JMLR: Workshop and Conference Proceedings vol 40:1–35, 2015

On Learning Distributions from their Samples

Sudeep Kamath SUKAMATH@PRINCETON.EDU
Princeton University

Alon Orlitsky ALON@UCSD.EDU
Venkatadheeraj Pichapati DHEERAJPV7@GMAIL.COM
Ananda Theertha Suresh ASURESH@UCSD.EDU
University of California, San Diego

Abstract

One of the most natural and important questions in statistical learning is: how well can a distribution be approximated from its samples. Surprisingly, this question has so far been resolved for only one loss, the KL-divergence and even in this case, the estimator used is ad hoc and not well understood. We study distribution approximations for general loss measures. For ℓ_2^2 we determine the best approximation possible, for ℓ_1 and χ^2 we derive tight bounds on the best approximation, and when the probabilities are bounded away from zero, we resolve the question for all sufficiently smooth loss measures, thereby providing a coherent understanding of the rate at which distributions can be approximated from their samples.

Keywords: Probability estimation, online learning, min-max loss, f-divergence

1. Introduction

1.1. Definitions and previous results

Many natural phenomena are believed to be of probabilistic nature. For example, written text, spoken language, stock prices, genomic composition, disease symptoms, physical characteristics, communication noise, traffic patterns, and many more, are commonly assumed to be generated according to some unknown underlying distribution.

It is therefore of practical importance to approximate an underlying distribution from its observed samples. Namely, given samples from an unknown distribution p, to find a distribution q that approximates p in a suitable sense. Yet surprisingly, despite many years of statistical research, very little is known about this problem.

The simplest rigorous formulation of this problem may be in terms of min-max performance. Any distribution $p = (p_1, \dots, p_k)$ over $[k] \overset{\text{def}}{=} \{1, \dots, k\}$ corresponds to an element of the simplex

$$\Delta_k \overset{\text{def}}{=} \left\{ p^k \in \mathbb{R}^k_{\geq 0} : \sum_{i=1}^{k} p_i = 1 \right\}.$$

For two distributions $p, q \in \Delta_k$, let $L(p, q)$ be the *loss* when the true distribution p is approximated by the estimate q. The right loss function typically depends on the application. For example, for compression and investment applications, the relevant loss is often the Kullback Leibeler (KL) divergence, for classification, the pertinent measure is typically the ℓ_1 loss, and other applications use ℓ_2, Hellinger, chi-squared and other losses.

Let $[k]^*$ be the set of finite sequences over $[k]$. A *distribution estimator* is a mapping $q : [k]^* \to \Delta_k$ associating with each observed sample $x^n \in [k]^*$ a distribution $q(x^n) = (q_1(x^n), \ldots, q_k(x^n))$ over $[k]$. The expected loss of q after observing n samples $X^n = X_1, \ldots, X_n$, generated *i.i.d.* according to an unknown distribution $p \in \Delta_k$ is

$$\mathop{\mathbb{E}}_{X^n \sim p} L(p, q(X^n)).$$

The loss of q for the worst distribution is

$$r^{\mathrm{L}}_{k,n}(q) \overset{\mathrm{def}}{=} \max_{p \in \Delta_k} \mathop{\mathbb{E}}_{X^n \sim p} L(p, q(X^n)).$$

We are interested in the least worst-case loss achieved by any estimator, also called the *min-max loss*,

$$r^{\mathrm{L}}_{k,n} \overset{\mathrm{def}}{=} \min_q r^{\mathrm{L}}_{k,n}(q) = \min_q \max_{p \in \Delta_k} \mathop{\mathbb{E}}_{X^n \sim p} L(p, q(X^n)).$$

Determining the min-max loss for a given loss function L, and the optimal estimator achieving it, is of significant practical importance. For example, an estimator with small KL-loss could improve compression and stock-portfolio selection, while an estimator with a small ℓ_1 loss could result in better classification.

Yet as above, very little is know about $r^{\mathrm{L}}_{k,n}$. The only loss function for which $r^{\mathrm{L}}_{k,n}$ has been determined even to the first order is KL-divergence[1],

$$\mathrm{KL}(p, q) \overset{\mathrm{def}}{=} \sum_{i=1}^{k} p_i \log \frac{p_i}{q_i}, \tag{1}$$

where after a sequence of papers, Cover (1972); Krichevsky (1998); Braess et al. (2002); Paninski (2004), just eleven years ago, Braess and Sauer (2004) showed that for fixed k, as n increases,

$$r^{\mathrm{KL}}_{k,n} = \frac{k-1}{2n} + o\left(\frac{1}{n}\right). \tag{2}$$

Even so, their estimator is somewhat impenetrable, and their proof for why their specific estimator works but similar estimators with different parameters do not, relied on automated computer calculations of the behavior of the loss at the boundaries of the simplex.

1.2. Relation to cumulative loss

The scarcity of results is even more surprising as more complex questions have been studied and resolved in much more detail. For example, several researchers in statistics, information theory, and online learning, have studied the more complex *min-max cumulative loss* that minimizes the sum of losses over n successive estimates,

$$R^{\mathrm{L}}_{k,n} \overset{\mathrm{def}}{=} \min_q \max_{p \in \Delta_k} \mathop{\mathbb{E}}_{X^n \sim p} \sum_{j=1}^{n} L(p, q(X^j)).$$

[1] All logarithms in this paper are natural logarithms.

Among the many results on the topic, Krichevsky and Trofimov (1981) showed that for KL loss,

$$R^{\text{KL}}_{k,n} = \frac{k-1}{2}\log n + o(\log n), \tag{3}$$

and a sequence of papers e.g. Xie and Barron (1997); Drmota and Szpankowski (2004) have subsequently determined $R^{\text{KL}}_{k,n}$ up to additive accuracy of $O(1/n)$.

The major difference between the loss $r^{\text{L}}_{k,n}$ and its cumulative counterpart $R^{\text{L}}_{k,n}$ is that $r^{\text{L}}_{k,n}$ minimizes the loss for any given number of observed samples, while $R^{\text{L}}_{k,n}$ minimizes only the sum of the losses. Hence $R^{\text{L}}_{k,n}$ does not provide a direct insight about $r^{\text{L}}_{k,n}$, while $r^{\text{L}}_{k,n}$ provides a clear upper bound on $R^{\text{L}}_{k,n}$. For any loss L, k, and n,

$$R^{\text{L}}_{k,n} \le \sum_{j=1}^{n} r^{\text{L}}_{k,j}.$$

For example, for KL divergence, this relation and (2) implies that for fixed k, as n increases,

$$R^{\text{KL}}_{k,n} \le \sum_{j=1}^{n} r^{\text{KL}}_{k,j} = \frac{k-1}{2}\log n + o(\log n),$$

recovering the behavior of the upper bound in (3)

Another difference between the loss $r^{\text{L}}_{k,n}$ and its cumulative counterpart $R^{\text{L}}_{k,n}$ is that the loss has also the meaning of how well one can learn the distribution from n observations, while the cumulative loss does not carry that meaning. Furthermore, while for KL divergence, $R^{\text{L}}_{k,n}$ can be interpreted as the *redundancy*, the additional number of bits required to represent the whole sequence X^n when the distribution is not known, for other loss functions, the cumulative loss $R^{\text{L}}_{k,n}$ may not have a clear meaning.

A natural question to ask may therefore be why characterizing the simpler quantity $r^{\text{KL}}_{k,n}$ took much longer than for the more complex $R^{\text{KL}}_{k,n}$. As described in the next subsection, the reason may be the simplicity of approaching the optimal cumulative loss.

1.3. Add-constant estimators

Many popular estimators assign to each symbol a probability proportional to its number of occurrences plus a positive constant. Let

$$T_i \stackrel{\text{def}}{=} T_i(X^n) \stackrel{\text{def}}{=} \sum_{j=1}^{n} \mathbb{1}(X_j = i)$$

denote the number of times symbol $i \in [k]$ appeared in a sample X^n. The add-β estimator $q_{+\beta}$ over $[k]$, assigns to symbol i a probability proportional to its number of occurrences plus β, namely,

$$q_i \stackrel{\text{def}}{=} q_i(X^n) \stackrel{\text{def}}{=} q_{+\beta,i}(X^n) \stackrel{\text{def}}{=} \frac{T_i + \beta}{n + k\beta},$$

where as above, we will often abbreviate $q_{+\beta,i}(X^n)$ by $q_i(X^n)$ and even just q_i. Well-known add-constant estimators include the *empirical frequency* estimator q_{+0}, the Laplace estimator q_{+1}, and the *Krichevsky-Trofimov (KT)* estimator $q_{+1/2}$ Krichevsky and Trofimov (1981). The last one, the *add-half* estimator, was the first estimator shown by Krichevsky and Trofimov (1981) to asymptotically achieve the min-max cumulative loss for KL divergence in (3).

Unlike the cumulative loss that is asymptotically achieved by the add-half estimator, as shown in Section 5, for $r^{\text{KL}}_{k,n}$, the asymptotically optimal estimator derived by Braess and Sauer (2004) is much more involved, perhaps explaining the lag in time it took to derive.

1.4. Results

We first consider three important loss functions and determine their loss either exactly or to the first order with correct constant. In Section 2, we consider the ℓ_2^2 distance

$$\ell_2^2(p,q) \stackrel{\text{def}}{=} \sum_{i=1}^{k}(p_i - q_i)^2.$$

Expected ℓ_2^2 distance of add-constant estimators is closely related to the variance of binomial distributions. This property lets us determine the exact min-max loss for ℓ_2^2 distance for every k and n, and in Theorem 3 we show that

$$r^{\ell_2^2}_{k,n} = r^{\ell_2^2}_{k,n}(q_{+\sqrt{n}/k}) = \frac{1-\frac{1}{k}}{(\sqrt{n}+1)^2},$$

and that furthermore, $q_{+\sqrt{n}/k}$ has the same expected ℓ_2^2 divergence for every distribution in Δ_k. Note that unlike min-max cumulative loss for KL-divergence where add-half is nearly optimal, for ℓ_2^2 the optimal min-max loss is achieved by an estimator that adds a constant that depends on the alphabet size k and the number of samples n, and can be arbitrarily large.

Observe also that the ℓ_2^2 loss decreases to 0 with n uniformly over all alphabet sizes k. For the remaining divergences we consider, the rate at which the loss decreases with the sample size n will depend on the alphabet size k.

In Section 3 we consider the chi-squared loss and analyze one of its several forms

$$\chi^2(p,q) \stackrel{\text{def}}{=} \sum_{i=1}^{k}\frac{(p_i - q_i)^2}{q_i}.$$

In Lemmas 4 and 5, we show that

$$\frac{k-1}{n+k+1} - \frac{k(k-1)(\log(n+1)+1)}{4(n+k)(n+k+1)} \le r^{\chi^2}_{k,n} \le \frac{k-1}{n+1},$$

where the upper bound is obtained by the Laplace estimator. In particular, this implies that for any fixed k, as n increases,

$$r^{\chi^2}_{k,n} = \frac{k-1}{n} + O\left(\frac{\log n}{n^2}\right).$$

One of the most important distances in machine learning is

$$\ell_1(p,q) \stackrel{\text{def}}{=} \sum_{i=1}^{k}|p_i - q_i|.$$

It can be easily shown that if distributions can be estimated to ℓ_1 distance δ, then an element can be classified to one of two unknown distributions with error probability that is at most 2δ above that achievable with prior knowledge of the distributions. In Section 4 we consider the ℓ_1 distance. It is part of folklore that $r^{\ell_1}_{k,n} = \Theta(\sqrt{\frac{k-1}{n}})$. In Corollary 9 we determine the first-order behavior, showing that for every fixed k, as n increases,

$$r^{\ell_1}_{k,n} = \sqrt{\frac{2(k-1)}{\pi n}} + O\left(\frac{1}{n^{\frac{3}{4}}}\right).$$

In Section 5 we consider the min-max loss with the commonly-used family of f-divergence loss functions, defined in Csiszár (1967). Let $f : \mathbb{R}_{\geq 0} \mapsto \mathbb{R}$ be convex and satisfy $f(1) = 0$, then

$$D_f(p||q) \overset{\text{def}}{=} \sum_{i=1}^{k} q_i \cdot f\left(\frac{p_i}{q_i}\right) \; .$$

Many important notions of loss are f-divergences. The most important among them are: the relative entropy from $f(x) = x \log x$; the χ^2 divergence from $f(x) = (x-1)^2$; the Hellinger divergence $H(p||q) = \sum_{i=1}^{k} \left(\sqrt{p_i} - \sqrt{q_i}\right)^2$ from $f(x) = (1 - \sqrt{x})^2$; the ℓ_1 distance (or total variation distance) from $f(x) = |x-1|$.

These are of predominant interest in various applications and are frequently the subject of study. For example, of the ten different notions of loss considered in Gibbs and Su (2002), there are only five relevant to distributions on discrete alphabets, four of which are f-divergences, being precisely the four listed above.

We first discuss the difficulty with providing a coherent general formula for all f-divergences and show that the challenge arises from distributions that are close to the boundary of the simplex Δ_k, specifically probability distributions that assign probability roughly $\frac{1}{n}$ to some elements. In Theorem 10 we show that under the common assumption that excludes these extreme distributions and considers only distributions bounded away from the boundary of the simplex, the min-max loss as well as the optimal estimators have a simple form. Let $r^{f}_{k,n}$ denote the min-max f-divergence for all distributions in Δ_k, and let $\hat{r}^{f}_{k,n}(\delta)$ denote the same for distributions in the simplex interior, i.e. satisfying $p_i \geq \delta > 0$, for all i. We show that under a mild smoothness condition on the convex function f, namely for all functions f that are sub-exponential and that are thrice differentiable in a neighborhood of $x = 1$, the asymptotic loss is determined by the second derivative of f at 1,

$$\hat{r}^{f}_{k,n}(\delta) = f''(1) \cdot \frac{k-1}{2n} + o\left(\frac{1}{n}\right).$$

This result provides a simple understanding of the min-max loss for a large family of f-divergences, in a unified fashion.

