0(i)}\right) - 1\right) + \sum_{i=2}^K \widehat{p}^0(i).$$

All in all, we have shown that there exists a p^*, dependent on $\widehat{p}$, such that

$$\mathbb{P}\Big(\operatorname{KL}(p^*\|\widehat{p}) \geq \frac{2\ln(1/\delta)}{3n}\left(\ln\left(\frac{2\ln(1/\delta)(K-1)}{3n\sum_{i=2}^K \widehat{p}^0(i)}\right) - 1\right) + \sum_{i=2}^K \widehat{p}^0(i)\Big) > \delta\,.$$

■

Theorem 11 *For $K \geq 2$ and any $\delta \in (0,1]$, the minimax rate is at least*

$$r_n^*(\delta) \geq \frac{2\ln(K-1)\ln(1/\delta)}{3n} \qquad \text{for all } n > \frac{4}{3}\ln(1/\delta).$$

Proof Consider the distributions corresponding to (16). For each of them the probability of the all ones sequence is

$$P_j^n(X_1 = \cdots = X_n = 1) = (1 - \tfrac{\alpha}{n})^n > \delta,$$

where the inequality holds because $\ln(1+z) \geq z - z^2$ for $z \geq -1/2$, so that

$$\ln\Big(1 - \frac{\alpha}{n}\Big) \geq -\frac{\alpha}{n} - \Big(\frac{\alpha}{n}\Big)^2 = -\frac{\alpha}{n}\Big(1 + \frac{\alpha}{n}\Big) > -\frac{3\alpha}{2n} = \frac{\ln\delta}{n},$$

where we have used that $n > 2\alpha$ by assumption.

For an arbitrary weak test Ψ, let j^* be the output of Ψ on the all ones sequence. Then condition (19) of Lemma 10 is fulfilled because

$$P_{j^*}^n(\Psi = j^*) \geq P_{j^*}^n(X_1 = \cdots = X_n = 1) = (1 - \tfrac{\alpha}{n})^n > \delta.$$

The lemma therefore tells us that

$$r_n^*(\delta) \geq \inf_P \max_j \operatorname{KL}(P_j\|P) \geq \inf_P \frac{1}{K-1}\sum_{j=2}^K \operatorname{KL}(P_j\|P) = \frac{1}{K-1}\sum_{j=2}^K \operatorname{KL}(P_j\|\bar{P}),$$

where $\bar{P} = \frac{1}{K-1}\sum_{j=2}^K P_j$. Since, for any j,

$$\operatorname{KL}(P_j\|\bar{P}) = \frac{\alpha}{n}\ln\frac{\alpha/n}{\alpha/n/(K-1)} = \frac{\alpha}{n}\ln(K-1) = \frac{2\ln(K-1)\ln(1/\delta)}{3n},$$

we conclude that $r_n^*(\delta) \geq \frac{2\ln(K-1)\ln(1/\delta)}{3n}$, as required.

■

Lemma 9 *Let $P_1, \ldots, P_M$ be distributions defined on a sample space $\mathcal{X}$, and let P_j^n be the distribution of n independent draws from P_j. Given any $\delta \in [0,1]$, suppose that*

$$\forall j \neq k: \quad \operatorname{KL}\Big(P_j\|\frac{P_j + P_k}{2}\Big) \geq s_n, \tag{17}$$

$$\inf_\Psi \max_j P_j^n(\Psi \neq j) > \delta, \tag{18}$$

where the infimum is over all possible hypothesis tests $\Psi : \mathcal{X}^n \to \{1, \ldots, M\}$. Then

$$\inf_{\widehat{P}} \max_j P_j^n\Big(\operatorname{KL}(P_j\|\widehat{P}) \geq s_n\Big) > \delta,$$

where the infimum is over all estimators based on n observations.

