On Competitive Distribution Estimation

Nadim Ghaddar, Vaishakh Ravindrakumar

September 12, 2018

Abstract

In this report, we discuss previous work on probability distribution estimation in the competitive sense, i.e. relative to an oracle that essentially knows the underlying distribution. Two natural limits on the oracle’s power are considered; in the first case, the oracle is assumed to know the distribution up to a permutation over the alphabet, and in the second, the oracle is assumed to know the exact underlying distribution but is restricted to assign the same probabilities for symbols that appear the same number of times in the sample. In both cases, it is shown that there exists an estimator that achieves a competitive regret of $\tilde{O}(\min(k/n, 1/\sqrt{n}))$, and that for the latter case, such a competitive regret is inevitable.

1 Introduction

Estimating an unknown distribution from samples that are drawn from the distribution is a fundamental problem in many statistics and machine learning applications. Probability estimation has thus been extensively researched for long. Yet, estimators which are practical and provably-optimal have remained elusive. The paper considered in this report [8] concerns itself with finding such estimators.

Let $p$ denote the underlying distribution on a discrete alphabet $[k] \triangleq \{1, 2, ..., k\}$. Further, $p$ is assumed to belong to a class of distributions $\mathcal{P}$ and through the rest of the report we assume $\mathcal{P} = \Delta_k$ or the simplex in $k$ dimensions. Samples $X_1, X_2, ..., X_n$ are then drawn from $p$, allowing the construction of an estimator $q_{X^n}$, also over $[k]$.

The performance of the estimator $q_{X^n}$ is evaluated via the popular KL divergence

$$D(p||q_{X^n}) = \sum_{i\in[k]} p(i) \log \frac{p(i)}{q_{X^n}(i)}.$$ 

Let $r_n(q, p) \triangleq \mathbb{E}_{X^n \sim p^n}[D(p||q_{X^n})]$. The min-max loss $r_n(\mathcal{P})$ under this setting is defined as

$$r_n(\mathcal{P}) = \min_q \max_p r_n(q, p).$$

For various regimes of $(k, n)$, add-constant estimators are known to be essentially min-max optimal. For instance, for $\frac{k}{n} \to 0$ as $n \to \infty$, an estimator similar to add-3/4 is optimal [3]
and for $\frac{k}{n} \to \infty$ as $n \to \infty$, add-log $\frac{n}{k}$ is shown to be optimal in [9].

However, add-constant estimators perform poorly in practice since the worst-case distribution for these estimators is the uniform distribution which is rarely encountered in practice. Thus, practitioners use other estimators such as absolute-discounting, cross-validation based estimators and those based on the Good-Turing estimator [5]. Of these three, the most popular has been the latter which consists of estimators of the form

$$q_{x^n}(x) = \frac{\phi(n(x) + 1)n(x) + 1}{\phi(n(x))}$$

where $n(x)$ denotes the number of times the symbol $x$ appears in the sample $x^n$ and $\phi(t)$ is the number of symbols that appear $t$ times in $x^n$.

Although Good-Turing based estimators have been widely used in practice, the theory behind it has been unfolding only recently, for instance in [6], [4], [7], [1]. The key contribution of the paper we consider [8] is to show that a Good-Turing based estimator is within $O\left(\min\left(\frac{1}{\sqrt{n}}, \frac{k}{n}\right)\right)$ of the best estimator in all regimes of interest.

This is established via the competitive setting, where unlike the KL divergence, the performance of an estimator is evaluated as the regret against an oracle. Two such oracles are considered, the first of which knows the underlying distribution upto a permutation and the second which knows the underlying distribution but is allowed to use only ‘natural’ estimators. We call the first oracle - permutation oracle and the second - natural oracle.