2. ℓ_2^2 distance

ℓ_2^2 is the simplest loss to analyze as the calculations resemble those for variance. For the empirical-frequency estimator, $T_i \sim B(p_i, n)$, the binomial distribution with parameters p_i and n, hence $\mathbb{E}(T_i) = np_i$ and $V(T_i) = np_i(1-p_i)$. The expected loss under the empirical estimator is therefore

$$\mathbb{E}||p - q(X^n)||_2^2 = \sum_{i=1}^{k} \mathbb{E}\left(\frac{T_i}{n} - p_i\right)^2 = \sum_{i=1}^{k} \frac{V(T_i)}{n^2} = \sum_{i=1}^{k} \frac{p_i \cdot (1-p_i)}{n} = \frac{1 - \sum_{i=1}^{k} p_i^2}{n} \leq \frac{1 - \frac{1}{k}}{n},$$

with equality when all p_i are $1/k$.

Similar calculations show that the min-max optimal estimator is add-$\sqrt{n}/k$, that it improves on empirical frequency only slightly, increasing the denominator from n to $n + 2\sqrt{n} + 1$, and that it has the same loss for each $p \in \Delta_k$. The proofs of Lemmas 1 and 2 below are in Appendix B.

Lemma 1 *For all $k \geq 2$ and $n \geq 1$,*

$$\min_{\beta \geq 0} r^{\ell_2^2}_{k,n}(q_{+\beta}) = r^{\ell_2^2}_{k,n}(q_{+\sqrt{n}/k}) = \frac{1 - \frac{1}{k}}{(\sqrt{n}+1)^2} \ .$$

Furthermore, $q_{+\sqrt{n}/k}$ has the same expected loss for every distribution $p \in \Delta_k$.

We obtain a matching lower bound.

Lemma 2 *For all $k \geq 2$ and $n \geq 1$,*

$$r^{\ell_2^2}_{k,n} \geq \frac{1 - \frac{1}{k}}{(\sqrt{n}+1)^2} \ .$$

The two lemmas exactly determine the min-max ℓ_2^2 loss and show that it is achieved by the add-$\sqrt{n}/k$ estimator.

Theorem 3 *For all $k \geq 2$ and $n \geq 1$,*

$$r^{\ell_2^2}_{k,n} = r^{\ell_2^2}_{k,n}(q_{+\sqrt{n}/k}) = \frac{1 - \frac{1}{k}}{(\sqrt{n}+1)^2} \ .$$

3. χ^2 divergence

We first upper bound the performance of the Laplace estimator, then show that for $n \gg k$ it is near optimal.

Lemma 4 *For every $k \geq 2$ and $n \geq 1$,*

$$r^{\chi^2}_{k,n} \leq r^{\chi^2}_{k,n}(q_{+1}) = \frac{k-1}{n+1}.$$

Proof Rewrite

$$\chi^2(p||q) = \sum_{i=1}^{k} \frac{(p_i - q_i)^2}{q_i} = -1 + \sum_{i=1}^{k} \frac{p_i^2}{q_i}.$$

Then, for the Laplace estimator,

$$\mathbb{E}\big(\chi^2(p||q_{+1}(X^n))\big) = \mathbb{E}\left(-1 + \sum_{i=1}^{k} \frac{p_i^2}{\frac{T_i+1}{n+k}}\right) = -1 + (n+k)\sum_{i=1}^{k} p_i^2 \mathbb{E}\left(\frac{1}{T_i+1}\right).$$

Now,

$$p_i^2 \mathbb{E}\left(\frac{1}{T_i+1}\right) = p_i^2 \sum_{t=0}^{n} \frac{1}{t+1}\binom{n}{t} p_i^t (1-p_i)^{n-t} = \frac{p_i(1-(1-p_i)^{n+1})}{n+1} \leq \frac{p_i}{n+1} \ .$$

Hence,

$$\mathbb{E}\big(\chi^2(p||q_{+1}(X^n))\big) \le -1 + (n+k)\sum_{i=1}^{k} \frac{p_i}{n+1} = \frac{k-1}{n+1}.$$

This bound holds for any distribution p and equality holds if p assigns a probability of 1 to any element, and the lemma follows. ■

We obtain a lower bound on the min-max loss that characterizes the first order term in its behavior for large n. To obtain this lower bound, we use a uniform prior over the distributions in the simplex Δ_k. The calculations are somewhat involved, and presented in Appendix C.

Lemma 5 *For every $k \ge 2$ and $n \ge 1$,*

$$r^{\chi^2}_{k,n} \ge \frac{k-1}{n+k+1} - \frac{k(k-1)\left(\log(n+1)+1\right)}{4(n+k)(n+k+1)}\ .$$

The next corollary follows.

Corollary 6 *As n increases and $k = o(n/\log n)$,*

$$r^{\chi^2}_{k,n} = \frac{k-1}{n} + o\bigg(\frac{k-1}{n}\bigg),$$

and for fixed k,

$$r^{\chi^2}_{k,n} = \frac{k-1}{n} + O\bigg(\frac{\log n}{n^2}\bigg).$$

The simple evaluation of the Laplace estimator in Lemma 4 may lead one to believe that other add-constant estimators will also achieve the $(k-1)/n$ asymptotic min-max loss. This is not the case. We now show that Laplace is the only add-constant estimator achieving this asymptotic behavior.

For simplicity consider the binary alphabet $k = 2$. Let $p = p_1$, $q = q_1$, and $T = T_1$, $\chi^2(p||q) = \frac{p^2}{q} + \frac{(1-p)^2}{1-q} - 1$. The expected divergence of the add-β estimator is therefore

$$\mathbb{E}\big(\chi^2(p||q_{+\beta}(X^n))\big) = \sum_{t=0}^{n} \binom{n}{t} p^t (1-p)^{n-t} \left(\frac{p^2}{\frac{t+\beta}{n+2\beta}} + \frac{(1-p)^2}{\frac{n-t+\beta}{n+2\beta}} - 1\right).$$

It can be shown that this expected loss behaves as $\Theta\left(\frac{1}{n}\right)$. To capture the behavior of n times the expected loss for the tiny probability $p = \frac{z}{n}$, define

$$\begin{aligned}\Phi^\beta(z) &\overset{\text{def}}{=} \lim_{n\to\infty} n \cdot \mathbb{E}\Big(\chi^2\Big(\frac{z}{n}||q_{+\beta}(X^n)\Big)\Big)\\ &= \lim_{n\to\infty} n \sum_{t=0}^{n} \binom{n}{t} \frac{z^t}{n^t}\Big(1-\frac{z}{n}\Big)^{n-t}\left(\frac{\frac{z^2}{n^2}}{\frac{t+\beta}{n+2\beta}} + \frac{\left(1-\frac{z}{n}\right)^2}{\frac{n-t+\beta}{n+2\beta}} - 1\right)\\ &= \sum_{t=0}^{\infty} \frac{e^{-z}z^t}{t!}\left(\beta + t - 2z + \frac{z^2}{t+\beta}\right) = \beta - z + z^2\,\mathbb{E}\left(\frac{1}{Y_z+\beta}\right), \end{aligned} \quad (4)$$

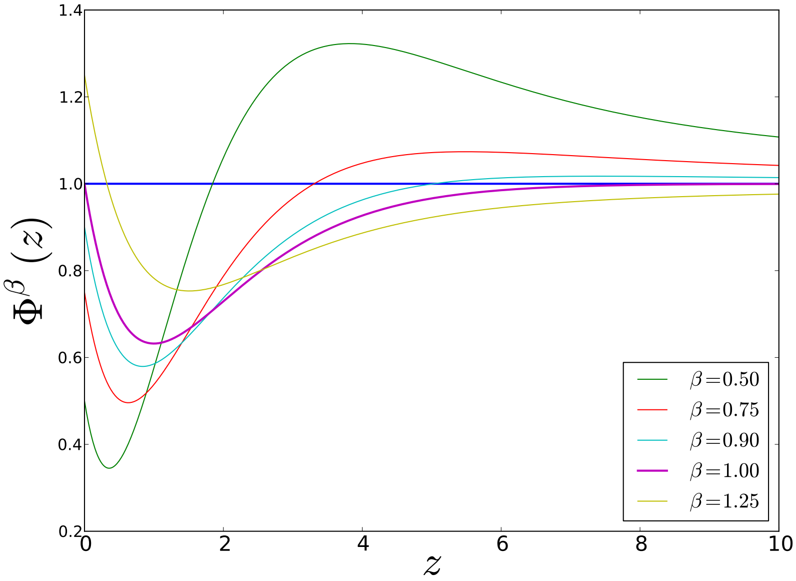

Figure 1: Plots of $\Phi^{\beta}(z)$ from (4) for different choices of β

where $Y_z \sim \text{Poisson}(z)$. It is easy to show that for any fixed $a > 0$, the function of z given by $n \cdot \mathbb{E}\big(\chi^2\big(\frac{z}{n}||q_{+\beta}(X^n)\big)\big)$ converges uniformly to $\Phi^{\beta}(z)$ over $z \in [0, a]$ as $n \to \infty$. We would therefore, at the very least need to have $\Phi^{\beta}(z) \leq 1$, $\forall z \in [0, a]$ so that our proposed estimator would be asymptotically optimal, achieving an expected loss of $\frac{1}{n} + o\left(\frac{1}{n}\right)$. For $\beta = 1$, we can get a closed form expression for $\Phi^{\beta=1}(z) = 1 - ze^{-z}$ which is indeed always bounded above by 1. But we find numerically that for absolutely any other choice of $\beta \neq 1$, we have $\Phi^{\beta}(z) > 1$ for some $z \in [0, 10]$. See Fig. 1 which plots the function $\Phi^{\beta}(z)$ for various choices of β. Thus, the maximum expected loss for the add-β estimator over $p \in [0, 1]$ behaves asymptotically as $\frac{c(\beta)}{n}$ for $\beta > 0$, where

$$c(\beta) \begin{cases} = 1 & \beta = 1, \\ > 1 & \beta \neq 1. \end{cases}$$

Thus, for min-max loss under chi squared loss, the Laplace estimator is uniquely asymptotically optimal among all add-β estimators.

4. ℓ_1 distance

We provide non-asymptotic upper and lower bounds on $r^{\ell_1}_{k,n}$. The expected ℓ_1 distance under the empirical estimator is

$$\mathbb{E}||p - q_{+0}(X^n)||_1 = \sum_{i=1}^{k} \mathbb{E}\left|p_i - \frac{T_i}{n}\right| \leq \sum_{i=1}^{k} \sqrt{\mathbb{E}\left|p_i - \frac{T_i}{n}\right|^2} = \sum_{i=1}^{k} \sqrt{\frac{p_i(1-p_i)}{n}} \leq \sqrt{\frac{k-1}{n}},$$

where the last inequality follows from Cauchy-Schwarz. This is a loose calculation though, and can be improved. If $Z \sim \mathcal{N}(0,1)$, then by the central limit theorem and uniform integrability, we can see that for large n,

$$\sqrt{n}\mathbb{E}\left|p_i - \frac{T_i}{n}\right| \approx \sqrt{p_i(1-p_i)}\mathbb{E}|Z| = \sqrt{\frac{2p_i(1-p_i)}{\pi}}\,. \tag{5}$$

Hence the above analysis loses a constant factor of $\sqrt{\frac{2}{\pi}}$. Interestingly, the expected absolute deviation of a binomial random variable from its mean has a rich history: De Moivre obtained an explicit expression for this quantity, see Diaconis and Zabell (1991) for a historical note on this. We prove our upper bound by providing uniform bounds on this rate of convergence for all p. These uniform bounds we shall prove will also be non-asymptotic in nature, in that, they will hold for every value of k and n.

Lemma 7 *For every $k \geq 2$ and $n \geq 1$,*

$$r^{\ell_1}_{k,n} \leq \sqrt{\frac{2(k-1)}{\pi n}} + \frac{4k^{1/2}(k-1)^{1/4}}{n^{3/4}}\,.$$

Proof We use the empirical estimator to obtain this upper bound. We use the Berry-Esseen theorem to give quantitative bounds on the approximation presented in (5). The details are provided in Appendix D. ■

We also prove a lower bound on the min-max loss.

Lemma 8 *For every $k \geq 2$ and $n \geq 1$,*

$$r^{\ell_1}_{k,n} \geq \sup_{\beta \geq 1} \sqrt{\frac{2(k-1)}{\pi n}}\left(1 - \frac{k}{2(k-1)\beta}\right) - \frac{4k^{\frac{1}{2}}(k-1)^{\frac{1}{4}}}{n^{3/4}} - \frac{k\,(1+k\beta)}{n+k\beta}\,,$$

where the supremum is explicitly attained at

$$\beta^* = \max\left\{1, \frac{n}{(2\pi n(k-1))^{1/4}(nk-n-k)^{1/2} - k}\right\} \sim \max\left\{1, \frac{n^{1/4}}{k^{3/4}}\right\}.$$

Proof The lower bound on min-max loss is proved using Bayes loss for the prior $\mathrm{Dir}(\beta, \beta, \ldots, \beta)$. Using the fact that the median and the mean of a beta distribution are close (Lemma 15 in Appendix D), we obtain the desired lower bound. The full details can be found in Appendix D. ■

We note that the expected loss for the empirical estimator is highest at the uniform distribution and smaller at all other distributions. We also observe that we can get a lower bound that matches the upper bound (up to the first-order term) in the case when the alphabet size k is held fixed when $\beta = n^{\frac{1}{4}}$, thus placing most of the probability mass around the uniform distribution, where the expected loss is the highest. This obtains the correct first order term for all values of $k \geq 2$. Previous results proving lower bounds on ℓ_1 min-max loss such as Han et al. (2014) do not yield the correct first order term.