Proof Let $\widehat{P}$ be any estimator, and let $\Psi^* = \arg\min_{j\in[M]} \mathrm{KL}(P_j\|\widehat{P})$ (breaking ties arbitrarily) be the corresponding minimum KL divergence hypothesis test. Then, for any j, we have

$$\mathrm{KL}(P_{\Psi^*}\|\widehat{P}) \le \mathrm{KL}(P_j\|\widehat{P}).$$

Hence, on the event that $\Psi^* \neq j$,

$$\begin{aligned}\mathrm{KL}(P_j\|\widehat{P}) &\ge \tfrac{1}{2}\mathrm{KL}(P_j\|\widehat{P}) + \tfrac{1}{2}\mathrm{KL}(P_{\Psi^*}\|\widehat{P}) \ge \min_P \left\{\tfrac{1}{2}\mathrm{KL}(P_j\|P) + \tfrac{1}{2}\mathrm{KL}(P_{\Psi^*}\|P)\right\}\\ &= \tfrac{1}{2}\mathrm{KL}\left(P_j\|\frac{P_j+P_{\Psi^*}}{2}\right) + \tfrac{1}{2}\mathrm{KL}\left(P_{\Psi^*}\|\frac{P_j+P_{\Psi^*}}{2}\right) \ge s_n,\end{aligned}$$

where the last inequality uses (17). Thus, for all j,

$$P_j^n\big(\mathrm{KL}(P_j\|\widehat{P}) \ge s_n\big) \ge P_j^n\big(\Psi^* \neq j\big).$$

Taking the maximum over j we find that

$$\max_j P_j^n\big(\mathrm{KL}(P_j\|\widehat{P}) \ge s_n\big) \ge \max_j P_j^n\big(\Psi^* \neq j\big) \ge \inf_\Psi \max_j P_j^n\big(\Psi \neq j\big) > \delta,$$

where the last inequality holds by assumption (18). The result follows by taking the infimum over $\widehat{P}$. ∎

Lemma 10 *Let $P_1, \ldots, P_M$ be distributions defined on a sample space $\mathcal{X}$, and let P_j^n be the distribution of n independent samples from P_j. Given any $\delta \in [0,1]$, suppose that*

$$\inf_\Psi \max_j P_j^n(\Psi = j) > \delta, \tag{19}$$

where the infimum is over all possible weak hypothesis tests $\Psi : \mathcal{X}^n \to \{1,\ldots,M\}$. Then

$$\inf_{\widehat{P}} \max_j P_j^n\Big(\mathrm{KL}(P_j\|\widehat{P}) \ge s_n\Big) > \delta \qquad \text{for } s_n = \inf_P \max_j \mathrm{KL}(P_j\|P). \tag{20}$$

Proof Let $\widehat{P}$ be any estimator, and let $\widehat{\Psi} = \arg\max_{j\in[M]} \mathrm{KL}(P_j\|\widehat{P})$ (breaking ties arbitrarily) be the corresponding maximum KL divergence weak hypothesis test. Then we have

$$\mathrm{KL}(P_{\widehat{\Psi}}\|\widehat{P}) = \max_j \mathrm{KL}(P_j\|\widehat{P}) \ge s_n,$$

for s_n as defined in (20). Hence, for any j,

$$P_j^n\Big(\mathrm{KL}(P_j\|\widehat{P}) \ge s_n\Big) \ge P_j^n\big(\widehat{\Psi} = j\big).$$

Taking the maximum over j, we find that

$$\begin{aligned}\max_j P_j^n\Big(\mathrm{KL}(P_j\|\widehat{P}) \ge s_n\Big) &\ge \max_j P_j^n\big(\widehat{\Psi} = j\big)\\ \inf_{\widehat{P}} \max_j P_j^n\Big(\mathrm{KL}(P_j\|\widehat{P}) \ge s_n\Big) &\ge \inf_\Psi \max_j P_j^n\big(\Psi = j\big) > \delta,\end{aligned}$$

where the last inequality holds by assumption (19). ∎

Lemma 13 *For any* $\alpha \in \left(0, \frac{1}{2K}\right]$ *and* $\epsilon > 0$*, the covering and packing entropies are bounded by*

$$N(\triangle_0^K, \epsilon, \mathrm{KL}) \leq K \ln\Big(\frac{\alpha 2\sqrt{2K}}{\epsilon} + 1\Big), \qquad M(\triangle_0^K, \epsilon, V) \geq K \ln \frac{\alpha}{\epsilon} + \frac{K}{2} \ln \frac{K\pi}{8}.$$

Proof We will first consider the upper bound on $N(\triangle_0^K, \epsilon, \mathrm{KL})$ before proving the lower bound on $M(\triangle_0^K, \epsilon, V)$.