Let $P_{\sigma}$ be the partition of $\Delta_k$ into classes of distributions that are equivalent up to a permutation. The competitive regret against the permutation oracle is thus defined as

$$r^p_{n}(\Delta_k) \triangleq \min_{q} \max_{P \in P_{\sigma}}(r_n(q, P) - r_n(P)),$$

where $r_n(q, P) \triangleq \max_{p \in P} r_n(q, p)$ and $r_n(P) \triangleq \min_{q} \max_{p \in P} r_n(q, p)$. For each class $P \in P_{\sigma}$, the permutation oracle knowing that $p \in P$, chooses the best estimator for that class $P$ and $r_n(P)$ is the resulting loss. In this context, we also define $r^p_{n}(q, \Delta_k) \triangleq \max_{P \in P}(r_n(p, q) - r_n(P))$.

On the other hand, a ‘natural’ estimator $q$ is one which is forced to assign the same probability to symbols that occur the same number of times. Then the regret against the natural oracle is defined as

$$r^p_{n}(\Delta_k) = \min_{q} \max_{p \in \Delta_k}(r_n(q, p) - r^p_{n}(p))$$

where $r^p_{n}(p) \triangleq \min_{q \in Q^{nat}} r_n(q, p)$. Here for any $p \in \Delta_k$, the sample $x^n$ induces a collection of ‘usable’ natural estimators $Q^{nat}$ from which the oracle can choose from. The oracle simply picks the one that minimizes the loss $r_n(q, p)$ and the resulting loss is $r^p_{n}(p)$. Just as for the permutation estimator, we also define $r^p_{n}(q, \Delta_k) = \max_{p \in \Delta_k}(r_n(q, p) - r^p_{n}(p))$.

We then have the following theorems which establish the near optimality of the Good-Turing based estimator.

**Theorem 1.** For any $k$ and $n$, a variant of the Good-Turing estimator $q'$ defined in [1, 2](and defined in Section 4) satisfies $r^p_{n}(q', \Delta_k) \leq r^p_{n}(q', \Delta_k) \leq O\left(\min\left(\frac{1}{\sqrt{n}}, \frac{k}{n}\right)\right)$.
Thus Theorem 1 validates the claims made above. We elucidate proof of this theorem in 3 stages. In Section 2 it is shown that for any estimator $q$, $r_n^{P_\sigma}(q, \Delta_k) \leq r_n^{nat}(q, \Delta_k)$ so that in particular, the result holds for $q'$. In Section 3 the problem of competing against the natural oracle is reduced to that of estimating the combined mass. This, in-turn, allows us to use the Good-Turing based estimator $q'$ defined in [1, 2] (and defined in Section 4). The latter is motivated by the fact that Good-Turing estimators estimate the combined mass to a high degree of accuracy, as shown in [6], [4], [7], [1]. This then completes proof of Theorem 1.

**Theorem 2.** For any $k$ and $n$, $r_n^{nat}(\Delta_k) \geq \tilde{\Omega}(\min(\frac{1}{\sqrt{n}}, \frac{k}{n}))$, $r_n^{P_\sigma}(\Delta_k) \geq \tilde{\Omega}(\min(\frac{1}{n^{2/3}}, \frac{k}{n}))$

The above theorem thus gives a matching lower bound for regret against the natural oracle. The lower bound however remains loose when pitted against the permutation oracle. We provide a description of the proof of the latter result (i.e. lower bound for the competitive regret against the permutation oracle) in Section 5.

2 Natural vs Permutation Class

In this section, we outline the proof of the first part of Theorem 1 i.e. that for any estimator $q$, we have

$$r_n^{P_\sigma}(q, \Delta_k) \leq r_n^{nat}(q, \Delta_k) \quad (1)$$

Recall that $r_n^{P_\sigma}(q, \Delta_k)$ is the competitive regret of the estimator $q$ with respect to an oracle that knows the underlying distribution up to a permutation ($P_\sigma$ is a partition of $\Delta_k$ into permutation equivalence classes, i.e. distributions $p$ and $p'$ belong to the same partition class of $P_\sigma$ if there exists a permutation $\sigma$ of $[k]$ such that $p'_{\sigma(i)} = p_i$). On the other hand, $r_n^{nat}(q, \Delta_k)$ is the competitive regret of the estimator $q$ with respect to an oracle that knows the underlying distribution but is forced to assign the same probability to symbols that appeared the same number of times in the sample.

