6.7800 Spring 2025: Bonus Lecture 03

Proving Sanov's Bound via Minimax Redundancy

Jongha (Jon) Ryu
jongha@mit.edu

April 18, 2025

Contents

1 An Alternative Proof of Sanov's Theorem

In the class, we have learned the celebrated Sanov's theorem, which provides an elegant characterization of large deviation for finite alphabet distributions. In what follows, we let $\mathcal{Y}$ be a finite alphabet of size $|\mathcal{Y}| = K$. Recall:

Theorem 1 (Sanov's Theorem)**.** *Let $\mathcal{S} \subset \mathcal{P}^{\mathcal{Y}}$ be an arbitrary set of distributions, and let $q \in \mathcal{P}^{\mathcal{Y}}$ be arbitrary. Then*

$$Q\{\mathcal{S} \cap \mathcal{P}_N^{\mathcal{Y}}\} \leq (N+1)^K 2^{-ND(p_* \,\|\, q)}, \tag{1}$$

where

$$p_* = \arg\min_{p \in \mathcal{S}} D(p \,\|\, q) \tag{2}$$

is the I-projection of q onto $\mathcal{S}$.

As we learned in the class, the standard proof is based on the method of types. In this lecture, we will derive a similar bound to Sanov's theorem via *minimax redundancy*, which is the driving concept of this course. Interestingly, we can even prove a (slightly) tighter bound as follows.

Theorem 2. *Under the same setting of Theorem 1,*

$$Q\{\mathcal{S} \cap \mathcal{P}_N^{\boldsymbol{y}}\} \le e^{o(1)} \frac{\Gamma(1/2)^K}{\Gamma(K/2)} \left(\frac{N}{2\pi}\right)^{\frac{K-1}{2}} 2^{-ND(p_* \,\|\, q)}. \tag{3}$$

Note that the Chernoff exponent $D(p_* \,\|\, q)$ remains the same as expected from the matching lower bound, the polynomial term $(N+1)^K$ becomes tighter as $N^{\frac{K-1}{2}}$.

1.1 Pointwise Minimax Redundancy

Before we prove the theorem, we recall the notion of *minimax redundancy*. For a class of distributions indexed by a parameter $x \in \mathcal{X}$, we define the (mean) minimax redundancy as

$$\bar{R}_N^* := \min_b \max_{x \in \mathcal{X}} D(p(y^N; x) \,\|\, b(y^N)).$$

In the proof below, we will require a stronger notion, which we call the *pointwise minimax redundancy*:

$$R_N^* := \min_b \max_{x \in \mathcal{X}} \max_{y^N} \log \frac{p(y^N; x)}{b(y^N)}.$$

Note that the expectation with respect to $p(y^N; x)$ is replaced by the maximum over all possible sequence y^N. This notion arises when we would like to analyze the extremely worst-case scenario, assuming that the sequence y^N can be an arbitrary sequence.

Note that, clearly, $R_N^* \ge \bar{R}_N^*$. For K-ary i.i.d. processes, i.e., when $\{p(y^N; x) \colon x \in \mathcal{X}\}$ is the class of all i.i.d. distributions, Xie and Barron [8] showed that

$$R_N^* = \frac{K-1}{2} \log \frac{N}{2\pi} + \log \frac{\Gamma(1/2)^K}{\Gamma(K/2)} + o(1), \tag{4}$$

$$\bar{R}_N^* = \frac{K-1}{2} \log \frac{N}{2\pi e} + \log \frac{\Gamma(1/2)^K}{\Gamma(K/2)} + o(1). \tag{5}$$

Remarkably, $\bar{R}_N^*$ and R_N^* are of the same order for this special case. These results have profound implications in information theory and statistics; see [8] for more details. Specifically, we will invoke Eq. (4) in the proof below.

