Large Deviations of Empirical Measures

Let $\mathcal{X} = \{1, \ldots, d\}$ be a discrete finite set and let $\{X_t, \ t = 1, 2, \ldots\}$ be a sequence of i.i.d. random variables taking in values in $\mathcal{X}$ with

$$\pi_i = \text{Prob}(X_1 = i), \quad i = 1, \ldots, d.$$ 

Define the unit vectors

$$\mathbf{e}_i = \left(0, \ldots, 0, 1, 0, \ldots, 0\right)^T, \quad i = 1, \ldots, d,$$

the $\mathbb{R}^d$-valued random variables

$$\mathbf{X}_t = \mathbf{e}_{X_t}, \quad t = 1, 2, \ldots,$$

and their sample mean

$$\hat{P}_n = \frac{1}{n} \sum_{t=1}^{n} \mathbf{X}_t, \quad n = 1, 2, \ldots.$$ 

Note that $\hat{P}_n$ is the empirical measure induced on $\mathcal{X}$ by the sequence $X_1, \ldots, X_n$.

From the laws of large numbers, we expect that as $n \to \infty$, $\hat{P}_n$ should converge to $\pi = (\pi_1, \ldots, \pi_d)^T$, the true density of the random variables. By treating $\hat{P}_n$ as a random variable in $\mathbb{R}^d$, we can apply the Gärtner-Ellis theorem to determine the rate of this convergence. In order to do so, we must compute the corresponding rate function $I(\mathbf{x})$. To this end, let $\mathbf{Y}_n = n\hat{P}_n$ and observe that

$$\psi_n(\theta) = \frac{1}{n} \log E\left[e^{\theta^T \mathbf{Y}_n}\right]$$

$$= \frac{1}{n} \log E\left[e^{\theta^T \sum_{t=1}^{n} \mathbf{e}_{X_t}}\right]$$
\[
\psi(\theta) = \log \sum_{i=1}^{d} \pi_i e^{\theta_i},
\]
and
\[
I(x) = \sup_{\theta \in \mathbb{R}^d} \theta^T x - \log \sum_{i=1}^{d} \pi_i e^{\theta_i}.
\]
To determine the supremum in this definition, consider the following cases.

- **Case**: \(x_i < 0\) for some \(i\) in \(\{1, \ldots, d\}\): By letting the corresponding component \(\theta_i\) of \(\theta\) go to \(-\infty\), it is easy to see that \(I(x) = +\infty\).

- **Case**: \(x = 0\): By letting \(\theta\) go to \(-\infty\), it is again easy to see that \(I(0) = +\infty\).

- **Case**: \(x \geq 0\) and \(\sum x_i = 1\): Differentiating the function \(\theta^T x - \log \sum_{i=1}^{d} \pi_i e^{\theta_i}\) with respect to \(\theta_j\) and setting the derivative to zero, we obtain the implicit equations

\[
x_j = \frac{\pi_j e^{\theta_j}}{\sum_{i=1}^{d} \pi_i e^{\theta_i}} = \frac{\pi_j e^{\theta_j}}{K(\theta)},
\]

or

\[
\theta_j = \log \frac{x_j}{\pi_j} + \log K(\theta).
\]

Substituting this back into the expression for \(I(\cdot)\), we get that
\[
I(x) = \sum_{j=1}^{d} x_j \log \frac{x_j}{\pi_j} = D(x\|\pi).
\]

- **Case**: \(x \geq 0\) and \(0 < \sum x_i = s \neq 1\): For any such \(x\) define \(\tilde{x} = s^{-1}x\). Determine the \(\theta^*\) which achieves the supremum in the definition of \(I(\tilde{x})\), as above, and note that

\[
D(\tilde{x}\|\pi) = \sum_{i=1}^{d} \theta^*_i \tilde{x}_i - \log \sum_{i=1}^{d} \pi_i e^{\theta^*_i} = \sum_{i=1}^{d} (\theta^*_i + c) \tilde{x}_i - \log \sum_{i=1}^{d} \pi_i e^{\theta^*_i + c},
\]

for any constant \(c \in \mathbb{R}\). Therefore, if \(\theta + c\) designates the addition of \(c\) to each component of \(\theta\), then
\[
(\theta^* + c)^T x - \log \sum_{i=1}^{d} \pi_i e^{\theta^*_i + c} = (\theta^* + c)^T \tilde{x} - \log \sum_{i=1}^{d} \pi_i e^{\theta^*_i + c} + \sum_{j=1}^{d} (\theta^*_j + c)x_j (1 - s^{-1})
\]

2
\[ D(\bar{x}||\pi) + (1 - s^{-1})g^s T x + c(s - 1), \]

and since the choice of \( c \) is arbitrary, the quantity above can be made arbitrarily large by letting \( c \) go to + or \(-\infty\), depending on the sign of \( s - 1 \). Therefore \( I(x) = +\infty \).