So what inequality 1 actually implies is that any estimator $q$ has lower regret to an oracle that knows a permutation of the distribution in comparison to a natural oracle-aided estimator. The key idea of the proof is to relate the regrets of the two classes, which is achieved by the following lemma.

**Lemma 1.** For every class part $P$ in $P_\sigma$, we have

$$r_n(P) \geq \max_{p \in P} r_n^{nat}(p)$$

Note that $r_n(P)$ is the min-max loss incurred by an oracle that knows that the underlying distribution lies in the partition class $P \in P_\sigma$, and $r_n^{nat}(p)$ is the minimum loss achieved by a natural estimator with respect to the distribution $p$. Intuitively, we expect this result to be true because if the oracle knows the underlying distribution up to a permutation, then the only information that can be used by the oracle are the multiplicities of each of the symbols, and thus its work emulates a natural estimator in that sense. So the incurred loss should be lower bounded by the worst-case loss of a natural estimator.

The key to prove this lemma is to show that there exists an optimal estimator that is natural,
namely the estimator that considers all possible permutations of the observed sequence \( x^n \). In mathematical terms, this optimal estimator looks something like the following:

\[
q''_{x^n}(y) = \frac{\sum_{\sigma \in \Sigma_k} p(\sigma(x^n y))}{\sum_{\sigma' \in \Sigma_k} p(\sigma'(x^n))}
\]

where \( \Sigma_k \) is the set of all permutations of \( k \) symbols. It is easy to see that any two symbols \( y \) and \( y' \) that appeared the same number of times in \( x^n \) will be assigned the same probability estimates under \( q'' \), i.e. \( q'' \) is natural. Note that the oracle knows the underlying distribution \( p \) up to a permutation, and thus has no problem in constructing the estimator \( q'' \).

Now, we show the optimality of this estimator. One observation that we can make is the fact that \( q''_{x^n}(y) = q''_{\sigma(x^n)}(\sigma(y)) \) for any permutation \( \sigma \) (simply because the definition of \( q'' \) invokes all possible permutations on \( x^n y \)), so we have that

\[
\max_{p \in \mathcal{P}} r_n(q''_{x^n}, p) = \frac{1}{k!} \sum_{\sigma \in \Sigma_k} r_n(q''_{x^n}, p(\sigma(\cdot)))
\]

simply because the estimator \( q'' \) incurs the same loss for every \( p \in \mathcal{P} \) (since all distributions in \( \mathcal{P} \) are equivalent up to a permutation).

On the other hand, another important observation that we can make is the following: if \( P \in \mathbb{P}_\sigma \), then for any estimator \( q \),

\[
\max_{p \in P} \mathbb{E}[D(p||q)] \geq \frac{1}{k!} \sum_{\sigma \in \Sigma_k} r_n(q''_{x^n}, p(\sigma(\cdot)))
\]

Several facts are used to show this. First, the maximum regret incurred by an estimator \( q \) with respect to a partition part \( P \) is greater than the average of the regrets incurred with respect to each of the distributions in \( P \). Also, by expanding the formula of the KL divergence and using the fact that all distributions in \( P \in \mathbb{P}_\sigma \) have the same entropy, with the fact that the KL divergence of any two probability distributions \( q \) and \( q'' \) is non-negative, then inequality \(3\) follows.

What equation \(2\) and inequality \(3\) imply is that the regret of any estimator \( q \) is greater than that of \( q'' \), and thus there exists a natural estimator that achieves the minimum loss for the oracle that knows a permutation of the distribution. This result, along with the max-min inequality of multivariate functions, imply Lemma \(1\).