1.2 Proof of Theorem 2

Proof of Theorem 2. Without loss of generality, we assume that $q \notin \mathcal{S}$, since $Q\{\mathcal{S} \cap \mathcal{P}_N^{\boldsymbol{y}}\} = 1$ and the right hand side is at least 1 otherwise. In this case, $\hat{p}(\cdot; \boldsymbol{y}) \in \mathcal{S}$ implies that

$$D(\hat{p}(\cdot; \boldsymbol{y}) \,\|\, q) \ge \arg\min_{p \in \mathcal{S}} D(p \,\|\, q) = D(p_* \,\|\, q) =: r. \tag{6}$$

We note that

$$\begin{aligned} Q\{\mathcal{S} \cap \mathcal{P}_N^{\boldsymbol{y}}\} &= \mathbb{P}_q\big(\hat{p}(\cdot;\boldsymbol{y}) \in \mathcal{S}\big) \\ &\overset{(a)}{\leq} \mathbb{P}_q\Big(D(\hat{p}(\cdot;\boldsymbol{y}) \,\|\, q) \geq r\Big) \\ &= \mathbb{P}_q\Big(\frac{1}{N}\log\frac{\hat{p}(y^N;\boldsymbol{y})}{q(y^N)} \geq r\Big) \\ &= \mathbb{P}_q\Big(\frac{\hat{p}(y^N;\boldsymbol{y})}{q(y^N)} \geq 2^{Nr}\Big), \end{aligned}$$

where (a) follows from (6). We now introduce an arbitrary probability $b(y^N)$ over the sequence y^N, and continue from the last inequality to obtain

$$\begin{aligned} Q\{\mathcal{S} \cap \mathcal{P}_N^{\boldsymbol{y}}\} &\leq \mathbb{P}_q\Big(\frac{b(y^N)}{q(y^N)} \geq 2^{Nr}\frac{b(y^N)}{\hat{p}(y^N;\boldsymbol{y})}\Big) \\ &\overset{(b)}{\leq} \mathbb{P}_q\Big(\frac{b(y^N)}{q(y^N)} \geq 2^{Nr}\min_{\tilde{y}^N}\min_{p\in\mathcal{P}^{\boldsymbol{y}}}\frac{b(\tilde{y}^N)}{p(\tilde{y}^N)}\Big) \\ &\overset{(c)}{\leq} \mathbb{E}_q\Big[\frac{b(y^N)}{q(y^N)}\Big]2^{-Nr}\max_{\tilde{y}^N}\max_{p\in\mathcal{P}^{\boldsymbol{y}}}\frac{p(\tilde{y}^N)}{b(\tilde{y}^N)} \\ &\overset{(d)}{=} 2^{-Nr}\max_{\tilde{y}^N}\max_{p\in\mathcal{P}^{\boldsymbol{y}}}\frac{p(\tilde{y}^N)}{b(\tilde{y}^N)}. \end{aligned} \tag{7}$$

Here, (c) follows from Markov's inequality, and (d) since $\mathbb{E}_q[\frac{b(y^N)}{q(y^N)}] = 1$. Since this holds for any arbitrary distribution $b(y^N)$, we can take a minimization over $b \in \mathcal{P}^{\mathcal{Y}^N}$ and get

$$Q\{\mathcal{S} \cap \mathcal{P}_N^{\boldsymbol{y}}\} \leq 2^{-Nr}\min_{b}\max_{\tilde{y}^N}\max_{p\in\mathcal{P}^{\boldsymbol{y}}}\frac{p(\tilde{y}^N)}{b(\tilde{y}^N)} = 2^{-Nr+R_N^*},$$

where we define

$$R_N^* := \min_{b}\max_{x\in\mathcal{X}}\max_{\tilde{y}^N}\log\frac{p(\tilde{y}^N;x)}{b(\tilde{y}^N)},$$

which is the *pointwise* minimax redundancy for K-ary i.i.d. probabilities. Invoking the established result on the pointwise minimax redundancy in Eq. (4), we conclude the proof. □

1.3 Remarks on the Optimal Strategy

Though we did not need to know what probability $b \in \mathcal{P}^{\mathcal{Y}^N}$ achieves the pointwise minimax redundancy in Eq. (4), it is informative to note the (near) optimal strategy. Recall from the class that the optimal strategy for the minimax redundancy in Eq. (5) is in the form of a *mixture* distribution: for a distribution over $w(\cdot)$ over $\mathcal{X}$, let

$$b_w(y^N) := \int p(y^N;x)w^*(x)dx.$$

For the i.i.d. processes, the optimal mixture distribution w^* is the Dirichlet distribution with concentration parameter $\boldsymbol{\alpha} = (\frac{1}{2}, \dots, \frac{1}{2})$, which results in the so-called KT mixture

$$q_{\mathsf{KT}}(y^N) := \int p(y^N; x) w(x) \, \mathrm{d}x = \frac{B(\mathbf{k}(y^N) + \boldsymbol{\alpha})}{B(\boldsymbol{\alpha})},$$

which is attributed to Krichevsky and Trofimov [2]. Here, for $y^N \in [K]^N$, define $k_i(y^N) := \sum_{t=1}^N \mathbb{1}\{y_t = i\} = (\text{number of } i\text{'s in } y^N)$[1] and $B(\cdot)$ is the multivariate beta function.[2]