Corollary 9 *For fixed k, as n increases,*

$$r^{\ell_1}_{k,n} = \sqrt{\frac{2(k-1)}{\pi n}} + O\left(\frac{1}{n^{3/4}}\right).$$

For additional work on density estimation of various non-parametric classes under ℓ_1 loss, see Chan et al. (2013).

5. General family of f-divergences

Based on the results so far, one can ask a natural question: For a fixed alphabet size k, is there a systematic way to understand the asymptotics of min-max loss and asymptotically optimal estimators for all f-divergences as simple properties of the function f? To study this question, let us look at the several popular f-divergences mentioned in the introduction, where the function f is smooth and thrice continuously differentiable, i.e. all but the ℓ_1 loss. As we have done for χ^2, for simplicity, we consider the binary alphabet $\{1, 2\}$, so $k = 2$, and the space of probability distributions can be represented by a single parameter $p = P(X = 2), 0 \le p \le 1$.

$\boldsymbol{\chi^2}$ $f(x) = x^2 - 1$, $\chi^2(p||q) = \frac{p^2}{q} + \frac{(1-p)^2}{(1-q)^2} - 1$. As we saw in Section 3, the Laplace estimator achieves the asymptotic min-max loss, and no other add-constant estimator does.

KL $f(x) = x \log x$, $KL(p, q) = p \log \frac{p}{q} + (1-p)\log\frac{1-p}{1-q}$. As shown by Krichevsky (1998), no add-constant estimator can achieve a natural lower bound of $r^{\text{KL}}_{2,n} \ge \frac{1}{2n} + o\left(\frac{1}{n}\right)$. But, Braess and Sauer (2004) showed that $r^{\text{KL}}_{2,n} = \frac{1}{2n} + o\left(\frac{1}{n}\right)$. and the asymptotically-optimal estimator is a *varying*-add-β estimator described as follows: If a symbol appears exactly r times in n samples, it is assigned a probability that is proportional to $r + \beta_r$ where β_r is a fixed sequence given by

$$\beta_0 = \frac{1}{2}, \qquad \beta_1 = 1, \qquad \beta_2 = \beta_3 = \ldots = \frac{3}{4}.$$

Hellinger $f(x) = (1-\sqrt{x})^2$, $H(p||q) = 2\left(1 - \sqrt{pq} - \sqrt{(1-p)(1-q)}\right)$.

We may perform a Bayes lower bound calculation similar to the lower bound of Lemma 5 presented in Appendix C on Hellinger divergence instead of the χ^2 divergence. This yields the following:

$$r^H_{2,n} = \min_{q(x^n)} \max_{p \in \Delta_2} \mathbb{E} H(p||q(X^n)) \ge \frac{1}{4n} + o\left(\frac{1}{n}\right).$$

(This lower bound may also be seen as a consequence of our Thm. 10 to follow). Suppose we try to obtain a matching upper bound using an add-β estimator. The expected loss when the true distribution parameter is p will be

$$\bar{F}^{\beta}_n(p) \overset{\text{def}}{=} \sum_{t=0}^{n} \binom{n}{t} p^t (1-p)^{n-t}\, 2\left(1 - \sqrt{p \cdot \frac{t+\beta}{n+2\beta}} - \sqrt{(1-p)\cdot\frac{n-t+\beta}{n+2\beta}}\right).$$

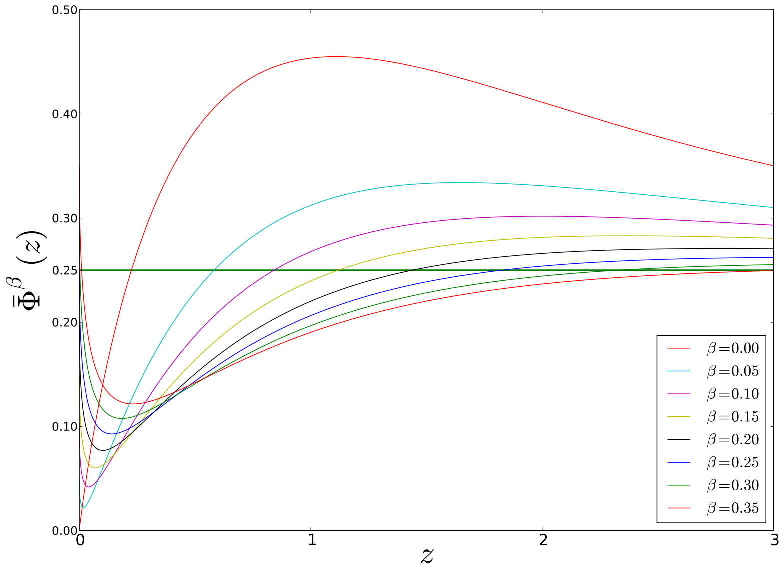

Figure 2: Plots of $\bar{\Phi}^{\beta}(z)$ from (6) for different choices of β. No choice of β yields a curve that's always under the straight line at 0.25.

Similar to (4) in Sec. 3, we define $\bar{\Phi}^{\beta}(z) \overset{\text{def}}{=} \lim_{n\to\infty} n\bar{F}_n^{\beta}\left(\frac{z}{n}\right)$. Thus, $\bar{\Phi}^{\beta}(z)$ captures the behavior of n times the expected loss for the *tiny* probability parameter $p = \frac{z}{n}$. An easy calculation gives

$$\bar{\Phi}^{\beta}(z) = \beta + 2z - 2\sqrt{z}\,\mathbb{E}\left[\sqrt{Y_z + \beta}\right] \,, \tag{6}$$

where $Y_z \sim \text{Poisson}(z)$, and that for any fixed $a > 0$, the function $n\bar{F}_n^{\beta}\left(\frac{z}{n}\right)$ converges uniformly to $\bar{\Phi}^{\beta}(z)$ over $z \in [0, a]$. We would therefore, like $\bar{\Phi}^{\beta}(z) \le \frac{1}{4}$, $\forall z \in [0, a]$ so that our proposed estimator would match the lower bound asymptotically. But numerical calculation shows that no choice of β achieves this goal (see Fig. 2). Thus, for the min-max Hellinger divergence loss, we may have to look for either a) better lower bounds or b) complicated estimators such as the varying-add-β estimators which were proposed by Braess and Sauer (2004) for the KL loss.

The above discussion helps us appreciate the difficulty of giving a coherent answer to the asymptotic min-max loss under an arbitrary f-divergence loss function: the *erratic behavior* of the expected loss at the boundaries of the simplex, a behavior that depends on the function f and the estimator in a complex fashion. We note though that in all three examples above, it is easy to show that the behavior of the expected loss for distributions bounded away from the boundary of the simplex, say $0.1 \le p \le 0.9$ can be shown to match the corresponding lower bounds for any 'reasonable' estimator, in particular any add-β estimator for any $\beta > 0$. Indeed, note from Fig. 1 and Fig. 2 that for any $\beta > 0$, $\Phi^{\beta}(z) \to 1, \bar{\Phi}^{\beta}(z) \to \frac{1}{4}$ as $z \to \infty$ which too is suggestive of being a counterpart upper bound to the lower bounds $r^{\chi^2}_{2,n} \geq \frac{1}{n} + o\left(\frac{1}{n}\right)$ and $r^{H}_{2,n} \geq \frac{1}{4n} + o\left(\frac{1}{n}\right)$. We generalize this observation to our main result in Theorem 10.

Define for any $0 < \delta < \frac{1}{k}$, the δ-bounded simplex $\hat{\Delta}^{\delta}_{k}$ as the set of probability distributions that satisfy $p_i \geq \delta \ \forall i = 1, 2, \ldots, k$. Define the min-max loss for the δ-bounded simplex under an f-divergence loss function:

$$\hat{r}^{f}_{k,n}(\delta) \overset{\text{def}}{=} \min_{q(x^n)} \max_{p \in \hat{\Delta}^{\delta}_{k}} \mathbb{E} D_f(p||q(X^n)). \tag{7}$$

Theorem 10 *Let f be convex and thrice differentiable with $f(1) = 0$ and $f''(1) > 0$. Further, suppose f is sub-exponential, namely $\limsup_{x \to \infty} \frac{|f(x)|}{e^{cx}} = 0, \ \forall c > 0$. Fix any alphabet size k and any $0 < \delta < \frac{1}{k}$. Then,*

$$\hat{r}^{f}_{k,n}(\delta) = \min_{q(x^n)} \max_{p \in \hat{\Delta}^{\delta}_{k}} \mathbb{E} D_f(p||q(X^n)) = \frac{(k-1)f''(1)}{2n} + o\left(\frac{1}{n}\right). \tag{8}$$

The first-order term in the asymptotic behavior is not affected by δ. Furthermore, any add-β estimator with fixed $\beta > 0$ achieves this asymptotic behavior.

Theorem 10 unifies the three examples of f-divergences we looked at in this section by giving us for any $0 < \delta < \frac{1}{2}$:

- $f(x) = x \log x, \qquad f''(1) = 1, \qquad \hat{r}^{\text{KL}}_{2,n}(\delta) = \frac{1}{2n} + o\left(\frac{1}{n}\right)$.
- $f(x) = x^2 - 1, \qquad f''(1) = 2, \qquad \hat{r}^{\chi^2}_{2,n}(\delta) = \frac{1}{n} + o\left(\frac{1}{n}\right)$.
- $f(x) = (1 - \sqrt{x})^2, \qquad f''(1) = \frac{1}{2}, \qquad \hat{r}^{H}_{2,n}(\delta) = \frac{1}{4n} + o\left(\frac{1}{n}\right)$.

Remark 11 *The proof will show that Theorem 10 also holds when f is assumed to be thrice differentiable in an open interval $(1-\theta, 1+\theta)$ for some $\theta > 0$, instead of thrice differentiable everywhere.*

Remark 12 *The sub-exponential assumption in Theorem 10 is necessary only to ensure that add-β estimators are asymptotically optimal. If the sub-exponential assumption is dropped, the asymptotic behavior in (8) still holds, but the asymptotically optimal estimators need to be modified. One such modified estimator may be as follows: We set $q_i = \frac{T_i + \beta}{n + k\beta}$ for $i = 1, 2, \ldots, k$ only if $\frac{T_i}{n} \geq \frac{\delta}{2}$ for all i, and we set $q_i = \frac{1}{k}$ for $i = 1, 2, \ldots, k$, if the stated condition is not true.*

Proof The intuition behind the theorem is as follows. If the true distribution is p, then the empirical distribution of the samples is distributed approximately normally around p within a distance of $O\left(\frac{1}{\sqrt{n}}\right)$. Furthermore, if p is bounded away from the boundaries of the simplex, then this empirical distribution has a finite amount of variance. Any add-β estimator with a fixed $\beta > 0$ moves the empirical distribution around by at most $O\left(\frac{1}{n}\right)$. The empirical distribution itself is moved by the randomness to a distance that is about $\Theta\left(\frac{1}{\sqrt{n}}\right)$. This means that there is not much change in the f-divergence due to the choice of β and a Taylor approximation of $f(x)$ around $x = 1$ captures the behavior of the min-max loss.

The full proof is placed in Appendix E. ■

References

D. Braess and T. Sauer. Bernstein polynomials and learning theory. *Journal of Approximation Theory*, 128:187–206, 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 *Algorithmic Learning Theory: Proceedings of the 13th International Conference ALT , Springer, Heidelberg*, pages 380–394, 2002.

Siu-on Chan, Ilias Diakonikolas, Rocco A. Servedio, and Xiaorui Sun. Learning mixtures of structured distributions over discrete domains. In *Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2013, New Orleans, Louisiana, USA, January 6-8, 2013*, pages 1380–1394, 2013.

T.M. Cover. Admissibility properties of Gilbert's encoding for unknown source probabilities. *Transactions on Information Theory*, 18(1):216–217, January 1972.

I. Csiszár. Information type measures of differences of probability distribution and indirect observations. *Studia Math. Hungarica*, 2:299–318, 1967.

P. Diaconis and S. Zabell. Closed form summation for classical distributions: Variations on a theme of De Moivre. *Statistical Science*, 6(3):284–302, 1991.

M. Drmota and W. Szpankowski. Precise minimax redundancy and regret. *Transactions on Information Theory*, 50(11):2686–2707, November 2004.

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

R. Groeneveld and G. Meeden. The mode, median, and mean inequality. *The American Statistician*, 31(3):120–121, August 1977.

Yanjun Han, Jiantao Jiao, and Tsachy Weissman. Minimax estimation of discrete distributions under l1 loss. *arXiv preprint arXiv:1411.1467*, 2014.

D. Kershaw. Some extensions of W. Gautschi's inequalities for the gamma function. *Mathematics of Computation*, 41(164):607–611, October 1983.

R.E. Krichevsky. The performance of universal encoding. *Transactions on Information Theory*, 44 (1):296–303, January 1998.

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

L. Paninski. Variational minimax estimation of discrete distributions under KL loss. In *Advances in Neural Information Processing Systems (NIPS)*, 2004.