In words, Lemma \(1\) says that the worst-case loss incurred by a natural oracle-aided estimator is still smaller than the loss incurred by an oracle that knows a permutation of the distribution. And hence, the competitive regret of an estimator \( q \) is larger with respect to a natural oracle-aided estimator than with respect to an oracle that is limited to knowing a permutation of the probability distribution, thereby implying inequality \(1\).

Note that the same result holds for any coarser partition of the collection of distributions (e.g. distributions with same entropy, or same sparse support). Intuitively, this is because a coarser partition means that the oracle knows less information about the underlying distribution and thus the competitive regret of any estimator \( q \) would be smaller. For instance, if
\( P_{\text{ent}} \) is the partition of \( \Delta_k \) into distributions of the same entropy, we have

\[
  r_n^{P_{\text{ent}}} (q, \Delta_k) \leq r_n^{P_{\sigma}} (q, \Delta_k)
\]

Also note that the above fact holds for a partition induced by any other symmetric property of the distribution (such as support size, \( l_2 \) norm, etc.). Thus is because all distributions within a given permutation class evaluate to the same symmetric property and is thus a coarser partition. Thus, competitive optimality against the permutation oracle also implies optimality against oracles which know the particular symmetric property. This, in-turn, provides another justification for choosing the permutation oracle to compete against in the paper considered \([8]\).

## 3 Combined Mass Reduction

In this section we reduce the problem of competing against the natural oracle to that of estimating the combined mass. The combined mass \( S(t) \) is the total mass of the elements that occur \( t \) times. Also recall that \( \phi(t) \) is used to denote the number of elements which occur \( t \) times. Then the KL divergence between \( S(t) \) and its estimate \( \hat{S}(t) \) is defined as:

\[
  D(S||\hat{S}) = \sum_{t=0}^{n} S(t) \log \frac{S(t)}{\hat{S}(t)}.
\]

**Lemma 2.** For a natural estimator \( q \), if we let \( \hat{S}(t) = \sum_{x:n(x)=t} q(x) \), \( r_n^{\text{nat}} (q, p) = \mathbb{E}[D(S||\hat{S})] \).

We first provide a sketch of the proof. We then justify it intuitively and also discuss its implications.

Proof of Lemma 2 involves two steps. The regret encountered by the natural oracle is computed first and this oracle regret is used in the next step to prove the lemma. Towards computing the natural oracle regret, it is shown that the best natural estimator that minimizes \( r_n^{\text{nat}} (p) \) is

\[
  q^*(x) = \frac{S(n(x))}{\phi(n(x))}
\]

Proof of the above cleverly uses the fact that \( q_{x^n} (x) \) is natural. It does so by considering the estimator dependent term of the regret expression: \( \sum_{x \in X} p(x) \log \frac{1}{q_{x,x}(x)} \) and then noting that we can sum together the probabilities \( p(x) \) of all those \( x \in X \) which occur the same number of times, i.e. \( \sum_{x \in X} p(x) \log \frac{1}{\hat{q}_{x^n(x)}} = \sum_{t=0}^{n} \sum_{x:n(x)=t} p(x) \log \frac{1}{\hat{q}_{x^n(x)}} \). Non-negativity of KL-divergence, which again is a technique that occurs throughout the paper, then establishes \( q^*(x) = \frac{S(n(x))}{\phi(n(x))} \). Simply plugging \( q^*(x) \) in the expression for regret, \( r_n (q, p) \), then gives

\[
  r_n^{\text{nat}} (p) = \mathbb{E} \left[ \sum_{t=0}^{n} S(t) \log \frac{\phi(t)}{S(t)} \right] - H(p).
\]

To complete proof of Lemma 2 by similar manipulations as above, it’s shown that for any natural estimator \( q \), \( \sum_{x \in X} p(x) \log \frac{1}{q_{x^n(x)}} = \sum_{t=0}^{n} S(t) \log \frac{S(t)}{\hat{S}(t)} + \sum_{t=0}^{n} S(t) \log \frac{\phi(t)}{S(t)} \). Then, subtracting \( r_n^{\text{nat}} (p) \) from this expression proves Lemma 2.