Therefore the rate function for this problem may be written as

\[
I(x) = \begin{cases} 
D(\bar{x}||\pi) & \text{if } x \in P^d, \text{ the unit simplex in } \mathbb{R}^d, \\
+\infty & \text{otherwise.}
\end{cases}
\]

**Sanov's Theorem (Interim Statement):** Let \( G \subset P^d \) be a set of probability distributions on \( X \). If \( G \) is the closure of its interior, then the empirical measure satisfies

\[
\lim_{n \to \infty} \frac{1}{n} \log \text{Prob}(\hat{P}_n \in G) = - \inf_{x \in G} D(\bar{x}||\pi).
\]

Therefore, for sufficiently large \( n \),

\[
\text{Prob}(\hat{P}_n \in G) \approx K(n, \pi, G) e^{-n \inf_{x \in G} D(\bar{x}||\pi)},
\]

where \( K(\cdot, \pi, G) \) varies slowly in \( n \). In its present form, this result follows, as remarked already, by an application of the Gärtner-Ellis theorem. We shall prove a more refined version of this theorem later using techniques which provide us with not only the exponentially decaying term, but also the more slowly varying (polynomial) terms in the probability.

Before introducing the methods for obtaining these finer asymptotics, let us briefly examine the special case when \( G \) is the semi-infinite interval \([a, \infty)\), with \( E[X_1] < a < d \). Assume that \( \pi_i > 0 \) for \( i = 1, \ldots, d \) and note that the set

\[
G = \left\{ x \in P^d : \sum_{i=1}^{d} ix_i \geq a \right\},
\]

is convex and closed in the Euclidean topology. Therefore \( D(\bar{x}||\pi) \), which is a convex function of \( x \), achieves its minimum value. Further, it can be shown that there is a \( \theta^* \in \mathbb{R} \) such that

\[
\inf_{x \in G} D(\bar{x}||\pi) = D(\bar{x}^*||\pi),
\]

where \( x_i^* = \frac{\pi_i e^{i \theta^*}}{\sum_{j=1}^{d} \pi_j e^{j \theta}} \), for \( i = 1, \ldots, d \), with \( \sum_{i=1}^{d} ix_i^* = a \).

**Types and Typical Sequences**

**Definition:** The type of a sequence \( x^n_1 = x_1, \ldots, x_n \) of symbols from a finite alphabet \( X \) is

\[
P_{x^n} = \left( \frac{k_1}{n}, \ldots, \frac{k_d}{n} \right),
\]
where \( k_i \) is the number of occurrences of the \( i \)-th element of \( \mathcal{X} \) in \( x_1^n \), \( i = 1, \ldots, d \), and \( d = |\mathcal{X}| \). The set of all sequences \( x_1^n \in \mathcal{X}^n \) of type \( P \) is denoted \( T_P^n \). The set of all types of \( n \)-length sequences is denoted by \( \mathcal{T}^n \).

**Lemma:** For a given \( \mathcal{X} \) and \( n \), the number of possible types is 
\[
|\mathcal{T}^n| = \binom{n + d - 1}{d - 1}.
\]

**Proof:** Consider an array of \( n + d - 1 \) consecutive slots and the task of placing \( d - 1 \) red and \( n \) blue objects, one in each slot. Clearly, the number of ways to do this, \( i.e. \), the number of ways in which one can first choose \( d - 1 \) slots to place the red objects and then fill up the remaining slots with the blue objects, is 
\[
\binom{n + d - 1}{d - 1}.
\]
For every arrangement of the red and blue objects, if we identify the number of blue objects to the left of the first red object with \( k_1 \), the number of blue objects between the first and second red objects with \( k_2 \), etc., then we get a *bona fide* type \( P \).

**Homework:** Show that 
\[
|\mathcal{T}^n| \leq (n + 1)^{|\mathcal{X}|}.
\]

**Theorem:** For any type \( P = (p_1, \ldots, p_d) \),
\[
\left( \frac{n + d - 1}{d - 1} \right)^{-1} 2^{nH(P)} \leq |T_P^n| \leq 2^{nH(P)},
\]
where \( H(P) = - \sum_{i=1}^d p_i \log_2 p_i \) is the Shannon entropy of \( P \).