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

Appendix A. Properties of Dirichlet prior

Most of the lower bound arguments in the paper use properties of the Dirichlet prior. A standard argument to lower bound min-max loss is

$$\min_q \max_{p \in \Delta_k} \mathbb{E} L(p, q) \geq \min_q \mathbb{E}_\pi \mathbb{E}_p L(p, q), \tag{9}$$

where π is any prior distribution over probabilities in Δ_k. A useful prior to use that makes the right hand side amenable to analysis is the Dirichlet prior. The Dirichlet prior is a density with positive parameters $\beta^k \stackrel{\text{def}}{=} (\beta_1, , \ldots ,, \beta_k)$ is

$$\text{Dir}_{\beta^k}(p) = \frac{1}{B(\beta^k)} \prod_{i=1}^{k} p_i^{\beta_i - 1},$$

where $B(\beta^k)$ is a normalization factor ensuring that the probabilities integrate to 1. One of the most useful properties of Dirichlet prior is that the posterior distribution upon observing a sequence with types t^k is

$$\text{Dir}_{\beta^k}(p | T^k = t^k) = \text{Dir}_{\beta^k + t^k}(p). \tag{10}$$

Furthermore, for any i,

$$\text{Dir}_{\beta^k}(p_i) = \text{Beta}_{\beta_i, \sum_{j=1, j \neq i}^{k} \beta_j}(p_i). \tag{11}$$

For X distributed as Beta(α, β),

$$\mathbb{E}X = \frac{\alpha}{\alpha + \beta}, \tag{12}$$

$$\mathbb{E}X^2 = \frac{\alpha(\alpha + 1)}{(\alpha + \beta)(\alpha + \beta + 1)}. \tag{13}$$

For lower bounding min-max chi-squared loss, the following moment calculations would be useful.

Lemma 13 *If p is generated from Dir_{1^k} and the type T^k is generated from p, then the following hold.*

$$\begin{aligned}
\mathbb{E}(T_1 + 1) &= \frac{n + k}{k} \\
\mathbb{E}(T_1 + 1)(T_2 + 1) &= \frac{(n + k)(n + k + 1)}{k(k + 1)} \\
\mathbb{E}(T_1 + 1)(T_1 + 2) &= \frac{2(n + k)(n + k + 1)}{k(k + 1)} \\
\mathbb{E}\frac{T_2 + 1}{T_1 + 1} &\leq \log(n + 1) + \frac{3}{2} \\
\mathbb{E}\frac{1}{T_1 + 1} &\leq 1.
\end{aligned}$$

Proof If p is generated from Dir_{1^k} and the type T^k is generated from p, then it can be shown that

$$\Pr(t^k) = \int \mathrm{Dir}_{1^k}(p)p(t^k)dp = \mathrm{Dir}_{1^k}((t/n)^k). \tag{14}$$

Since Dir_{1^k} is same as uniformly sampling from the simplex, we have

$$T^k = (T_1, T_2, \ldots, T_k) \ \sim \text{Uniform on } \{(t_1, t_2, \ldots, t_k) : \sum_{i=1}^{k} t_i = n, t_i \geq 0, \forall i\} .$$

Thus,

$$P(T_1 = a, T_2 = b) = \frac{\binom{n-a-b+k-3}{k-3}}{\binom{n+k-1}{k-1}}, \quad \text{for } 0 \leq a, b, a+b \leq n,$$
$$P(T_1 = a) = \frac{\binom{n-a+k-2}{k-2}}{\binom{n+k-1}{k-1}}, \quad \text{for } 0 \leq a \leq n.$$

By symmetry, we have $\mathbb{E}T_1 = \frac{n}{k}$, so

$$\mathbb{E}(T_1 + 1) = \frac{n+k}{k} \ .$$

Some simple calculations of the combinatorial sums using Mathematica gives:

$$\mathbb{E}(T_1 + 1)(T_1 + 2) = \frac{2(n+k)(n+k+1)}{k(k+1)} \ ,$$
$$\mathbb{E}(T_1 + 1)(T_2 + 1) = \frac{(n+k)(n+k+1)}{k(k+1)} \ .$$

Defining $H_n = \sum_{k=1}^{n} \frac{1}{k}$ and $\gamma = 0.5772...$ as the Euler-Mascheroni constant, and noting that $H_n \leq \log n + \gamma + \frac{1}{2}$ for $n \geq 1$,

$$\begin{aligned}\mathbb{E}\frac{T_2+1}{T_1+1} &= \sum_{a=0}^{n}\sum_{b=0}^{n-i} \frac{\binom{n-a-b+k-3}{k-3}}{\binom{n+k-1}{k-1}} \frac{b+1}{a+1} \\ &= \sum_{a=0}^{n} \frac{\binom{n-a+k-1}{k-1}}{\binom{n+k-1}{k-1}} \frac{1}{a+1} \\ &\leq \sum_{a=0}^{n} \frac{1}{a+1} = H_{n+1} \leq \log(n+1) + \gamma + \frac{1}{2} \leq \log(n+1) + \frac{3}{2},\end{aligned}$$

and since $T_1 \geq 0$,

$$\mathbb{E}\frac{1}{T_1+1} \leq 1.$$

■

Appendix B. Proofs of Lemma 1 and Lemma 2.

First, we present the proof of Lemma 1.

Proof By definition of variance, $\mathbb{E}(X^2) = V(X) + (\mathbb{E}(X))^2$. Hence,

$$\mathbb{E}\left(p_i - \frac{T_i+\beta}{n+\beta k}\right)^2 = \frac{1}{(n+k\beta)^2}\cdot\mathbb{E}\Big(T_i - np_i - \beta(kp_i-1)\Big)^2$$
$$= \frac{1}{(n+k\beta)^2}\cdot\Big(V(T_i) + \beta^2(kp_i-1)^2\Big) = \frac{1}{(n+\beta k)^2}\cdot\Big(np_i(1-p_i) + \beta^2(kp_i-1)^2\Big).$$

The loss of the add-β estimator for a distribution p is therefore,

$$\mathbb{E}||p - q_{+\beta}(X^n)||_2^2 = \sum_{i=1}^k \mathbb{E}\left(p_i - \frac{T_i+\beta}{n+k\beta}\right)^2 = \frac{1}{(n+k\beta)^2}\cdot\left(n - \beta^2 k - (n-\beta^2k^2)\sum_{i=1}^k p_i^2\right).$$

The expected ℓ_2^2 loss of an add-β estimator is therefore determined by just the sum of squares $\sum_{i=1}^k p_i^2$ that ranges from $1/k$ to 1. For $\beta \le \sqrt{n}/k$, the expected loss is maximized when the square sum is $1/k$, and for $\beta \ge \sqrt{n}/k$, when the square sum is 1, yielding

$$r_{k,n}^{\ell_2^2}(q_{+\beta}) = \max_{p\in\Delta_k}\mathbb{E}||p - q_{+\beta}(X^n)||_2^2 = \frac{1}{(n+k\beta)^2}\cdot\begin{cases} n(1-\frac{1}{k}) & \text{for } \beta \le \frac{\sqrt{n}}{k},\\ \beta^2 k(k-1) & \text{for } \beta \ge \frac{\sqrt{n}}{k}.\end{cases}$$

For $\beta \le \sqrt{n}/k$, the expected loss decreases as β increases, and for $\beta \ge \sqrt{n}/k$, it increases as β increases, hence the minimum worst-case loss is achieved for $\beta = \sqrt{n}/k$. Furthermore, $q_{+\sqrt{n}/k}$ has the same expected loss for every underlying distribution p, yielding the lemma. ■

Now, we present the proof of Lemma 2.

Proof For any prior π over distributions in Δ_k,

$$r_{k,n}^{\mathrm{L}} = \min_{q\in\Delta_k}\max_{p\in\Delta_k}\mathop{\mathbb{E}}_{X^n\sim p} L(p, q(X^n)) \ge \min_{q\in\Delta_k}\mathop{\mathbb{E}}_{P\sim\pi, X^n\sim P} L(P, q(X^n)).$$

For every random variable X, $\mathbb{E}(X-x)^2$ is minimized by $x = \mathbb{E}(X)$. Similarly, for every random pair (X,Y), given Y, $\mathbb{E}(X - x(Y))^2$ is minimized by $x(y) = \mathbb{E}(X|y)$. Hence for ℓ_2^2 loss, for every prior π the right-hand-side above,

$$\min_{q\in\Delta_k}\mathop{\mathbb{E}}_{P\sim\pi, X^n\sim P}||P - q(X^n)||_2^2 = \min_{q\in\Delta_k}\mathop{\mathbb{E}}_{P\sim\pi, X^n\sim P}\sum_{i=1}^k (P_i - q_i(X^n))^2,$$

is minimized by the estimator q^* that assigns to each symbol i its expected probability

$$q_i^*(x^n) = \mathop{\mathbb{E}}_{P\sim\pi, X^n\sim P}(P_i|x^n).$$

As described in equations (10), (11) and (12) in Appendix A, for the Dirichlet prior with parameter $\beta^k = (\beta, \ldots, \beta)$, upon observing x^n of type $t^k = t_1, \ldots, t_k$, the posterior distribution of P is

$\mathrm{Dir}_{\beta^k+t^k}$, hence

$$\begin{aligned} q_i^*(x^n) &= \mathop{\mathbb{E}}_{P\sim\mathrm{Dir}_{\beta^k}, X^n\sim P}(P_i|x^n) = \mathop{\mathbb{E}}_{P\sim\mathrm{Dir}_{\beta^k}, X^n\sim P}(P_i|T^k = t^k) \\ &= \mathop{\mathbb{E}}_{P\sim\mathrm{Dir}_{\beta^k+t^k}}(P_i) = \frac{\beta+t_i}{k\beta+\sum_{j=1}^k t_i} = \frac{\beta+t_i}{k\beta+n}. \end{aligned}$$

Namely, with Dir_{β^k} prior, the expected loss is minimized by $q_{+\beta}$. If we take $\beta = \sqrt{n}/k$, then the expected loss is minimized by $q_{+\sqrt{n}/k}$, and Lemma 1 showed that the resulting loss is the same, $\frac{1-\frac{1}{k}}{(\sqrt{n}+1)^2}$, for all distributions in Δ_k, and the lemma follows. ■

Appendix C. Proof of Lemma 5

The non-asymptotic upper bound $r^{\chi^2}_{k,n} \le \frac{k-1}{n+1}$ was shown using the Laplace estimator in Section 3.

As stated in Appendix A, to get lower bounds on $r^{\chi^2}_{k,n}$, we use the fact that the Bayes loss for the uniform prior is a lower bound on the min-max loss. Let $\pi = \mathrm{Dir}_{1^k}$. For every estimator q,

$$\begin{aligned} &\max_{p\in\Delta_k}\mathbb{E}_p\left(-1+\sum_{i=1}^k\frac{p_i^2}{q_i}\right) \\ &\overset{(a)}{\ge} \mathbb{E}_\pi\mathbb{E}_p\left(-1+\sum_{i=1}^k\frac{P_i^2}{q_i}\right) && (15) \\ &\overset{(b)}{=} \mathbb{E}_\pi\mathbb{E}_p\mathbb{E}\left(-1+\sum_{i=1}^k\frac{P_i^2}{q_i}\middle| X^n\right) && (16) \\ &\overset{(c)}{=} \mathbb{E}_\pi\mathbb{E}_p\left[-1+\sum_{i=1}^k\frac{\frac{(T_i+1)(T_i+2)}{(n+k)(n+k+1)}}{q_i}\right] && (17) \\ &\overset{(d)}{\ge} \mathbb{E}_\pi\mathbb{E}_p\left[-1+\sum_{i=1}^k\frac{\frac{(T_i+1)(T_i+2)}{(n+k)(n+k+1)}}{\frac{\sqrt{(T_i+1)(T_i+2)}}{\sum_{r=1}^k\sqrt{(T_r+1)(T_r+2)}}}\right] && (18) \\ &= -1+\frac{\mathbb{E}_\pi\mathbb{E}_p\sum_{i=1}^k\sum_{r=1}^k\sqrt{(T_i+1)(T_i+2)}\sqrt{(T_r+1)(T_r+2)}}{(n+k)(n+k+1)} && (19) \\ &\overset{(e)}{=} -1+\frac{\left(k\mathbb{E}(T_1+1)(T_1+2)+k(k-1)\mathbb{E}\sqrt{(T_1+1)(T_1+2)(T_2+1)(T_2+2)}\right)}{(n+k)(n+k+1)}, && (20) \end{aligned}$$

where (a) follows from Equation (9), (b) follows from Tower law of expectation, (c) follows from Equations (10), (11), and (13), (d) follows from a simple Lagrange multiplier calculation which shows that under the constraint $\{q_i \ge 0, i = 1, 2, \ldots, k, \sum_{i=1}^k q_i = 1\}$, the quantity $\sum_{i=1}^k \frac{a_i}{q_i}$ is minimized by $q_i \propto \sqrt{a_i}$, (e) follows from symmetry which implies that T_i for $i = 1, 2, \ldots, k$, have the same distribution, and (T_i, T_r) for $i, r = 1, 2, \ldots, k, i \ne r$ also have the same distribution.

Now, using the Taylor approximation $\sqrt{1+x} \geq 1 + \frac{x}{2} - \frac{x^2}{8}$ for $x \geq 0$,

$$\begin{aligned}
&\mathbb{E}\sqrt{(T_1+1)(T_1+2)(T_2+1)(T_2+2)} \\
=&\mathbb{E}(T_1+1)(T_2+1)\sqrt{1+\frac{1}{T_1+1}}\sqrt{1+\frac{1}{T_2+1}} \\
\geq&\mathbb{E}(T_1+1)(T_2+1)\left(1+\frac{1}{2(T_1+1)}-\frac{1}{8(T_1+1)^2}\right)\left(1+\frac{1}{2(T_2+1)}-\frac{1}{8(T_2+1)^2}\right) \\
=&\mathbb{E}(T_1+1)(T_2+1)+\mathbb{E}(T_1+1)+\frac{1}{4}-\frac{1}{4}\mathbb{E}\frac{T_2+1}{T_1+1}-\frac{1}{8}\mathbb{E}\frac{1}{T_1+1}+\frac{1}{64}\mathbb{E}\frac{1}{(T_1+1)(T_2+1)} \\
\geq&\mathbb{E}(T_1+1)(T_2+1)+\mathbb{E}(T_1+1)+\frac{1}{4}-\frac{1}{4}\mathbb{E}\frac{T_2+1}{T_1+1}-\frac{1}{8}\mathbb{E}\frac{1}{T_1+1}\ , \qquad (21)
\end{aligned}$$

where the last equality used symmetry again: the distribution of T_1 and T_2 are the same, as are the distributions of (T_1, T_2) and (T_2, T_1).