To see why Lemma 2 might be true, as shown above, we may split the KL loss of any estimator into two terms \( \sum_{t=0}^{n} S(t) \log \frac{S(t)}{\hat{S}(t)} \) and \( \sum_{t=0}^{n} S(t) \log \frac{\phi(t)}{S(t)} \). We call the former ‘Combined mass divergence (CMD)’ and the latter ‘Multiplicity Error (ME)’.
Then, note that ME is incurred by the natural oracle as well. This happens because the natural oracle is forced to use a natural estimator. Thus, if a set of elements occur the same number of times, the oracle is forced to assign the same probability to all these elements and ME captures the resulting error. On the other hand, if each element occurs a distinct number of times with probability 1,

$$E[ME] = E[\sum_{t=0}^{n} S(t) \log \frac{\phi(t)}{S(t)}] = H(p)$$

so that $$r_n^{nat}(p) = 0$$ as we might expect. Thus, CMD is the loss that remains when we remove the error incurred by the natural oracle thereby justifying Lemma 2.

Lemma 2 allows for a reduction of the problem at hand to that of estimating the combined mass. As motivated in Section 1, it is in the regime where $$\frac{k}{n} \to \infty$$, where k is the alphabet size and n is the number of samples, that traditional empirical estimators perform poorly.

On the other hand, estimating the combined mass of symbols which occur a given number of times is independent of the underlying alphabet size. Thus the reduction provided by Lemma 2 allows us to overcome a significant obstacle - by enabling the use of Good-Turing based estimators which are known to perform well for the reduced problem.

### 4 Good-Turing Optimality

As we outlined in Section 3, finding the best competitive natural estimator is equivalent to finding the best estimator for the combined probability mass. Acharya et al. proposed in [2] an estimator for $$S$$ that achieves a regret that is upper bounded as depicted in Theorem 1. In this section, we show a proof sketch of this upper bound (please note that this is a different paper from the one considered for this report, and that’s why only a high-level description is given).

The proposed estimator $$S(t)$$ is defined as a combination of the Good-Turing estimator $$G(t)$$, the empirical estimator $$E(t)$$ and a “genie-like-aided” estimator $$F(t)$$. We skip defining the Good-Turing estimator and the empirical estimator as they are well-known, but show the definition of the so-called genie-like-aided estimator:

$$F(t) = \phi(t) \frac{t + 1}{n} \frac{E[\phi(t+1)]}{E[\phi(t)]]}$$

For this estimator, the ratio $$\frac{E[\phi(t+1)]}{E[\phi(t)]}$$ is approximated through exploiting properties of Poisson functionals, and it is shown that the bias in this estimate is small. The proposed unnormalized estimator looks something like the following:

$$\hat{S}^{un}(t) = \begin{cases} 
\max(G(0), \frac{1}{n}) & \text{if } t = 0 \\
E(t) & \text{if } \phi(t) \leq \log^2 n \\
\max(G(t), \frac{1}{n}) & \text{if } t \leq \log^2 n \text{ and } \phi(t) > \log^2 n \\
\min \left( \max \left( F(t), \frac{1}{n^2} \right), 1 \right) & \text{otherwise}
\end{cases}$$

And the proposed normalized estimator is

$$S(t) = \frac{\hat{S}^{un}(t)}{\sum_{t'=0}^{n} \hat{S}^{un}(t')}.$$
One reason for proposing such an estimator is that for small values of $\phi(t)$, the empirical estimator performs well, while for small values of $t$, the Good-Turing estimator has better performance. Also, for large values of $t$, the genie-like-aided estimator avoids the large variance of the Good-Turing estimator and achieves an estimation error for a specific multiplicity $t$ that is sub-linear in both $t$ and $\mathbb{E}[\phi(t)]$ (both the Good-Turing and empirical estimators have errors that are linear in one of $t$ and $\phi(t)$ and sub-linear in the other).