Upper Bound: Let $\chi^2(p,q) = \sum_{i=1}^K \frac{(p(i)-q(i))^2}{q(i)}$ denote the χ^2 divergence. Then $\mathrm{KL}(p\|q) \leq \chi^2(p,q)$ for any p, q in $\triangle^K$ (Tsybakov, 2009). Consequently, for any $p, q \in \triangle_0^K$,

$$\mathrm{KL}(p\|q) \leq \chi^2(p,q) = \sum_{i=1}^K \frac{(p(i)-q(i))^2}{q(i)} = 2K \sum_{i=1}^K (p(i)-q(i))^2 = 2K\|p-q\|_2^2.$$

We may equate $\triangle_0^K$ with $B_2(\alpha, u) = u + \alpha B_2$, where B_2 is the ℓ_2 unit ball in $\mathbb{R}^K$. Then by convexity of B_2, any optimal ℓ_2 ϵ-cover of $B_2(\alpha,u)$ will consist of points inside $B_2(\alpha,u)$ and therefore induces a KL $(2K\epsilon^2)$-cover of $\triangle_0^K$. Hence we get the following upper bound on the covering entropy:

$$\begin{aligned} N(\triangle_0^K, 2K\epsilon^2, \mathrm{KL}) &\leq N(B_2(\alpha,u), \epsilon, \|\cdot\|_2) = N(\alpha B_2, \epsilon, \|\cdot\|_2) \\ &= N(B_2, \frac{\epsilon}{\alpha}, \|\cdot\|_2) \leq K \ln\Big(\frac{2\alpha}{\epsilon} + 1\Big). \\ N(\triangle_0^K, \epsilon^2, \mathrm{KL}) &\leq K \ln\Big(\frac{\alpha 2\sqrt{2K}}{\epsilon} + 1\Big). \end{aligned}$$

Lower Bound: Since the packing entropy $M(\triangle_0^K, \epsilon, V)$ is always an upper bound on the covering entropy $N(\triangle_0^K, \epsilon, V)$, it is sufficient to find a lower bound on the covering entropy. Let B_1 be the unit ℓ_1 ball in $\mathbb{R}^K$ and note that $V(p,q) = \|p-q\|_1$. Then

$$M(\triangle_0^K, \epsilon, V) \geq N(\triangle_0^K, \epsilon, V) = N(\alpha B_2, \epsilon, \|\cdot\|_1).$$

In order to construct a cover of αB_1 by m ℓ_1-balls of radius ϵ, we need that the total volume $m\,\mathrm{Vol}(\epsilon B_1)$ of the balls in the cover is at least the volume $\mathrm{Vol}(\alpha B_2)$ of the set being covered, so that

$$N(\alpha B_2, \epsilon, \|\cdot\|_1) \geq \ln \frac{\mathrm{Vol}(\alpha B_2)}{\mathrm{Vol}(\epsilon B_1)} = \ln \frac{\alpha^K \mathrm{Vol}(B_2)}{\epsilon^K \mathrm{Vol}(B_1)}.$$

The volume of the unit ℓ_p ball (for $p \geq 1$) in dimension K is

$$\mathrm{Vol}(B_p) = 2^K \frac{\Gamma\Big(1 + \frac{1}{p}\Big)^K}{\Gamma(1 + \frac{K}{p})}.$$

Consequently (using that $\Gamma(2) = 1$ and $\Gamma(3/2) = \sqrt{\pi}/2$),

$$\begin{aligned} \ln \frac{\mathrm{Vol}(B_2)}{\mathrm{Vol}(B_1)} &= \ln \frac{\Gamma(1+K)}{\Gamma(1+\frac{K}{2})} - K \ln \frac{2}{\sqrt{\pi}} \\ &\geq \frac{K}{2} \ln \frac{K}{2} - K \ln \frac{2}{\sqrt{\pi}} = \frac{K}{2} \ln \frac{K\pi}{8}. \end{aligned}$$