Substituting results from Lemma 13 and Equation (21) to bound chi-squared loss,

$$\begin{aligned}
r^{\chi^2}_{k,n} \geq& -1 + \frac{1}{(n+k)(n+k+1)}\Bigg[k\cdot\frac{2(n+k)(n+k+1)}{k(k+1)} \\
&+k(k-1)\cdot\left(\frac{(n+k)(n+k+1)}{k(k+1)}+\frac{n+k}{k}+\frac{1}{4}-\frac{1}{4}(\log(n+1)+\frac{3}{2})-\frac{1}{8}\right)\Bigg] \\
=&\frac{k-1}{n+k+1}-\frac{k(k-1)\left(\log(n+1)+1\right)}{4(n+k)(n+k+1)}.
\end{aligned}$$

Appendix D. Proof of Lemmas 7 and 8

In this section, we provide the proofs of lemmas 7 and 8 together.

First, we present the following lemma that establishes a non-asymptotic bound on the rate of convergence of expected absolute deviations of a binomial random variable from its mean to the corresponding expected absolute value of a Gaussian random variable.

Lemma 14 *Let* $T \sim \text{Binomial}(n, p)$. *For* $\alpha, \beta \geq 0$, *we have*

$$\left|\mathbb{E}\left|p-\frac{T+\alpha}{n+\alpha+\beta}\right|-\sqrt{\frac{2p(1-p)}{\pi n}}\right| \leq \frac{4p^{1/4}(1-p)^{1/4}}{n^{3/4}}+\frac{\alpha+\beta}{n+\alpha+\beta}.$$

Proof Let $X \sim \text{Bernoulli}(p)$. Then, $\mathbb{E}X = p, \sigma^2 \stackrel{\text{def}}{=} E\left(X_1-p\right)^2 = p(1-p)$ and $\rho \stackrel{\text{def}}{=} E\left|X_1-p\right|^3 = p(1-p)(1-2p+2p^2)$. Let $Y = \frac{T-np}{\sigma\sqrt{n}}$ and let $Z \sim \mathcal{N}(0,1)$.

By the Berry-Esseen theorem, we have

$$|\Pr(|Y| \leq t) - \Pr(|Z| \leq t)| \leq \frac{2\rho}{\sigma^3\sqrt{n}}.$$

Since $\mathbb{E}Y^2 = \mathbb{E}Z^2 = 1$, we have for any $R > 0$,

$$\mathbb{E}|Y|1_{|Y|\geq R} \leq \mathbb{E}\frac{|Y|^2}{R}1_{|Y|\geq R} \leq \frac{\mathbb{E}Y^2}{R} = \frac{1}{R},$$

and similarly $\mathbb{E}|Z|1_{|Z|\geq R} \leq \frac{1}{R}$. Thus,

$$\begin{aligned}
&|\mathbb{E}|Y| - \mathbb{E}|Z||\\
=&\left|\int_0^\infty \Pr(|Y| \geq t) - \Pr(|Z| \geq t)dt\right|\\
\leq&\int_0^R |\Pr(|Y| \geq t) - \Pr(|Z| \geq t)|dt + \int_R^\infty \Pr(|Y| \geq t) + \Pr(|Z| \geq t)dt\\
\leq&\frac{2\rho}{\sigma^3\sqrt{n}}R + \mathbb{E}|Y|1_{|Y|\geq R} + \mathbb{E}|Z|1_{|Z|\geq R}\\
\leq&2\left(\frac{\rho}{\sigma^3\sqrt{n}}R + \frac{1}{R}\right).
\end{aligned}$$

We optimize the upper bound by choosing $R = \sqrt{\frac{\sigma^3\sqrt{n}}{\rho}}$. Thus, we get

$$|\mathbb{E}|Y| - \mathbb{E}|Z|| \leq 4\sqrt{\frac{\rho}{\sigma^3\sqrt{n}}}\,.$$

Multiplying both sides by $\frac{\sigma}{\sqrt{n}}$, we get

$$\begin{aligned}
\left|\mathbb{E}\left|p - \frac{T}{n}\right| - \frac{\sigma}{\sqrt{n}}\mathbb{E}|Z|\right| &\leq 4\sqrt{\frac{\rho}{\sigma^3\sqrt{n}}}\frac{\sigma}{\sqrt{n}}\\
&= \frac{4p^{1/4}(1-p)^{1/4}(1-2p+2p^2)^{1/2}}{n^{3/4}}\\
&\leq \frac{4p^{1/4}(1-p)^{1/4}}{n^{3/4}}\,.
\end{aligned}$$

Using $\mathbb{E}|Z| = \sqrt{\frac{2}{\pi}}$ completes the proof for the case $\alpha = \beta = 0$. To complete the proof for general $\alpha, \beta \geq 0$, note that

$$\begin{aligned}
\left|p - \frac{T+\alpha}{n+\alpha+\beta}\right| &\leq \left|p - \frac{T}{n}\right| + \left|\frac{T}{n} - \frac{T+\alpha}{n+\alpha+\beta}\right|\\
&= \left|p - \frac{T}{n}\right| + \left|\frac{T\beta + (T-n)\alpha}{n(n+\alpha+\beta)}\right|\\
&\leq \left|p - \frac{T}{n}\right| + \frac{\alpha+\beta}{n+\alpha+\beta}\,.
\end{aligned}$$

■

Using Lemma 14 with $\alpha, \beta = 0$, we have

$$\sum_{i=1}^{k} \mathbb{E}\left|p_i - \frac{T_i}{n}\right| \leq \sum_{i=1}^{k} \sqrt{\frac{2p_i(1-p_i)}{\pi n}} + \frac{4p_i^{1/4}(1-p_i)^{1/4}}{n^{3/4}}$$
$$\leq \sqrt{\frac{2(k-1)}{\pi n}} + \frac{4k^{1/2}(k-1)^{1/4}}{n^{3/4}},$$

where the last inequality follows by observing that the uniform distribution maximizes both terms. This shows that the empirical estimator achieves the performance leading to the desired upper bound on $r_{k,n}^{\ell_1}$. This completes the proof of Lemma 7.

To get the lower bound, we bound the min-max loss by the Bayes loss. Choose a Bayesian prior on P as $\mathrm{Dir}(\beta, \beta, \ldots, \beta)$ where $\beta \geq 1$ may be chosen later depending on n, k, so $\beta = \beta(n, k) \geq 1$. Then, the conditional law $P|(T_1, T_2, \ldots, T_k)$ works out to:

$$P|(T_1 = t_1, \ldots, T_k = t_k) \sim \mathrm{Dir}(t_1 + \beta, t_2 + \beta, \ldots, t_k + \beta).$$

Note that this means $P_i|(T_1 = t_1, \ldots, T_k = t_k) \sim \mathrm{Beta}(t_i + \beta, n - t_i + (k-1)\beta)$. We will use the following two lemmas:

Lemma 15 *Groeneveld and Meeden (1977) For $\alpha, \beta > 1$, the median of the $\mathrm{Beta}(\alpha, \beta)$ distribution is sandwiched between the mean $\frac{\alpha}{\alpha+\beta}$ and the mode $\frac{\alpha-1}{\alpha+\beta-2}$ so that the distance between the mean and median is at most $\frac{|\beta-\alpha|}{(\alpha+\beta)(\alpha+\beta-2)}$.*

Lemma 16 *Kershaw (1983) For $y \geq 2$, $0 < r \leq 1$,*

$$\left(y - \frac{1}{2}\right)^r \leq \frac{\Gamma(y+r)}{\Gamma(y)} \leq y^r.$$

The Bayes loss can then be lower bounded in the following way:

$$\begin{aligned}
&\min_q \mathbb{E}\sum_{i=1}^k |P_i - q_i(T_1, T_2, \ldots, T_k)| \\
&\geq \sum_{i=1}^k \min_q \mathbb{E}|P_i - q_i(T_1, T_2, \ldots, T_k)| \qquad \text{(where } q_i \text{ now need not add up to 1)} \\
&= \sum_{i=1}^k \min_q \mathbb{E}\left[\mathbb{E}[|P_i - q_i(T_1, T_2, \ldots, T_k)|\,|\, T_1, T_2, \ldots, T_k]\right] \\
&\overset{(a)}{=} \sum_{i=1}^k \min_q \mathbb{E}\left[\mathbb{E}[|P_i - \text{Median}(\text{Beta}(\beta + T_i, (k-1)\beta + n - T_i))|\,|\, T_1, T_2, \ldots, T_k]\right] \\
&= \sum_{i=1}^k \mathbb{E}\left[|P_i - \text{Median}(\text{Beta}(\beta + T_i, (k-1)\beta + n - T_i))|\right] \\
&\overset{(b)}{\geq} \sum_{i=1}^k \left[\mathbb{E}\left|P_i - \frac{T_i + \beta}{n + k\beta}\right| - \left|\frac{(k-2)\beta + n}{(n+k\beta)(n+k\beta-2)}\right|\right] \\
&\overset{(c)}{\geq} \sum_{i=1}^k \left[\sqrt{\frac{2}{\pi n}}\mathbb{E}P_i^{1/2}(1-P_i)^{1/2} - \frac{4}{n^{3/4}}\mathbb{E}P_i^{1/4}(1-P_i)^{1/4} - \frac{k\beta}{n+k\beta} - \frac{(k-2)\beta + n}{(n+k\beta)(n+k\beta-2)}\right] \\
&\geq \sum_{i=1}^k \left[\sqrt{\frac{2}{\pi n}}\mathbb{E}P_i^{1/2}(1-P_i)^{1/2} - \frac{4}{n^{3/4}}\mathbb{E}P_i^{1/4}(1-P_i)^{1/4} - \frac{k\beta}{n+k\beta} - \frac{1}{n+k\beta}\right],
\end{aligned}$$

where (a) follows because the optimal q_i is the median of the posterior distribution $P_i \sim \text{Beta}(\beta + T_i, (k-1)\beta + n - T_i)$, (b) follows from Lemma 15, and (c) follows from Lemma 14.

Now, using the fact that $P_i \sim \text{Beta}(\beta, (k-1)\beta)$, we have for $r = \frac{1}{2}, \frac{1}{4}$,

$$\mathbb{E}P_i^r(1-P_i)^r = \frac{\Gamma(k\beta)}{\Gamma(\beta)\Gamma((k-1)\beta)} \cdot \frac{\Gamma(\beta + r)\Gamma((k-1)\beta + r)}{\Gamma(k\beta + 2r)},$$

and using Lemma 16, we obtain

$$\begin{aligned}
\mathbb{E}P_i^{\frac{1}{2}}(1-P_i)^{\frac{1}{2}} &\geq \frac{(k-1)^{\frac{1}{2}}}{k}\left(1 - \frac{1}{2\beta}\right)^{\frac{1}{2}}\left(1 - \frac{1}{2(k-1)\beta}\right)^{\frac{1}{2}}, \\
\mathbb{E}P_i^{\frac{1}{4}}(1-P_i)^{\frac{1}{4}} &\leq \frac{(k-1)^{\frac{1}{4}}}{k^{\frac{1}{2}}} \cdot \frac{1}{\left(1 - \frac{1}{2k\beta}\right)^{\frac{1}{2}}}.
\end{aligned}$$

$$
\begin{aligned}
&\min_q \mathbb{E}\sum_{i=1}^{k} |P_i - q_i(T_1, T_2, \ldots, T_k)| \\
\geq & \sqrt{\frac{2(k-1)}{\pi n}} \left(1 - \frac{1}{2\beta}\right)^{\frac{1}{2}} \left(1 - \frac{1}{2(k-1)\beta}\right)^{\frac{1}{2}} - \frac{4k^{\frac{1}{2}}(k-1)^{\frac{1}{4}}}{n^{3/4}} \cdot \frac{1}{\left(1 - \frac{1}{2k\beta}\right)^{\frac{1}{2}}} \\
& - \frac{k(1+k\beta)}{n+k\beta} \\
\geq & \sqrt{\frac{2(k-1)}{\pi n}} \left(1 - \frac{1}{2\beta}\right)^{\frac{1}{2}} \left(1 - \frac{1}{2(k-1)\beta}\right)^{\frac{1}{2}} - \frac{4k^{\frac{1}{2}}(k-1)^{\frac{1}{4}}}{n^{3/4}} \cdot \frac{1}{\left(1 - \frac{1}{2k\beta}\right)^{\frac{1}{2}}} \\
& - \frac{k(1+k\beta)}{n+k\beta} \\
\geq & \sqrt{\frac{2(k-1)}{\pi n}} \left(1 - \frac{1}{2\beta}\right) \left(1 - \frac{1}{2(k-1)\beta}\right) - \frac{4k^{\frac{1}{2}}(k-1)^{\frac{1}{4}}}{n^{3/4}} \\
& - \frac{k(1+k\beta)}{n+k\beta} \qquad\qquad (\text{using } \sqrt{1-x} \geq 1 - x, 0 \leq x \leq 1) \\
\geq & \sqrt{\frac{2(k-1)}{\pi n}} \left(1 - \frac{k}{2(k-1)\beta}\right) - \frac{4k^{\frac{1}{2}}(k-1)^{\frac{1}{4}}}{n^{3/4}} - \frac{k\,(1+k\beta)}{n+k\beta}\,.
\end{aligned}
$$

This proves the desired lower bound for any chosen $\beta \geq 1$. Since $\beta \geq 1$ is arbitrary, this completes the proof of the lower bound in Lemma 8.