To show the upper bound, the error of the genie-like-aided estimator is first bounded, through a simple application of the triangle inequality and the union bound. Also, bounds on the estimation errors of the empirical estimators and the Good-Turing estimator follow from an application of Bernstein’s inequality and the Chernoff bound. The exact bounds are shown in Lemmas 10, 11 and 26 of [2].

From these bounds, the error of the proposed unnormalized estimator for a specific multiplicity $t$ can be upper bounded by a factor that is sub-linear in the product $t\phi(t)$ and decays like $\tilde{O}(1/n)$, namely it is shown that with high probability, we have

$$|S(t) - \hat{S}^{\text{un}}(t)| = \mathcal{O}\left(\frac{\min(\sqrt{\phi(t)(t + 1)}, \phi(t)^{7/11}\sqrt{t + 1})}{n \log^{-3} n}\right)$$

It can also be seen that this error is smaller than that of both the Good-Turing and empirical estimators by up to polylog($n$) factors. Furthermore, by repeated application of Cauchy-Schwarz inequality, we can show that the normalizing constant $N = \sum_{t=0}^{n} \hat{S}^{\text{un}}(t)$ is very close to 1, namely that, with very high probability, we have $|N - 1| = \tilde{O}\left(\frac{1}{n^{1/4}}\right)$. Finally, using these two facts, the KL divergence between $S(t)$ and $\hat{S}(t)$ can be upper bounded by $\tilde{O}\left(\frac{1}{n^{1/2}}\right)$.

A slight modification of the proof and using the fact that $\sum_{t=1}^{n} \sqrt{\phi(t)} \leq \sum_{t=1}^{n} \phi(t) \leq k$, it can be shown that the error of the estimator $\hat{S}$ is also upper bounded by $\tilde{O}\left(\frac{k}{n}\right)$, and thus the upper bound in Theorem 1 follows, i.e. for any $k$ and $n$, the proposed estimator satisfies

$$\max_{p \in \Delta_k} \mathbb{E}[D(S||\hat{S})] = \tilde{O}\left(\min\left(\frac{1}{\sqrt{n}}, \frac{k}{n}\right)\right)$$

Note that a matching lower bound is also shown in [2] and thus the proposed estimator has near-optimal competitive regret relative to natural estimators.

5 Lower Bound

In this section, we give a proof sketch for the second case of Theorem 2, or that $r_{n,s}^\text{P} (\Delta_k) \geq \Omega \left(\min\left(\frac{1}{n^{2/3}}, \frac{k}{n}\right)\right)$. To prove this, it’s first noted that to lower bound $r_{n,s}^\text{P} (\Delta_k)$, it’s sufficient to lower bound $r_{n,s}^\text{P} (\mathcal{P})$ for any any subset $\mathcal{P} \in \Delta_k$. The lower bound argument uses Fano’s inequality and Gilbert Varshamov bounds.

The idea here is to construct a set $\mathcal{P} \in \Delta_k$ that’s hard to learn. For this, we first define a ‘root’ distribution of alphabet size $m = O(\min(k, (n/(\log^2 n))^{1/3})$, with $p_0(i) \triangleq$
Proof of this uses a candidate estimator \( q \) that sorts the elements observed, and assigns the \( i \)th frequently occurred symbol the probability \( p_0(i) \). Proof then uses fact that \( q \) is a natural estimator and thus the resulting regret \( r_n(P_0) \) can be bounded as \( r_n(P_0) \leq p_{i,j} \log(n) \) where \( p_{i,j} \) is the probability that two elements \( i, j \) ‘cross over’ in their multiplicities. Finally, \( p_{i,j} \) is bounded using bounds on the values of \( p_0(i), p_0(j) \) and \( |p_0(i) - p_0(j)| \) in terms of \( c_{m,n}, m \) from how \( P \) was constructed, thereby proving \( \square \).

