EE376A/STATS376A Information Theory Lecture 14 - 02/22/2018

Lecture 14: Sanov's Theorem

Lecturer: Tsachy Weissman *Scribe: T Diamandis, R Gabrielsson, A Mohamed, G Murray*

In this lecture, we will introduce and prove Sanov's theorem, a useful tool in probability and statistics that is relevant for many key characterizations and theorems throughout the course. We will start with a recap of the method of types then proceed to discuss the main theorem.

1 Recap of the Method of Types

Consider the sequence $x^n \in \mathcal{X}^n$, where $\mathcal{X}$ is a finite alphabet. Let P_{x^n} be the empirical distribution such that $P_{x^n}(a) = \frac{N(a|x^n)}{n}$, where $N(a|x^n)$ denotes the number of times the symbol a appeared in the sequence x^n. Let $\mathbb{P}_n$ be the set of all empirical distributions over sequences of length n. Then we define the type class to be:

$$T(P) = \{x^n : P_{x^n} = P\} \text{ for } P \in \mathbb{P}_n$$

We have shown the following results:

- $|\mathbb{P}_n| \leq (n+1)^{|\mathcal{X}|-1}$
- $Q(x^n) = 2^{-n[H(P_{x^n})+D(P_{x^n}||Q)]}$
- For $P \in \mathbb{P}_n$: $\frac{1}{(n+1)^{|\mathcal{X}|-1}} 2^{nH(P)} \leq |T(P)| \leq 2^{nH(P)}$
  - Equivalently: $|T(P)| \doteq 2^{nH(P)}$ (see Section 2)
- For $P \in \mathbb{P}_n, Q$, where Q describes the true source of X: $\frac{1}{(n+1)^{|\mathcal{X}|-1}} 2^{-nD(P||Q)} \leq Q(T(P)) \leq 2^{-nD(P||Q)}$
  - Equivalently: $Q(T(P)) \doteq 2^{-nD(P||Q)}$ (see Section 2)
  - This follows from the previous two results

2 Notation

We write $\alpha_n \doteq \beta_n$ to denote equality on an exponential scale, or equality to first order in the exponent. More precisely, we have

$$\alpha_n \doteq \beta_n \iff \frac{1}{n} \log \frac{\alpha_n}{\beta_n} = \frac{1}{n} \log \alpha_n - \frac{1}{n} \log \beta_n \to 0 \text{ as } n \to \infty$$

Example:

$$\alpha_n \doteq 2^{n\gamma} \iff \alpha_n = 2^{n(\gamma+\epsilon_n)}, \text{ where } \epsilon_n \to 0 \text{ as } n \to \infty$$

Convention for empty sets: The maximum over an empty set is negative infinity; the minimum is positive infinity.

3 Sanov's Theorem

The version of Sanov's Theorem we consider bounds the probability that a function's empirical mean exceeds some value α. We begin by introducing some notation and stating the theorem.

Notation:
We let $\mathcal{M}(\mathcal{X})$ denote all pmf's on $\mathcal{X}$. Then for $P \in \mathcal{M}(\mathcal{X})$ and $f : \mathcal{X} \to \mathbb{R}$ we define the inner product:

$$\langle P, f\rangle \ = \sum_{a \in \mathcal{X}} P(a) f(a) = \mathbb{E}_{X \sim P}[f(X)]$$

Theorem 1. ***A Version of Sanov's Theorem:***
For X_i, iid $\sim Q$, and a function $f : \mathcal{X} \to \mathbb{R}$:

$$\frac{1}{(n+1)^{|\mathcal{X}|-1}} 2^{-nD_n^*(\alpha)} \leq Pr\left(\frac{1}{n}\sum_{i=1}^{n} f(X_i) \geq \alpha\right) \leq (n+1)^{|\mathcal{X}|-1} 2^{-nD_n^*(\alpha)}$$

where

$$D_n^*(\alpha) = \min_{P \in \mathbb{P}_n : \langle P, f\rangle \geq \alpha} D(P||Q)$$

As $n \to \infty$, the set of $\mathbb{P}_n$, which has components that are integer multiples of $\frac{1}{n}$ is dense in the set of all probability mass functions. Specifically, we can approximate any $P \in \mathcal{M}(\mathcal{X})$ arbitrarily well with a $P_n \in \mathbb{P}_n$ for large enough n. Thus, we have

$$\Pr\left(\frac{1}{n}\sum_{i=1}^{n} f(X_i) \geq \alpha\right) \doteq 2^{-nD^*(\alpha)}$$

where

$$D^*(\alpha) = \min_{P \in \mathcal{M}(\mathcal{X}) : \langle P, f\rangle \geq \alpha} D(P||Q)$$

3.1 Geometric Picture

For this example, let $|\mathcal{X}| = 3$, so our probability mass function lies on a plane in $\mathbb{R}^3$.

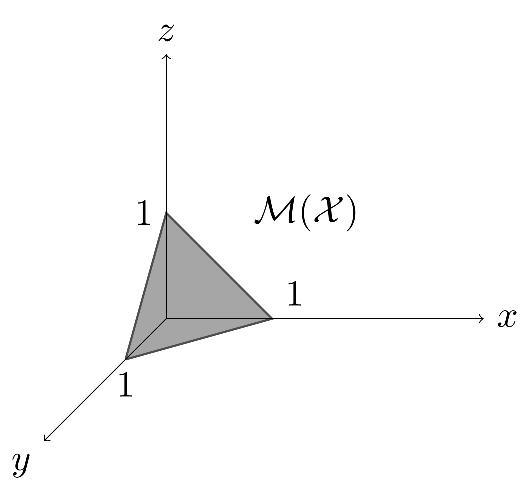

Figure 1: Set of pmf vectors in $\mathbb{R}^3$

We can look more closely at this equilateral triangle representing $\mathcal{M}(\mathcal{X})$.