Putting all inequalities together, we conclude that

$$M(\triangle_0^K, \epsilon, V) \geq K \ln \frac{\alpha}{\epsilon} + \frac{K}{2} \ln \frac{K\pi}{8},$$

as claimed. ■

Theorem 14 *There exists an absolute constant $C > 0$ such that, for any $K \geq 2$ and $\delta \in (0, 1/2)$, the minimax rate is at least*

$$r_n^*(\delta) \geq \frac{CK}{n} \qquad \text{for all } n \geq \tfrac{\ln 2}{2} K^2.$$

Proof By Lemma 13, we can apply Theorem 12 with

$$N(\epsilon) = K \ln\Big(\frac{\alpha 2\sqrt{2K}}{\epsilon} + 1\Big), \qquad\qquad M(\epsilon) = K \ln \frac{\alpha}{\epsilon} + \frac{K}{2} \ln \frac{K\pi}{8}.$$

Let $\alpha = C_1/\sqrt{n}$ for $C_1 \in (0, \frac{\sqrt{n}}{2K}]$ to be determined. Then

$$\epsilon_n^2 = \frac{N(\epsilon_n)}{n} = \frac{K}{n} \ln\Big(\frac{C_1 2\sqrt{2K}}{\epsilon_n \sqrt{n}} + 1\Big)$$

has solution $\epsilon_n = C_2\sqrt{\frac{K}{n}}$ for $C_2 > 0$ such that

$$C_2^2 = \ln\Big(\frac{C_1 2\sqrt{2}}{C_2} + 1\Big).$$

Now we find $\underline{\epsilon}_n = C_3\sqrt{K}\alpha$ for some $C_3 > 0$ such that

$$\begin{aligned}
M(\underline{\epsilon}_n) &\geq 4n\epsilon_n^2 + 2\ln 2\\
K \ln \frac{\alpha}{\underline{\epsilon}_n} + \frac{K}{2} \ln \frac{K\pi}{8} &\geq 4C_2^2 K + 2\ln 2\\
K \ln \frac{1}{C_3\sqrt{K}} + \frac{K}{2} \ln \frac{K\pi}{8} &\geq 4C_2^2 K + 2\ln 2\\
\frac{K}{2} \ln \frac{\pi}{8} &\geq (4C_2^2 + \ln C_3)K + 2\ln 2,
\end{aligned}$$

for which it is sufficient if

$$\frac{1}{2} \ln \frac{\pi}{32} \geq 4C_2^2 + \ln C_3.$$

The constants cannot be optimized in closed form to maximize $\underline{\epsilon}_n = C_1 C_3\sqrt{\frac{K}{n}}$, but a reasonable choice is to take $C_1 = \frac{\sqrt{\ln 2}}{2\sqrt{2}}$ (which falls in the allowed range by assumption on n) such that $C_2 = \sqrt{\ln 2}$. This leads to $C_3 = \exp\Big(\frac{1}{2}\ln\frac{\pi}{32} - 4\ln 2\Big)$.

Having satisfied the conditions of Theorem 12, it tells us that

$$\inf_{\widehat{p}} \sup_{p^* \in \triangle_0^K} \mathbb{P}_{p^*}\, n\Big(V(p^*, \widehat{p}) \geq \frac{C_1 C_3}{2}\sqrt{\frac{K}{n}}\Big) \geq \frac{1}{2}.$$

By Pinsker's inequality this implies that

$$\inf_{\widehat{p}} \sup_{p^* \in \triangle_0^K} \mathbb{P}_{p^*}\left(\mathrm{KL}(p^* \| \widehat{p}) \geq \frac{C_1^2 C_3^2}{8} \frac{K}{n}\right) \geq \frac{1}{2} > \delta,$$

and therefore $r_n^*(\delta) \geq C \frac{K}{n}$ for $C = \frac{C_1^2 C_3^2}{8} \approx 4.1 \times 10^{-6}$. ■