Interestingly, if $\boldsymbol{\alpha} = (1, \dots, 1)$, the Dirichlet distribution becomes the uniform distribution. This is a maximally ignorant prior, but one can check that the uniform mixture only results in a suboptimal redundancy $O((K-1)\log N)$, missing the factor of $1/2$. In words, to achieve the optimal redundancy, we need to put more emphasis on the boundary of the simplex.

A remarkable property of this minimax optimal strategy for the minimax redundancy in Eq. (5) is that it is also nearly minimax optimal for the pointwise minimax redundancy in Eq. (4). We note that the (exactly) minimax optimal strategy for the pointwise case is the *normalized maximum likelihood* defined as

$$b_{\mathsf{NML}}(y^N) := \frac{\max_{x \in \mathcal{X}} p(y^N; x)}{\sum_{\tilde{y}^N} \max_{x \in \mathcal{X}} p(\tilde{y}^N; x)}.$$

Note that this is the *equalizer*. This is well-defined, but due to the normalization constant, it is often not practical. Fortunately, we can use the KT strategy to enjoy almost the minimax optimal redundancy even under this scenario; see [8] for more details.

Lastly, we wish to remark the *predictive* form of the KT mixture distribution, or the mixture distribution induced by a Dirichlet distribution. Thanks to the conjugacy of the Dirichlet distribution to categorical distributions, we readily have

$$q_w(\cdot \mid y^N) = \frac{\mathbb{k}(y^N) + \boldsymbol{\alpha}}{N + \mathbb{1}^\intercal \boldsymbol{\alpha}}$$

when $w(x) = \mathsf{Dir}(x|\boldsymbol{\alpha})$. When $\boldsymbol{\alpha} = 0$, this recovers the maximum likelihood estimate. For positive vectors $\boldsymbol{\alpha}$, it suggests that the prediction should be done with *smoothed* empirical counts, where $\boldsymbol{\alpha}$ can be understood as *pseudocounts*. To achieve the optimal redundancy, we should perform the prediction as if there were some pseudocounts for each symbol (even though we have not observed anything yet), and 0.5 happens to be the *optimal* smoothing!

2 Confidence Set

Sanov's bound quantifies the probability of rare events. The story of large deviation is closely related to the phenomenon of *concentration of measures*, and we provide this alternative view here.

[1] As a vector $\mathbf{k}(y^N)$ can be understood as the count vector of all symbols in y^N.

[2] Here, for a positive vector $\boldsymbol{\alpha} \in \mathbb{R}^K_{>0}$, we denote the multivariate beta function as

$$B(\boldsymbol{\alpha}) := \frac{\prod_{i=1}^K \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^K \alpha_i)},$$

where $\Gamma(\alpha) := \int_0^\infty t^{\alpha-1} e^{-t} \, \mathrm{d}t$ $(\alpha > 0)$ denotes the gamma function.

Suppose that a K-ary process $y_1, \ldots, y_N$ is drawn i.i.d. from a distribution $p(y^n; \boldsymbol{\mu})$ with unknown mean parameter

$$\boldsymbol{\mu} = (\mathbb{E}[\mathbb{1}\{y = i\}])_{i=1}^{K} = \mathbb{E}[\mathbf{e}_y].$$

In this discrete case, we expect the empirical distribution $\hat{\boldsymbol{\mu}}_N := \frac{\mathbb{k}(y^N)}{N}$ (i.e., the type of the sequence) to converge to the true data generating distribution as N becomes large. Using the same technique introduced above in a slightly different manner, we can derive a *confidence set* for the true data generating process (or equivalently the mean vector). Formally, for $\delta \in (0, 1)$, we say that a set-valued function $C_\delta(y^N)$ is a *confidence set* for the mean parameter $\boldsymbol{\mu}$ at level (or coverage) $1 - \delta$, if