**Proof:** By the definition of types, we can write \( p_i = \frac{k_i}{n} \), \( i = 1, \ldots, d \), where \( k_1 + \ldots + k_d = n \), and
\[
|T_P^n| = \frac{n!}{k_1! \ldots k_d!}.
\]

Therefore we can equivalently show that
\[
\left( \frac{n + d - 1}{d - 1} \right)^{-1} \prod_{i=1}^d \left( \frac{n}{k_i} \right)^{k_i} \leq |T_P^n| \leq \prod_{i=1}^d \left( \frac{n}{k_i} \right)^{k_i}.
\]

Now, by the multinomial expansion formula,
\[
n^n = (k_1 + \ldots + k_d)^n = \sum_{(j_1, \ldots, j_d): \; j_1 + \ldots + j_d = n} \frac{n!}{j_1! \ldots j_d!} k_1^{j_1} \ldots k_d^{j_d}.
\]
The largest term in the expansion is \( \frac{n!}{k_1! \ldots k_d!} k_1^{k_1} \ldots k_d^{k_d} \), and from the previous lemma, we know that the number of terms in this sum is \( \binom{n + d - 1}{d - 1} \). Therefore

\[
\frac{n!}{k_1! \ldots k_d!} k_1^{k_1} \ldots k_d^{k_d} \leq n^n \leq \left( \frac{n + d - 1}{d - 1} \right) \binom{n + d - 1}{d - 1} \frac{n!}{k_1! \ldots k_d!} k_1^{k_1} \ldots k_d^{k_d},
\]

and the assertion of the theorem follows by rearrangement. 

**Homework:** Verify that \( \frac{n!}{k_1! \ldots k_d!} k_1^{k_1} \ldots k_d^{k_d} \) is indeed the largest term in the multinomial expansion.

Next, if \( \{X_t, \ t = 1, 2, \ldots \} \) is a sequence of i.i.d. random variables with common measure \( Q \) on \( \mathcal{X} = \{1, \ldots, d\} \), then for every \( x^n \in \mathcal{X}^n \),

\[
Q(X_1 = x_1, \ldots, X_n = x_n) = \prod_{t=1}^{n} Q(X_t = x_t) = \prod_{i=1}^{d} |Q(X_1 = i)|^{k_i} = 2\sum_{i=1}^{d} k_i \log Q(i) = 2^{-n} \sum_{i=1}^{d} \frac{k_i}{n} \log \frac{k_i}{n} - \frac{k_i}{n} \log \frac{k_i}{n} = 2^{-n} \left( H(P_{x^n}^n) + D(P_{x^n}^n || Q) \right).
\]

Therefore, for a type \( P \), the \( Q \)-probability of all sequences in \( T_P^n \) is the same and depends only on \( P \). Combining this with the cardinality of a type-class, we obtain the following important bounds on the probability of a type-class under an i.i.d. measure.

**Lemma:** For any distribution \( Q \) on \( \mathcal{X} \) and any type \( P \),

\[
\left( \frac{n + d - 1}{d - 1} \right)^{-1} 2^{-nD(P||Q)} \leq Q(T_P^n) \leq 2^{-nD(P||Q)}.
\]

**Remarks:** The concepts developed so far provide the following very useful view of \( \mathcal{X}^n \).

- The types in \( T^n \) partition the space into equivalence classes, and sequences in an equivalence class are equiprobable under every measure \( Q \) on \( \mathcal{X} \).

- The number of cells in the partition (\( = |T^n| \)) grows polynomially with \( n \), but the number of \( n \)-length sequences in each cell (\( \approx 2^{nH(P)} \)) grows exponentially with \( n \).
• The $Q$-probability of each cell depends on the I-divergence between the type representing the cell and the measure $Q$ – cells with a type “close” to $Q$ get more probability mass than those whose type is far away from $Q$.

• As $n$ grows, the grid corresponding to $\mathcal{T}^n \subset \mathbb{P}^d$ becomes finer, and for every $Q \in \mathbb{P}^d$, there are types in $\mathcal{T}^n$ which are arbitrarily “close” to $Q$.

The significant observation is that although there are many more types which are far away from $Q$ than types close to it, the number of distinct types of either kind grows at most polynomially with $n$, while the probability of a distant type decays exponentially faster than that of a type close to $Q$. Thus most of the total mass is eventually concentrated on types close to $Q$. 