Appendix E. Proof of Theorem 10

In the two subsections that follow, we prove an asymptotic upper bound and the matching asymptotic lower bound respectively on $\hat{r}^{f}_{k,n}(\delta)$.

E.1. Asymptotic upper bound on f-divergence loss

Fix any $\beta > 0$ and consider any add-β estimator: $q_i = \frac{T_i+\beta}{n+k\beta}$, $i = 1, 2, \ldots, k$. The expected f-divergence loss multiplied by n is then:

$$
n\mathbb{E}\sum_{i=1}^{k} \frac{T_i+\beta}{n+k\beta} f\left(\frac{p_i(n+k\beta)}{T_i+\beta}\right) = \frac{1}{1+\frac{k\beta}{n}} \sum_{i=1}^{k} \mathbb{E}(T_i+\beta) f\left(\frac{p_i(n+k\beta)}{T_i+\beta}\right).
$$

Let us fix any $\epsilon > 0$ satisfying $\epsilon \leq \frac{\delta}{2}$. By Hoeffding's inequality, we have

$$
P(|T_i - np_i| > \epsilon n) \leq 2e^{-2\epsilon^2 n}.
$$

We evaluate the quantity $\mathbb{E}(T_i+\beta) f\left(\frac{p_i(n+k\beta)}{T_i+\beta}\right)$ by breaking the space into two regions $\{|T_i - np_i| \leq \epsilon n\}$ and $\{|T_i - np_i| > \epsilon n\}$. Over the latter, the absolute value of the contribution of the expectation is upper bounded as

$$\begin{aligned}
&\left|\mathbb{E}(T_i+\beta) f\left(\frac{p_i(n+k\beta)}{T_i+\beta}\right) 1_{|T_i-np_i|>\epsilon n}\right| \\
\leq &(n+\beta) \max\left\{\left|f\left(\frac{p_i(n+k\beta)}{\beta}\right)\right|, \left|f\left(\frac{p_i(n+k\beta)}{n+\beta}\right)\right|\right\} \cdot 2e^{-2\epsilon^2 n} \\
&\qquad \text{(as f is convex, the maximum is attained at $T_i = 0$ or $T_i = n$)} \\
\leq &(n+\beta) \max\left\{\left|f\left(\frac{n+k\beta}{\beta}\right)\right|, \left|f\left(\frac{\delta(n+k\beta)}{n+\beta}\right)\right|\right\} \cdot 2e^{-2\epsilon^2 n}, \\
\leq &(n+\beta) \max\left\{\left|f\left(\frac{n+k\beta}{\beta}\right)\right|, |f(\delta)|\right\} \cdot 2e^{-2\epsilon^2 n},
\end{aligned}$$

which vanishes, using the sub-exponential property of f. Note that the upper bound has no dependence on p and converges uniformly to zero over all p satisfying $p_i \geq \delta$, $\forall i$.

Now, to estimate the contribution of the expectation over $\{|T_i - np_i| \leq \epsilon n\}$, define $g(x) \overset{\text{def}}{=} f\left(\frac{1}{1+x}\right)$ so that

$$\begin{aligned}
g(x) &= g(0) + g'(0)x + \frac{g''(0)}{2}x^2 + \frac{g'''(y)}{6}x^3, \quad \text{for some } y, |y| \leq |x|, \\
&= g(0) + g'(0)x + \frac{g''(0)}{2}x^2 \pm \frac{M(x)}{6}x^3, \quad \text{where } M(x) := \sup_{y:|y|\leq|x|} |g'''(y)|,
\end{aligned}$$

where the notation $a = b \pm c$ will mean that a is sandwiched between $b - c$ and $b + c$, i.e. $b - c \leq a \leq b + c$.

We observe

$$\begin{aligned}
g(x) &= f\left(\frac{1}{1+x}\right), & g(0) &= f(1) = 0, \\
g'(x) &= \frac{-1}{(1+x)^2} f'\left(\frac{1}{1+x}\right), & g'(0) &= -f'(1), \\
g''(x) &= \frac{2}{(1+x)^3} f'\left(\frac{1}{1+x}\right) + \frac{1}{(1+x)^4} f''\left(\frac{1}{1+x}\right), & g''(0) &= 2f'(1) + f''(1), \\
g'''(x) &= -\frac{6}{(1+x)^4} f'\left(\frac{1}{1+x}\right) - \frac{6}{(1+x)^5} f''\left(\frac{1}{1+x}\right) \\
&\quad - \frac{1}{(1+x)^6} f'''\left(\frac{1}{1+x}\right).
\end{aligned}$$

$$
\begin{aligned}
&\mathbb{E}(T_i+\beta) f\left(\frac{p_i(n+k\beta)}{T_i+\beta}\right) 1_{|T_i-np_i|\le \epsilon n}\\
=&\mathbb{E}(T_i+\beta) g\left(\frac{(T_i-np_i)+\beta(1-kp_i)}{p_i(n+k\beta)}\right) 1_{|T_i-np_i|\le \epsilon n}\\
=&\mathbb{E}(T_i+\beta)\left[g'(0)\frac{(T_i-np_i)+\beta(1-kp_i)}{p_i(n+k\beta)}+\frac{g''(0)}{2}\left(\frac{(T_i-np_i)+\beta(1-kp_i)}{p_i(n+k\beta)}\right)^2\right.\\
&\qquad\left.\pm\frac{M\left(\frac{1.1\epsilon}{\delta}\right)}{6}\left|\frac{(T_i-np_i)+\beta(1-kp_i)}{p_i(n+k\beta)}\right|^3\right] 1_{|T_i-np_i|\le \epsilon n}\\
&\left(\text{since for all } n\ge \tfrac{10\beta}{\epsilon}, \text{ we have } \left|\tfrac{(T_i-np_i)+\beta(1-kp_i)}{p_i(n+k\beta)}\right|\le \tfrac{1.1\epsilon}{\delta} \text{ over } \{|T_i-np_i|\le \epsilon n\}.\right)
\end{aligned}
$$

Now, we bound individual terms by using the following standard moments of the binomial distribution:

$$
\mathbb{E}T_i-np_i=0,\qquad \mathbb{E}(T_i-np_i)^2=np_i(1-p_i),\qquad \mathbb{E}(T_i-np_i)^3=np_i(1-p_i)(1-2p_i).
$$

The first term evaluates to:

$$
\begin{aligned}
&\mathbb{E}(T_i+\beta)\frac{(T_i-np_i)+\beta(1-kp_i)}{p_i(n+k\beta)}1_{|T_i-np_i|\le \epsilon n}\\
=\ &\mathbb{E}\left((T_i-np_i)+(np_i+\beta)\right)\frac{(T_i-np_i)+\beta(1-kp_i)}{p_i(n+k\beta)}(1-1_{|T_i-np_i|>\epsilon n})\\
=\ &\mathbb{E}\frac{(T_i-np_i)^2+(np_i+\beta(2-kp_i))(T_i-np_i)+(np_i+\beta)\beta(1-kp_i)}{p_i(n+k\beta)}(1-1_{|T_i-np_i|>\epsilon n})\\
=\ &\frac{np_i(1-p_i)+0+(np_i+\beta)\beta(1-kp_i)}{p_i(n+k\beta)}\\
&\qquad\pm\frac{n^2+(n+\beta(2+k))\,n+(n+\beta)\beta(1+k)}{\delta(n+k\beta)}\cdot 2e^{-2\epsilon^2 n}
\end{aligned}
$$

The second term evaluates to:

$$
\begin{aligned}
&\mathbb{E}(T_i+\beta)\left(\frac{(T_i-np_i)+\beta(1-kp_i)}{p_i(n+k\beta)}\right)^2 1_{|T_i-np_i|\le \epsilon n}\\
=\ &\mathbb{E}\left((T_i-np_i)+(np_i+\beta)\right)\frac{(T_i-np_i)^2+2\beta(1-kp_i)(T_i-np_i)+\beta^2(1-kp_i)^2}{p_i^2(n+k\beta)^2}1_{|T_i-np_i|\le \epsilon n}\\
=\ &\mathbb{E}\frac{1-1_{|T_i-np_i|>\epsilon n}}{p_i^2(n+k\beta)^2}\left[(T_i-np_i)^3+(np_i+\beta(3-2kp_i))(T_i-np_i)^2\right.\\
&\qquad\left.+\left(2\beta(1-kp_i)(np_i+\beta)+\beta^2(1-kp_i)^2\right)(T_i-np_i)+(np_i+\beta)\beta^2(1-kp_i)^2\right]\\
=\ &\frac{np_i(1-p_i)(1-2p_i)+(np_i+\beta(3-2kp_i))\,np_i(1-p_i)+0+(np_i+\beta)\beta^2(1-kp_i)^2}{p_i^2(n+k\beta)^2}\\
&\pm\frac{n^3+(n+\beta(3+2k))n^2+(2\beta(1+k)(n+\beta)+\beta^2(1+k)^2)n+(n+\beta)\beta^2(1+k)^2}{\delta^2(n+k\beta)^2}\cdot 2e^{-2\epsilon^2 n}
\end{aligned}
$$

The third term may be bounded as:

$$\begin{aligned}&\left|\mathbb{E}(T_i+\beta)\left|\frac{(T_i-np_i)+\beta(1-kp_i)}{p_i(n+k\beta)}\right|^3 1_{|T_i-np_i|\le\epsilon n}\right|\\ \le\;& (n+\beta)\mathbb{E}\frac{|T_i-np|^3+3|T_i-np|^2\beta(1-kp_i)+3\beta^2(1-kp_i)^2|T_i-np_i|+\beta^3|1-kp_i|^3}{p_i^3(n+k\beta)^3}.\end{aligned}$$

By the central limit theorem and uniform integrability, we have $\mathbb{E}\left|\frac{T_i-np_i}{\sqrt{np_i(1-p_i)}}\right|^i \to \mathbb{E}|Z|^i$ for $i=1,2,3$, where $Z\sim\mathcal{N}(0,1)$. Thus, the third term vanishes in n at a rate that can be bounded (using Berry-Eseen theorem) in terms of only k,ϵ,δ,β without dependence on p_i, only using that $p_i\ge\delta$ $i=1,2,\ldots,k$.

Thus, the limit as $n\to\infty$, of n times the expected f-divergence loss for the add-β estimator is:

$$\begin{aligned}&\lim_{n\to\infty} n\mathbb{E}\sum_{i=1}^{k}\frac{T_i+\beta}{n+k\beta}f\left(\frac{p_i(n+k\beta)}{T_i+\beta}\right)\\ =&\sum_{i=1}^{k}g'(0)(1-p_i+\beta(1-kp_i))+\frac{g''(0)}{2}(1-p_i)\\ =&g'(0)(k-1)+\frac{g''(0)}{2}(k-1)\\ =&-f'(1)(k-1)+\frac{2f'(1)+f''(1)}{2}(k-1)\\ =&\frac{(k-1)f''(1)}{2},\end{aligned}$$

independent of p and $\beta>0$. This proves $\hat{r}^f_{k,n}(\delta)\le\frac{(k-1)f''(1)}{2n}+o\left(\frac{1}{n}\right)$.

E.2. Asymptotic lower bound on f-divergence loss

In the previous subsection, we provided an upper bound on the expected loss for a large class of f-divergences. We now prove a matching lower bound for this class. Note that we don't need the sub-exponential assumption for the lower bound.

The proof consists of two parts. We first strengthen the known lower-bound proof for min-max loss under ℓ_2^2 loss by showing that essentially the same asymptotic lower bound holds even if we restrict the distributions to a very small subset of the k-simplex Δ_k. We then use this result to prove a tight lower bound for more general f-divergences.

The common technique for lower-bounding expected loss assumes a prior on the collection of distributions, and lower-bounds the expected loss over this prior. ℓ_2^2 loss has the convenient property that for any prior Π over the possible distributions, the estimator minimizing the expected ℓ_2^2 loss is exactly the mean of the posterior distribution given the observations.

The Dirichlet prior is particularly convenient as the posterior distribution is also of Dirichlet form, and the optimal estimator is an add-constant estimator. For our purpose, the simplest form of the Dirichlet prior, the uniform distribution, will suffice. Since our distributions are restricted to a subset of the simplex, we truncate the uniform distribution to a ball, and approximate the posterior by the add-constant estimator by showing that posterior distribution for full Dirichlet prior does not assign much probability outside the truncation with high probability. Following this, we use the add-constant estimator to lower bound the expected ℓ_2^2 loss between the optimal estimator and a set of distributions in the ball by $\frac{1-\frac{1}{k}}{n} - o\big(\frac{1}{n}\big)$. Then we show that this set of distributions has near full probability under the truncated prior. And therefore the lower bound of $\frac{1-\frac{1}{k}}{n} - o\big(\frac{1}{n}\big)$ also holds for the entire ball.

Finally, we relate the loss under general f-divergence with ℓ_2^2 loss by taking a Taylor series expansion. And then we use the lower bound for ℓ_2^2 loss to obtain a lower bound for the general f-divergence.

We explain these steps in much broader detail below.