$$\mathsf{P}\{\boldsymbol{\mu} \in C_\delta(y^N)\} \geq 1 - \delta.$$

Recall that we applied Markov's inequality in Eq. (7) to the random variable $\frac{b(y^N)}{q(y^N)}$. We again apply Markov's inequality as

$$\mathsf{P}_q\Big(\frac{b(y^N)}{q(y^N)} \geq \frac{1}{\delta}\Big) \leq \delta,$$

since $\mathbb{E}_q[\frac{b(y^N)}{q(y^N)}] = 1$ for some $\delta > 0$ and for any choice of distribution $b(\cdot) \in \mathcal{P}^{\mathcal{Y}^N}$. We can rewrite this inequality as

$$\mathsf{P}_q\Big(D(\hat{\boldsymbol{\mu}}_N \parallel q) \geq \frac{1}{N}\log\frac{1}{\delta} + \frac{1}{N}\frac{p(y^N; \hat{\boldsymbol{\mu}}_N)}{b(y^N)}\Big) \leq \delta.$$

This implies that if we define

$$C_\delta(y^N) := \Big\{\mathbf{m} \in \mathcal{P}^{\mathcal{Y}} : D(\hat{\boldsymbol{\mu}}_N \parallel \mathbf{m}) > \frac{1}{N}\log\frac{1}{\delta} + \frac{1}{N}\frac{p(y^N; \hat{\boldsymbol{\mu}}_N)}{b(y^N)}\Big\}, \tag{8}$$

then $C_\delta(y^N)$ is a confidence set with level $1 - \delta$. Geometrically, this confidence set is a KL divergence ball centered around the empirical mean $\hat{\boldsymbol{\mu}}_N$, where the radius is $\frac{1}{N}\log\frac{1}{\delta} + \frac{1}{N}\frac{p(y^N; \hat{\boldsymbol{\mu}}_N)}{b(y^N)}$. Note that this is a *meta* algorithm, which defines a confidence set for each probability $b \in \mathcal{P}^{\mathcal{Y}^N}$.

Since we do not know the data generating process, it is natural to plug-in the KT mixture $q_{\mathsf{KT}}(y^N)$, to minimize the radius of the KL ball. If we apply Stirling's approximation on the second term $\frac{1}{N}\frac{p(y^N; \hat{\boldsymbol{\mu}}_N)}{q_{\mathsf{KT}}(y^N)}$, we can show that, for N sufficiently large,

$$\mathsf{P}_q\Big(D(\hat{\boldsymbol{\mu}}_N \parallel q) \geq \frac{1}{N}\log\frac{1}{\delta} + \frac{K-1}{2N}\log N + O(1)\Big) \leq \delta.$$

We remark that, in the same language, Sanov's bound can be translated into as

$$\mathsf{P}_q\Big(D(\hat{\boldsymbol{\mu}}_N \parallel q) \geq \frac{1}{N}\log\frac{1}{\delta} + \frac{K}{N}\log(N+1)\Big) \leq \delta.$$

3 Time-Uniform Guarantee

Interestingly, it turns out that we can claim a stronger guarantee for the constructed confidence set, almost for free. Recall that our key technique so far is simply Markov's inequality on the random variable $\frac{b(y^N)}{q(y^N)}$, noting that $\mathbb{E}_q[\frac{b(y^N)}{q(y^N)}] = 1$. This relation indeed holds in a stronger sense. That is, we can show that the stochastic process $\mathsf{W}_N := \frac{b(y^N)}{q(y^N)}$ for $N \geq 1$ with $\mathsf{W}_0 = 1$ is a *martingale*.

Lemma 3. *Suppose that* $\mathbb{E}[\mathbf{e}_{y_N}|y^{N-1}] = \boldsymbol{\mu}$ *for any* $N \geq 1$. *For any causal gambling strategy,*

$$\mathbb{E}_q[\mathsf{W}_N|y^{N-1}] = \mathsf{W}_{N-1}$$

for any $N \geq 1$.

The proof is left as an exercise. This states that the sequence of probability ratio is a *martingale* sequence. In general, if a sequence $(\mathsf{W}_t)_{t=1}^N$ with respect to a stochastic outcome y^N satisfies that $\mathbb{E}[\mathsf{W}_N|y^{N-1}] \leq \mathsf{W}_{N-1}$ for any $N \geq 1$, then it is called *supermartingale*, and if the inequality holds with equality it is called *martingale*. Intuitively, this is a realistic model for gambling in real-world casino, where the gambling is statistically not favorable to gamblers.

