### Remarks on Large Deviations and Hypothesis Testing

Assume that each of 7 mannequins wears a monocolored dress from the finite alphabet $A = \{a_1, a_2, a_3, a_4\} = \{red, white, green, blue\}$. Irrespective of probabilistic models, the dress colors of the 7 mannequins $x_1^7 = (x_1, \dots, x_7) \in A^7$ can represent $\binom{7+4-1}{4-1} = 120$ different fashion trends, which are called types (note that if two mannequins interchange their dresses, the type will not change). Another example: *anna* and *nana* are of the same type as 4-types over the 2-letter alphabet $A = \{a, n\}$.

Now, if we have a probability distribution (population distribution) $\mathbb{P}$ over the colors, say

$$P(a_1) = \frac{1}{2}, \quad P(a_2) = \frac{1}{4}, \quad P(a_3) = \frac{1}{8}, \quad P(a_4) = \frac{1}{8},$$

then we want to select $n$ girls (we can call them mannequins, but they are not necessarily beautiful) who represent the population trend of fashion. The distribution $\mathbb{P}$ is called $n$-type if it is the type of some $x_1^n$, i.e., if we can select $n$ girls such that the empirical distribution of their dress colors is $\mathbb{P}$. Obviously, the above $\mathbb{P}$ cannot be 7-type, but it can be 8-type. Indeed, if $n = 8$, then 4 girls in red, 2 in white, and 1-1 in green and blue will produce the above color distribution. Observe, that such 8-type sequences are also typical in the sense that they give the mode of the polynomial distribution with parameters 8; $\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}$.

Since $H(\mathbb{P}) = 1.75$, and $\binom{8+4-1}{4-1} = 165$, Lemma 2.2. gives the following upper and lower bound for the cardinality of the type-class $\mathcal{T}_{\mathbb{P}}^8$ (which consists of 8-element samples with empirical distribution $\mathbb{P}$):

$$99 \approx \frac{1}{165} 2^{8 \times 1.75} \leq |\mathcal{T}_{\mathbb{P}}^8| \leq 2^{8 \times 1.75} = 16384$$

which is not too tight, as $|\mathcal{T}_{\mathbb{P}}^8| = \frac{8!}{4!2!1!1!} = 840$. However, when $H(\mathbb{P})$ is small and $n$ is large we can get tighter estimates.

Revisiting the TOTO example, there is no 13-type of the distribution $\mathbb{P} : \frac{1}{2}, \frac{1}{3}, \frac{1}{6}$ over the alphabet $\{1, 2, x\}$. However, the upper bound $2^{13H(\mathbb{P})}$ works even in the situation when this distribution is valid only for some permutation of $\{1, 2, x\}$ (it will not change the entropy). The message is that the smaller the entropy, the smaller the number of possibilities for $n$-length sequences approximately producing this distribution is.

The idea of the proofs of lemmas and theorems is that even if $\mathbb{P}$ cannot be $n$-type (for example, there are irrational entries in the pmf as a fashion dictator recommends, say, $\frac{1}{\sqrt{2}}$ for the probability of *red*), for large $n$ it is close to an $n$-type distribution, which in turn is close to the empirical distribution of an $n$-element sample that we select with high probability.

**Stein lemma:** The set of probability measures is $\mathcal{P} = \{\mathbb{P}_0, \mathbb{P}_1\}$, and denote by $f_0(x)$ and $f_1(x)$ the the corresponding pdf's (we work with absolutely continuous distributions). Let $\mathbf{X} = (X_1, \dots, X_n)^T$ be i.i.d. sample with likelihood function (joint pdf) $L_0(\mathbf{x}) = \prod_{i=1}^n f_0(x_i)$ (if they are from $\mathbb{P}_0$) and $L_1(\mathbf{x}) = \prod_{i=1}^n f_1(x_i)$ (if they are from $\mathbb{P}_1$), where $\mathbf{x} = (x_1, \dots, x_n)^T$. Assume that the relative entropy of $\mathbb{P}_0$ with respect to $\mathbb{P}_1$ exists, i.e.,

$$D(\mathbb{P}_0 || \mathbb{P}_1) = D(f_0 || f_1) = \mathbb{E}_{\mathbb{P}_0} \log_2 \frac{f_0(X_1)}{f_1(X_1)} < \infty$$

(we know that it is nonnegative).

Based on the above sample we want to decide about the following alternative:

$$H_0 : \mathbb{P} = \mathbb{P}_0 \quad \text{versus} \quad H_1 : \mathbb{P} = \mathbb{P}_1.$$

Let $\mathcal{X}_a^{(n)} \subseteq \mathbb{R}^n$ denote the acceptance region, and $\mathcal{X}_c^{(n)} \subseteq \mathbb{R}^n$ the critical (rejection) region of a statistical test for the above alternative (they are complements of each other, and in the absolutely continuous case, exhaust the sample space with probability 1). Further, denote by $\alpha_n = \mathbb{P}_0(\mathbf{X} \in \mathcal{X}_c^{(n)})$ the probability of the Type I error and by $\beta_n = \mathbb{P}_1(\mathbf{X} \in \mathcal{X}_a^{(n)})$ the probability of the Type II error.

Let $0 < \varepsilon < \frac{1}{2}$ be an arbitrary size of the test (supremum of the Type I errors) and

$$\beta_{n,\varepsilon} := \min_{\mathcal{X}_a^{(n)} \subseteq \mathbb{R}^n, \, \alpha_n < \varepsilon} \beta_n$$

be the minimal Type II error among the tests with size at most $\varepsilon$. Then

$$\lim_{n \to \infty} \frac{1}{n} \log_2 \beta_{n,\varepsilon} = -D(\mathbb{P}_0 || \mathbb{P}_1).$$

This was the theorem. Remarks:

- Under the above conditions, the Type II error exponentially decreases with $n$ (consistency of the test).
- The UMP test with $\mathcal{X}_c^{(n)} = \{\frac{L_1(\mathbf{X})}{L_0(\mathbf{X})} \geq c\}$ satisfies these requirements with appropriate $c = c(\varepsilon)$.