We restrict $\mathcal{P}$ to just distributions close to $(\frac{1}{k}\frac{1}{k}, \ldots, \frac{1}{k})$. For $\epsilon > 0$, consider the L_∞ ball of radius ϵ around $(1/k, \ldots, 1/k)$,

$$B_k(\epsilon) \overset{\text{def}}{=} \{p \in \Delta_k : ||p - (1/k, \ldots, 1/k)||_\infty < \epsilon\} = \left\{p \in \Delta_k : \left|p_i - \frac{1}{k}\right| < \epsilon \text{ for all } i\right\}.$$

We will use the following nested balls,

$$\begin{aligned} B_k^1 &\overset{\text{def}}{=} B_k\left(\frac{1}{n^{1/5}}\right), \\ B_k^2 &\overset{\text{def}}{=} B_k\left(\frac{1}{n^{1/5}} - \frac{3\log n}{\sqrt{n}}\right), \\ B_k^3 &\overset{\text{def}}{=} B_k\left(\frac{1}{n^{1/5}} - \frac{5\log n}{\sqrt{n}}\right), \\ B_k^4 &\overset{\text{def}}{=} B_k\left(\frac{k}{n^{1/5}}\right). \end{aligned}$$

Denote the min-max f-divergence loss for distributions in B_k^1 by

$$\hat{r}^f_{B_k^1,n} \overset{\text{def}}{=} \min_{q(x^n)} \max_{p \in B_k^1} \mathop{\mathbb{E}}_{X^n \sim p} D_f(p||q(X^n)).$$

E.2.1. LOWER BOUND FOR THE ℓ_2^2-LOSS

First, we state the general result of how to calculate optimal estimator that minimizes expected ℓ_2^2 loss under a prior. Let $\mathcal{P}$ be a collection of distributions over a set $\mathcal{X}$ and let Π be a prior over $\mathcal{P}$.

Given an observation x^n the posterior distribution over $\mathcal{P}$ is

$$\Pi(p|x^n) \overset{\text{def}}{=} \frac{\Pi(p) \cdot p(x^n)}{\int_{p' \in \mathcal{P}} \Pi(p') \cdot p'(x^n) dp'}.$$

$\Pi(p|x^n)$ in turn determines a posterior distribution over $\mathcal{X}$,

$$\hat{p}_i(x^n) \overset{\text{def}}{=} \int_{p \in \mathcal{P}} \Pi(p|x^n) \cdot p_i dp\ .$$

The following is well-known.

Lemma 17 *Let $\mathcal{P}$ be a collection of distributions over $[k]$ and let Π be a prior over $\mathcal{P}$, then for every observed x^n, $\mathbb{E}_\Pi \mathbb{E}_{X^n} \ell_2^2(p, q(X^n))$ is minimized by $\hat{p}(x^n)$.*

To lower bound $\hat{r}^{\ell_2^2}_{B_k^1,n}$, we use a uniform prior Π^1 over distributions in B_k^1. Let $t_i(x^n)$ be the number of times symbol i appears in x^n. After observing x^n, the posterior distribution of p is

$$\Pi^1(p|x^n) \propto \begin{cases} \prod_{i=1}^k p_i^{t_i(x^n)}, & p \in B_k^1, \\ 0, & \text{otherwise}, \end{cases}$$

and the posterior distribution over $\mathcal{X}$ is

$$\hat{p}_i^1(x^n) \stackrel{\text{def}}{=} \int_{p \in B_k^1} \Pi^1(p|x^n) \cdot p_i \, dp.$$

An explicit expression for $\hat{p}^1$ is hard to find, so instead, we show that $\hat{p}^1$ is very close to optimal estimator $\hat{p}$ for uniform prior over full simplex Δ_k. Then we use $\hat{p}$ to bound the expected ℓ_2^2 loss.

Consider the uniform prior Π over Δ_k. Then the posterior is the Dirichlet distribution,

$$\Pi(p|x^n) = \Gamma(n+k) \cdot \prod_{i=1}^k \frac{p_i^{t_i(x^n)}}{\Gamma(t_i(x^n)+1)}. \tag{22}$$

Therefore the estimator minimizing the expected ℓ_2^2 loss under the prior Π is

$$\begin{aligned} \hat{p}_i(x^n) &= \int_{p \in \Delta_k} \Gamma(n+k) \cdot \prod_{i=1}^k \frac{p_i^{t_i(x^n)}}{\Gamma(t_i(x^n)+1)} p_i dp \\ &= \frac{\Gamma(n+k)\Gamma(t_i(x^n)+2)}{\Gamma(n+k+1)\Gamma(t_i(x^n)+1)} \\ &= \frac{t_i(x^n)+1}{n+k} \\ &= \frac{t_i(x^n)}{n} + O\left(\frac{1}{n}\right). \end{aligned}$$

Let

$$nB_k^2 \stackrel{\text{def}}{=} \{n \cdot p : p \in B_k^2\},$$

then $x^n \in nB_k^2$ iff its type $\left(\frac{t_1(x^n)}{n}, \ldots, \frac{t_k(x^n)}{n}\right) \in B_k^2$. We show that for $x^n \in nB_k^2$, the posterior for prior Π^1 is not much higher than the posterior for Π.

Lemma 18 *For all sufficiently large n, for every $x^n \in nB_k^2$,*

$$\Pi(B_k^1|x^n) = 1 - o\left(\frac{1}{n}\right).$$

Proof

To bound the probability outside B_k^1, observe that since $x^n \in nB_k^2$ and $\left|\frac{t_i(x^n)}{n} - \frac{t_i(x^n)}{n+k-2}\right| \le \frac{\log n}{\sqrt{n}}$ $\forall i$,

$$
\begin{aligned}
\Pi\Big(\Big|p_i - \frac{k}{n}\Big| > \frac{1}{n^{1/5}}\Big| x^n\Big) &\le \Pi\Big(\Big|p_i - \frac{t_i(x^n)}{n}\Big| > \frac{\log n}{\sqrt{n}}\Big| x^n\Big)\\
&\le \Pi\Big(\Big|p_i - \frac{t_i(x^n)}{n+k-2}\Big| > \frac{2\log n}{\sqrt{n}}\Big| x^n\Big)\\
&= \int_0^{\frac{t_i(x^n)}{n+k-2}-\frac{2\log n}{\sqrt{n}}} \Pi(p_i|x^n)dp_i + \int_{\frac{t_i(x^n)}{n+k-2}+\frac{2\log n}{\sqrt{n}}}^1 \Pi(p_i|x^n)dp_i.
\end{aligned}
$$

From integrating (22),

$$
\Pi(p_i|x^n) = \frac{(n+k-1)!}{t_i(x^n)!(n-t_i(x^n)+k-2)!} p_i^{t_i(x^n)}(1-p_i)^{n-t_i(x^n)+k-2}.
$$

Now, observe that

$$
\begin{aligned}
p_i^t(1-p_i)^{n-t+k-2} &= \left(\frac{t}{n+k-2}\right)^t\left(1-\frac{t}{n+k-2}\right)^{n-t+k-2} e^{-(n+k-2)D_{KL}\left(\frac{t}{n+k-2}||p_i\right)}\\
&\le \left(\frac{t}{n+k-2}\right)^t\left(1-\frac{t}{n+k-2}\right)^{n-t+k-2} e^{-2(n+k-2)\left(\frac{t}{n+k-2}-p_i\right)^2}\\
&\le \left(\frac{t}{n+k-2}\right)^t\left(1-\frac{t}{n+k-2}\right)^{n-t+k-2} e^{-2n\left(\frac{t}{n+k-2}-p_i\right)^2}
\end{aligned}
$$

where the first inequality follows from Pinsker's inequality. Therefore, using Stirling's approximation and bounding Gaussian tail probability,

$$
\begin{aligned}
\int_0^{\frac{t}{n+k-2}-\frac{2\log n}{\sqrt{n}}} p_i^t(1-p_i)^{n-t}dp_i &\le \left(\frac{t}{n+k-2}\right)^t\left(1-\frac{t}{n+k-2}\right)^{n-t+k-2}\int_{\frac{2\log n}{\sqrt{n}}}^{\infty} e^{-2nx^2}dx\\
&\le \frac{t!(n-t+k-2)!}{(n+k-2)!}\frac{n^3}{4\log n\sqrt{n}}e^{-8\log^2 n}\\
&\le \frac{t!(n-t+k-2)!}{(n+k-1)!}\frac{1}{n^4}.
\end{aligned}
$$

Hence,

$$
\int_0^{\frac{t_i(x^n)}{n+k-2}-\frac{2\log n}{\sqrt{n}}} \Pi(p_i|x^n)dp_i \le \frac{1}{n^4}.
$$

Similarly,

$$
\int_{\frac{t_i(x^n)}{n+k-2}+\frac{2\log n}{\sqrt{n}}}^1 \Pi(p_i|x^n)dp_i \le \frac{1}{n^4}.
$$

Therefore $\forall i$,

$$
\Pi\Big(\Big|p_i - \frac{k}{n}\Big| > \frac{1}{n^{1/5}}\Big| x^n\Big) \le \frac{2}{n^4}.
$$

Using the union bound,

$$
\Pi\big(\Delta_k/B_k^1|x^n\big) \le \frac{2k}{n^4} \le \frac{1}{n^3}.
$$

■

Lemma 19 *For all sufficiently large n, for every $x^n \in nB_k^2$, for all $p \in B_k^1$,*

$$\Pi(p|x^n) = \Pi^1(p|x^n) \cdot \left(1 + o\left(\frac{1}{n}\right)\right).$$

Proof Since for $p \in B_k^1$, $\Pi(p|x^n)$ and $\Pi^1(p|x^n)$ have same expressions except for the normalizing constant, from Lemma 18 it follows that for $p \in B_k^1$,

$$\Pi(p|x^n) = \Pi^1(p|x^n) \cdot \left(1 + o\left(\frac{1}{n}\right)\right).$$

■

Next we show that $\hat{p}(x|x^n)$ and $\hat{p}^1(x|x^n)$ are close for $x^n \in nB_k^2$.

Lemma 20 *For all sufficiently large n, for $x^n \in nB_k^2$,*

$$\hat{p}_i^1(x^n) = \hat{p}_i(x^n) + o\left(\frac{1}{n}\right).$$

Proof From Lemmas 18 and 19,

$$\begin{aligned}
\hat{p}_i^1(x^n) - \hat{p}_i(x^n) &= \int_{p \in \Delta_k} (\Pi^1(p|x^n) - \Pi(p|x^n))p_i dp - \int_{p \notin \Delta_k} \Pi(p|x^n) p_i dp \\
&= \int_{p \in \Delta_k} \Pi^1(p|x^n) o\left(\frac{1}{n}\right) p_i dp + o\left(\frac{1}{n}\right) \\
&= \hat{p}_i^1(x^n) o\left(\frac{1}{n}\right) + o\left(\frac{1}{n}\right) \\
&= o\left(\frac{1}{n}\right).
\end{aligned}$$

■

We now lower bound the expected ℓ_2^2 loss between $\hat{p}^1(X^n)$ and p for any $p \in B_k^3$.

Lemma 21 *For $p \in B_k^3$,*

$$\mathbb{E}_{X^n} \ell_2^2(p, \hat{p}^1) \geq \frac{1 - \frac{1}{k}}{n} - o\left(\frac{1}{n}\right).$$

Proof First we show that x^n generated according to any $p \in B_k^3$ will belong to nB_k^2 with high probability. Using Union and Chernoff bounds, for $p \in B_k^3$,

$$\begin{aligned}
p(nB_k^2) &\geq 1 - k \cdot p(|t_i(x^n) - np_i| \geq 2 \log n \sqrt{n}) \\
&\geq 1 - 2ke^{-4 \log^2 n} \\
&\geq 1 - \frac{1}{n^3}.
\end{aligned}$$

Next using the above result and Lemma 20, we lower bound the expected ℓ_2^2 loss between $\hat{p}^1$ and $p \in B_k^3$.

$$
\begin{aligned}
\mathbb{E}\left(\sum_{i=1}^{k}(p_i-\hat{p}_i^1)^2\right) &\geq \mathbb{E}\sum_{i=1}^{k}\big((p_i-\hat{p}_i^1)^2 \big| \, X^n \in nB_k^2\big)p(nB_k^2)\\
&= \mathbb{E}\sum_{i=1}^{k}\left(\left(p_i-\hat{p}_i-o\left(\frac{1}{n}\right)\right)^2 \middle| \, X^n \in nB_k^2\right)p(nB_k^2)\\
&= \mathbb{E}\sum_{i=1}^{k}\left(\left(p_i-\hat{p}_i-o\left(\frac{1}{n}\right)\right)^2\right)-o\left(\frac{1}{n}\right)\\
&= \mathbb{E}\sum_{i=1}^{k}\left(\left(p_i-\frac{T_i}{n}-O\left(\frac{1}{n}\right)\right)^2\right)-o\left(\frac{1}{n}\right)\\
&\geq \mathbb{E}\sum_{i=1}^{k}\left(\left(p_i-\frac{T_i}{n}\right)^2-O\left(\frac{1}{n}\right)\left|p_i-\frac{T_i}{n}\right|\right)-o\left(\frac{1}{n}\right)\\
&= \sum_{i=1}^{k}p_i(1-p_i)-o\left(\frac{1}{n}\right)\\
&= \left(\frac{1}{k}+o\left(\frac{1}{n}\right)\right)\sum_{i=1}^{k}(1-p_i)-o\left(\frac{1}{n}\right)\\
&= \frac{1-\frac{1}{k}}{n}-o\left(\frac{1}{n}\right).
\end{aligned}
$$

■

Now, we show that Π^1 assigns most of the probability to B_k^3.