In this case, a stronger bound holds in place of Markov's inequality. The following inequality is due to Ville [6].

Theorem 4 (Ville's inequality)**.** *Let* $(W_t)_{t=0}^\infty$ *be a nonnegative supermartingale sequence. Then, for any* $\delta > 0$, *we have*

$$\mathsf{P}\Big(\sup_{t\geq 1} \frac{W_t}{W_0} \geq \frac{1}{\delta}\Big) \leq \delta.$$

Its proof is based on a simple application of the optional stopping theorem; see, e.g., the linked note. In words, this states that the probability that a gambler's wealth against a fair or adversarial casino (that always results in a (super-)martingale) ever goes beyond a certain threshold $1/\delta$ is at most δ. Then, we can readily get a time-uniform guarantee as follows for the confidence set (8): Since

$$\mathsf{P}_q\Big(\sup_{t\geq 1} \frac{b(y^N)}{q(y^N)} \geq \frac{1}{\delta}\Big) \leq \delta$$

for any causal strategy b, we obtain

$$\mathsf{P}_q\big\{\forall t \geq 1, \boldsymbol{\mu} \in C_\delta(y^t)\big\} \geq 1 - \delta.$$

Note the difference from the previous guarantee without time-uniformity:

$$\forall t \geq 1,\ \ \mathsf{P}_q\big\{\boldsymbol{\mu} \in C_\delta(y^t)\big\} \geq 1 - \delta.$$

We note that such a time-uniform confidence set is more suitable for sequential decision making, as it allows a user to make a decision at any time collecting samples gradually.

4 Concluding Remarks

The alternative proof of Sanov's bound here is not available in the literature, but the technique developed in this note is a simplified and adapted version for Sanov's bound of that studied by Ryu and Wornell [4]. The time-uniform confidence sequence has recently gained increasing attention in the statistics and computer science community, due to its versatility in real-world sequential decision making problems such as A/B testing, bandits, and election auditing, to name a few.

The *universal prediction* technique we used in this note can be also applied to constructing confidence sets for *continuous* random variables, but it requires to introduce the notion of *gambling*, where we will define a martingale process as a sequence of wealth, instead of the probability ratio $\frac{b(y^N)}{q(y^N)}$. In this case, a good strategy we can again use is the *universal portfolio* developed by Cover [1], which is a gambling strategy *induced* by the KT mixture. We refer an interested reader to a thought-provoking paper of Waudby-Smith and Ramdas [7], which delved into the idea of gambling for constructing time-uniform confidence sequence for bounded scalar random variables, and the paper of Orabona and Jun [3] which proposed to apply Cover's universal portfolio. Ryu and Wornell [4] recently extended this framework to vector-valued observations such as categorical data, probability-valued observations, and bounded-valued observations. The technique can be also applied to derive a new concentration inequality for nonnegative random variables; see [5].

References

[1] Thomas M Cover. Universal portfolios. *Math. Financ.*, 1(1):1–29, 1991.

[2] Raphail Krichevsky and Victor Trofimov. The performance of universal encoding. *IEEE Trans. Inf. Theory*, 27(2):199–207, 1981.

[3] Francesco Orabona and Kwang-Sung Jun. Tight concentrations and confidence sequences from the regret of universal portfolio. *IEEE Trans. Inf. Theory*, 70(1):436–455, 2024. doi: 10.1109/TIT.2023.3330187. arXiv:2110.14099.

[4] J. Jon Ryu and Gregory W. Wornell. Gambling-based confidence sequences for bounded random vectors. In *ICML*, 2024. arXiv:2402.03683.

[5] J Jon Ryu, Jeongyeol Kwon, Benjamin Koppe, and Kwang-Sung Jun. Improved offline contextual bandits with second-order bounds: Betting and freezing. *arXiv preprint arXiv:2502.10826*, 2025.

[6] Jean Ville. Etude critique de la notion de collectif. *Bull. Amer. Math. Soc*, 45(11):824, 1939.

[7] Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *J. R. Stat. Soc. B*, 86(1):1–27, 2024.

[8] Qun Xie and Andrew R Barron. Asymptotic minimax regret for data compression, gambling, and prediction. *IEEE Trans. Inf. Theory*, 46(2):431–445, 2000.