Finally \( P \) is constructed by including all permutations of elements in \( P' \). Now within the class \( P \), consider the permutation class induced by a particular \( p_0 \), that is the collection of permutations of \( p_0 \). Call this permutation class \( P_0 \). It’s then shown that

\[
r_n(P_0) \leq \frac{1}{n}, \tag{4}
\]

Proof of this uses a candidate estimator \( q \) that sorts the elements observed, and assigns the \( i \)th frequently occurred symbol the probability \( p_0(i) \). Proof then uses fact that \( q \) is a natural estimator and thus the resulting regret \( r_n(P_0) \) can be bounded as \( r_n(P_0) \leq p_{i,j} \log(n) \) where \( p_{i,j} \) is the probability that two elements \( i, j \) ‘cross over’ in their multiplicities. Finally, \( p_{i,j} \) is bounded using bounds on the values of \( p_0(i), p_0(j) \) and \( |p_0(i) - p_0(j)| \) in terms of \( c_{m,n}, m \) from how \( P \) was constructed, thereby proving \( \square \).

Next, the KL divergence between \( p_0 \) and \( p_{0'} \) is bounded \( \forall \bar{v} \neq \bar{v}' \in P' \). In particular, it’s shown that

\[
\frac{d^2m}{2n} \leq \frac{1}{2} \|p_0 - p_{0'}\|_1^2 \leq D(p_0\|p_{0'}) \leq \frac{48m}{n}.
\]

The upper bound on \( D(p_0\|p_{0'}) \) follows by bounding the KL divergence with Chi-squared distance and from the fact that our choice of \( \epsilon \approx \sqrt{\frac{1}{mn}} \ll \frac{1}{m} \). The lower bound on \( \|p_0 - p_{0'}\|_1^2 \) simply uses the fact that that the Hamming distance between \( \bar{v}, \bar{v}' \geq d'm \). Pinsker’s inequality then completes the proof.

We also have the Fano’s inequality for estimation, which states that for distributions \( p_1, p_2, ..., p_{r+1} \) such that \( D(p_i\|p_j) \leq \beta, \|p_i - p_j\|_1 \geq \alpha \forall i \neq j \), if \( q(X^n) \) is any estimator,

\[
\sup \mathbb{E}[\|p_i - q\|_1] \geq \frac{\alpha}{2} \left( 1 - \frac{n\beta + \log(2)}{\log(r)} \right).
\]

Intuitively, this bound tells us that if we have a large collection of distributions- \( r \), all of them sufficiently far apart in \( \ell_1 \) distance- \( \alpha \), and if we have that the KL Divergence between any pair is small- \( \beta \), it’s difficult to distinguish between the distributions in \( \ell_1 \) distance. This is the key Lemma that’s used in the proof.

Finally, to bound \( r^p_n(P) \), using \( r_n(P_0) \leq \frac{1}{n} \) and Pinsker’s inequality, it’s easy to obtain \( r^p_n(P) \geq \min_n \max_{q \in P'} \frac{1}{2} \mathbb{E}[\|p - q\|_1^2] - \frac{1}{n} \). Since for the class \( P' \), values of \( r, \alpha, \beta \) have been obtained above, Fano’s inequality is applied to prove that

\[
r^p_n(P) = \Omega \left( \frac{m}{n} \right).
\]
Substituting the value of $m$ proves the second case of Theorem 2.

6 Conclusion

This completes description of the paper considered [8]. For further reading, we point the interested reader to [10] which considers the problem competitive estimation under $\ell_1$ loss.

References