Lemma 22

$$\Pi^1(B_k^3) \geq 1-o(1).$$

Proof We relate $\Pi^1(p_i)$ to the volume of a set. Then using a property of these kind of sets, we show that $\Pi^1(p_i)$ decreases with increasing $\left|p_i-\frac{1}{k}\right|$. Thereby, we show that $\Pi^1(p_i)$ is concentrated around $\frac{1}{k}$. Then using the union bound, we lower bound $\Pi^1(B_k^3)$. First, we define the set and prove a property of volume of these sets. Let

$$
\begin{aligned}
S_k^a(\delta) &\overset{\text{def}}{=} \{(l_1,\ldots,l_k) : \forall i \, |l_i-a| \leq \epsilon \text{ and} \sum_{i=1}^{k} l_i = \delta\},\\
V_k^a(\delta) &\overset{\text{def}}{=} \text{Volume of } S_k^a.
\end{aligned}
$$

Because of symmetry, $V_k^a(ka-\gamma) = V_k^a(ka+\gamma)$ for any γ. We then show that $V_k^a(\delta)$ increases in the range $[0, ka]$ and decreases in the range $[ka, \infty)$. It is easy to see that the claim is true for $k = 1$. We now prove that the claim is true for k assuming that the claim is true for $k-1$. For

$\delta_1 < \delta_2 < ak$,

$$\begin{aligned}
V_k^a(\delta_1) - V_k^a(\delta_2) &= \int_{a-\epsilon}^{a+\epsilon} \big(V_{k-1}^a(\delta_1 - l_1) - V_{k-1}^a(\delta_2 - l_1)\big) dl_1 \\
&= \int_{a-\epsilon}^{\delta_1+\delta_2-a+\epsilon-2(k-1)a} \big(V_{k-1}^a(2(k-1)a - (\delta_1 - l_1)) - V_{k-1}^a(\delta_2 - l_1)\big) dl_1 \\
&+ \int_{\delta_1+\delta_2-a+\epsilon-2(k-1)a}^{a+\epsilon} \big(V_{k-1}^a(\delta_1 - l_1) - V_{k-1}^a(\delta_2 - l_1)\big) dl_1 \\
&\overset{(b)}{\le} \int_{a-\epsilon}^{\delta_1+\delta_2-a+\epsilon-2(k-1)a} \big(V_{k-1}^a(2(k-1)a - (\delta_1 - l_1)) - V_{k-1}^a(\delta_2 - l_1)\big) dl_1 \\
&= \int_{2(k-1)a+a-\epsilon-\delta_1}^{\delta_2-a+\epsilon} V_{k-1}^a(x) dx - \int_{2(k-1)a+a-\epsilon-\delta_1}^{\delta_2-a+\epsilon} V_{k-1}^a(x) dx \\
&= 0.
\end{aligned}$$

where (b) step follows since for $l_1 \in [\delta_1+\delta_2-a+\epsilon-2(k-1)a, a+\epsilon]$, $\delta_1 - l_1 \le \delta_2 - l_1 \le (k-1)a$.

Therefore for any k, a and $|ka - \delta_1| \ge |ka - \delta_2|$,

$$V_k^a(\delta_1) \le V_k^a(\delta_2).$$

Now we relate $\Pi^1(p_i)$ to volume of a set.

$$\Pi^1(p_i) \propto V_{k-1}^{1/k}(1-p_i) \text{ if } \left|p_i - \frac{1}{k}\right| \le \frac{1}{n^{1/5}}.$$

Therefore $\Pi^1(p_i)$ decreases with increasing $|p_i - \frac{1}{k}|$. Therefore Π^1 assigns a higher probability to B_k^3.

$$\begin{aligned}
\Pi^1(B_k^3) &\ge 1 - k\Pi^1\left(\left|p_i - \frac{1}{k}\right| \ge \frac{1}{n^{1/5}} - \frac{5\log n}{\sqrt{n}}\right) \\
&\ge 1 - \frac{5k\log n}{n^{3/10}} \\
&= 1 - o(1).
\end{aligned}$$

■

The following theorem follows.

Theorem 23

$$\hat{r}_{B_k^1,n}^{\ell_2^2} \ge \frac{1-\frac{1}{k}}{n} - o\left(\frac{1}{n}\right).$$

Proof Using Lemmas 21,22,

$$\begin{aligned}\hat{r}^{\ell_2^2}_{B_k^1,n} &\geq \mathbb{E}_{\Pi^1}\mathbb{E}_{X^n}\ell_2^2(p,\hat{p}^1)\\ &\geq \mathbb{E}_{\Pi^1}(\mathbb{E}_{X^n}\ell_2^2(p,\hat{p}^1)|p\in B_k^3)\Pi^1(p\in B_k^3)\\ &\geq \mathbb{E}_{\Pi^1}\left(\frac{1-\frac{1}{k}}{n}-o\left(\frac{1}{n}\right)\middle| p\in B_k^3\right)(1-o(1))\\ &= \frac{1-\frac{1}{k}}{n}-o\left(\frac{1}{n}\right).\end{aligned}$$

■

E.2.2. LOWER BOUND FOR GENERAL f-DIVERGENCE

As previously, we consider that p is from family of distributions B_k^1. We first show that any optimal estimator $q(X^n)$ will assign a distribution from B_k^4 because of convexity of f. Then we relate the loss under general f-divergence to ℓ_2^2 loss and lower bound loss under general f-divergence using the result we proved in previous subsection.

We first prove that the distance between p and q decreases as we move q closer to p.

Lemma 24 *For $p_1 > q_1, p_2 < q_2$ and $d \leq \min(p_1-q_1, q_2-p_2)$,*

$$q_1 f\left(\frac{p_1}{q_1}\right)+q_2 f\left(\frac{p_2}{q_2}\right)\geq (q_1+d)f\left(\frac{p_1}{q_1+d}\right)+(q_2-d)f\left(\frac{p_2}{q_2-d}\right).$$

Proof Let

$$g(y)=(q_1+y)f\left(\frac{p_1}{q_1+y}\right)+(q_2-y)f\left(\frac{p_2}{q_2-y}\right).$$

Then we show that $g'(y)\leq 0\,\forall\, 0\leq y\leq d$ from which the result follows.

$$g'(y)=f\left(\frac{p_1}{q_1+y}\right)-\frac{p_1}{q_1+y}f'\left(\frac{p_1}{q_1+y}\right)-\left(f\left(\frac{p_2}{q_2-y}\right)-\frac{p_2}{q_2-y}f'\left(\frac{p_2}{q_2-y}\right)\right).$$

Now let $h(x) = f(x) - xf'(x)$. We can see that $h(x)$ is a decreasing function since $h'(x) = -f''(x) \leq 0$.

Since $\frac{p_1}{q_1+y}\geq 1\geq \frac{p_2}{q_2-y}$, $g'(y)\leq 0 \quad \forall 0\leq y\leq d$. Hence for some $z\in[0,d]$,

$$g(d)-g(0)=g'(z)d\leq 0.$$

■

Now, using above lemma we show that optimal q is always from B_k^4.

Lemma 25 *The optimal $q(X^n)$ that minimizes $\max_{p\in B_k^1}\mathbb{E}_{X^n}D_f(p||q(X^n))$ will always be from B_k^4.*

$$argmin_{q(X^n)}\max_{p\in B_k^1}\mathbb{E}_{X^n}D_f(p||q(X^n))\in B_k^4.$$

Proof We first prove that an optimal estimator q exists such that for any x^n, either $q_i(x^n) \geq \frac{1}{k} - \frac{1}{n^{1/5}} \,\forall i$ or $q_i(x^n) \leq \frac{1}{k} + \frac{1}{n^{1/5}} \,\forall i$.

Suppose for some x^n, $q_i(x^n) > \frac{1}{k} + \frac{1}{n^{1/5}}$ and $q_j(x^n) < \frac{1}{k} - \frac{1}{n^{1/5}}$. Then we will show that for any $p \in B_k^1$, we decrease $D_f(p||q(x^n))$ by pushing $q_i(x^n)$ and $q_j(x^n)$ closer to the boundaries of interval $\left[\frac{1}{k} - \frac{1}{n^{1/5}}, \frac{1}{k} + \frac{1}{n^{1/5}}\right]$. Let $d = \min\left(q_i(x^n) - \left(\frac{1}{k} + \frac{1}{n^{1/5}}\right), \frac{1}{k} - \frac{1}{n^{1/5}} - q_j(x^n)\right)$. Then consider $q'(X^n)$ which is same as $q(X^n)$ for all $X^n \neq x^n$. And for x^n, $q'_k(x^n) = q_k(x^n) \,\forall k \neq i, j$ and $q'_i(x^n) = q_i(x^n) - d$, $q'_j(x^n) = q_j(x^n) + d$.

Using Lemma 24, we can show that for any $p \in B_k^1$,

$$q'_i(x^n) f\left(\frac{p_i}{q'_i(x^n)}\right) + q'_j(x^n) f\left(\frac{p_j}{q'_j(x^n)}\right) \leq q_i(x^n) f\left(\frac{p_i}{q_i(x^n)}\right) + q_j(x^n) f\left(\frac{p_j}{q_j(x^n)}\right).$$

And therefore,

$$D_f(p||q'(x^n)) \leq D_f(p||q(x^n)).$$

And similarly, we can do same process until there exists any l, m such that $q_l(x^n) > \frac{1}{k} + \frac{1}{n^{1/5}}$ and $q_m(x^n) < \frac{1}{k} - \frac{1}{n^{1/5}}$. And hence in the end, for any x^n, either $q_i(x^n) \geq \frac{1}{k} - \frac{1}{n^{1/5}} \,\forall i$ or $q_i(x^n) \leq \frac{1}{k} + \frac{1}{n^{1/5}} \,\forall i$.

If for x^n, $q_i(x^n) \geq \frac{1}{k} - \frac{1}{n^{1/5}} \,\forall i$, then $q_i(x^n) \leq \frac{1}{k} + \frac{k}{n^{1/5}} \,\forall i$. Similarly the other way. Therefore, for all x^n, $q(x^n) \in B_k^4$. ■

We now express $D_f(p||q)$ in terms of $\ell_2^2(p, q)$ for any $p, q \in B_k^4$.

Lemma 26 *For any $p, q \in B_k^4$,*

$$D_f(p||q) = (f''(1) + o(1))\frac{k}{2}\ell_2^2(p, q).$$

Proof

We take a taylor series expansion of $D_f(p||q)$. For some $\alpha_i \in \left[\frac{n^{1/5}-1}{n^{1/5}+1}, \frac{n^{1/5}+1}{n^{1/5}-1}\right]$,

$$\begin{aligned}
D_f(p||q) &= \sum_{i=1}^{k} q_i f(\frac{p_i}{q_i}) \\
&= \sum_{i=1}^{k} \left(q_i \left(\frac{p_i}{q_i} - 1\right) f'(1) + \frac{q_i}{2}\left(\frac{p_i - q_i}{q_i}\right)^2 f''(1) + \frac{q_i}{3!}\left(\frac{p_i - q_i}{q_i}\right)^3 f'''(\alpha_i) \right) \\
&= \sum_{i=1}^{k} \left(\frac{(p_i - q_i)^2}{2q_i} f''(1) + \frac{(p_i - q_i)^3}{6q_i^2} f'''(\alpha_i) \right).
\end{aligned}$$

Let $M = \max_{x \in \left[\frac{n^{1/5}-1}{n^{1/5}+1}, \frac{n^{1/5}+1}{n^{1/5}-1}\right]} f'''(x)$. Then,

$$\begin{aligned}
\left| D_f(p||q) - f''(1) \sum_{i=1}^{k} \frac{(p_i - q_i)^2}{2q_i} \right| &\le \sum_{i=1}^{k} |p_i - q_i| \frac{(p_i - q_i)^2}{6q_i^2} f'''(\alpha_i) \\
&\le \sum_{i=1}^{k} \frac{2}{n^{1/5}} M k^2 \frac{(p_i - q_i)^2}{2q_i} \\
&= \frac{2Mk^2}{n^{1/5}} \sum_{i=1}^{k} \frac{(p_i - q_i)^2}{2q_i} \\
&= o(1) \sum_{i=1}^{k} \frac{(p_i - q_i)^2}{2q_i}.
\end{aligned}$$

Therefore, for any f-divergence,

$$\begin{aligned}
D_f(p||q) &= (f''(1) + o(1)) \sum_{i=1}^{k} \frac{(p_i - q_i)^2}{2q_i} \\
&= (f''(1) + o(1)) \sum_{i=1}^{k} \frac{k(p_i - q_i)^2}{2kq_i} \\
&= (f''(1) + o(1)) \frac{k}{2} \sum_{i=1}^{k} (p_i - q_i)^2 (1 + o(1)) \\
&= (f''(1) + o(1)) \frac{k}{2} \ell_2^2(p, q).
\end{aligned}$$

$\blacksquare$

Now, we lower-bound $\hat{r}^{f}_{B_k^1, n}$ using the relationship between $D_f(p||q)$ and $\ell_2^2(p, q)$ and the fact that for optimal q, $q(X^n) \in B_k^4$.

Theorem 27

$$\hat{r}^{f}_{B_k^1, n} \ge \frac{f''(1)(k-1)}{2n} - o\left(\frac{1}{n}\right).$$

Proof From Lemma 26 and Theorem 23,

$$\begin{aligned}
\hat{r}^{f}_{B_k^1, n} &= \min_{q(x^n)} \max_{p \in B_k^1} \mathbb{E} D_f(p||q(X^n)) \\
&= \min_{q(x^n)} \max_{p \in B_k^1} \mathbb{E} (f''(1) + o(1)) \frac{k}{2} \ell_2^2(p, q(X^n)) \\
&= (f''(1) + o(1)) \frac{k}{2} \hat{r}^{\ell_2^2}_{B_k^1, n} \\
&\ge \frac{f''(1)(k-1)}{2n} - o\left(\frac{1}{n}\right).
\end{aligned}$$

■

Since, for any $0 \leq \delta < \frac{1}{k}$, $\delta < \frac{1}{k} - \frac{1}{n^{1/5}}$ for sufficiently high n,

$$\hat{r}^f_{k,n}(\delta) \geq \hat{r}^f_{B^1_k,n} \geq \frac{f''(1)(k-1)}{2n} - o\left(\frac{1}{n}\right).$$