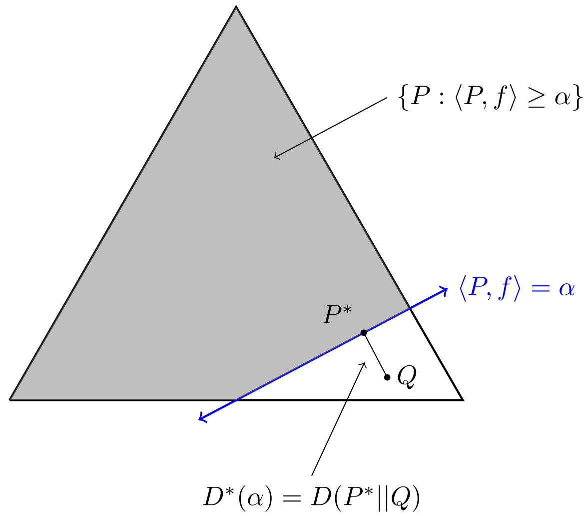

Figure 2: Set of possible pmfs $\mathcal{M}(\mathcal{X})$

The slope of the line $\langle P, f\rangle = \alpha$, shown above in blue, is determined by $f \in \mathbb{R}^3$, and the offset is determined by $\alpha \in \mathbb{R}$. We look for the point P^* in the feasible set (in gray) that is closest to Q under relative entropy, i.e. $D^*(\alpha) = D(P^*||Q)$. Note that a larger α will shrink the feasible set by moving the line in blue upwards. Thus, P^* will be further from Q, implying that the event in question has smaller probability.

By the LLN:

$$\Pr\left(\frac{1}{n}\sum_{i=1}^{n} f(X_i) \approx \langle Q, f\rangle\right) \approx 1.$$

In other words, this sum will be very close to the expected value of f under Q. We can conclude

$$\Pr\left(\frac{1}{n}\sum_{i=1}^{n} f(X_i) \geq \alpha\right)$$

is non-decaying for all $\alpha \leq \langle Q, f\rangle$, as the probability will go to 1 (the exponential decay rate is 0). Geometrically, this corresponds to a α such that Q is already in the feasible region, so $D(P^*||Q) = 0$ for $\alpha \leq \langle Q, f\rangle$.

On the other hand, if $\alpha > \langle Q, f\rangle$, we know that the probability will vanish. Sanov's Theorem tells us that it will vanish very (exponentially) rapidly and characterizes the exponent.

3.2 Example

Let X_i iid $\sim \text{Ber}(\frac{1}{2})$. We wish to find the exponential behavior of the probability that the fraction of 1's generated exceeds some level α:

$$\Pr\left(\frac{1}{n}\sum_{i=1}^{n} X_i \geq \alpha\right)$$

By LLN, if $\alpha \leq \frac{1}{2}$, this probability goes to 1, and if $1 \geq \alpha > \frac{1}{2}$, the probability is vanishing. However, we do not know how fast. Finally if $\alpha > 1$, the probability is 0, so the associated exponent is infinite.

By Sanov's Theorem applied to $Q = \text{Ber}(\frac{1}{2}),\ f(0) = 0,\ f(1) = 1$,

$$\Pr\left(\frac{1}{n}\sum_{i=1}^{n} X_i \geq \alpha\right) = \Pr\left(\frac{1}{n}\sum_{i=1}^{n} f(X_i) \geq \alpha\right) \doteq 2^{-nD^*(\alpha)}$$

where

$$\begin{aligned}
D^*(\alpha) &\stackrel{a}{=} \min_{0\leq p\leq 1,\ \langle \text{Ber}(p), f\rangle \geq \alpha} D\left(\text{Ber}(p) \mid\mid \text{Ber}\left(\tfrac{1}{2}\right)\right) \\
&\stackrel{b}{=} \min_{0\leq p\leq 1,\ p\geq\alpha} D\left(\text{Ber}(p) \mid\mid \text{Ber}\left(\tfrac{1}{2}\right)\right) \\
&= \min_{\alpha\leq p\leq 1} D\left(\text{Ber}(p) \mid\mid \text{Ber}\left(\tfrac{1}{2}\right)\right) \\
&= \begin{cases} 0 & \alpha \leq \frac{1}{2} \\ D\left(\text{Ber}(p) \mid\mid \text{Ber}\left(\frac{1}{2}\right)\right) & \frac{1}{2} < \alpha \leq 1 \\ \infty & 1 < \alpha \end{cases} \\
&= \begin{cases} 0 & \alpha \leq \frac{1}{2} \\ \alpha \log \frac{\alpha}{\frac{1}{2}} + (1-\alpha)\log \frac{1-\alpha}{\frac{1}{2}} & \frac{1}{2} < \alpha \leq 1 \\ \infty & 1 < \alpha \end{cases} \\
D^*(\alpha) &= \begin{cases} 0 & \alpha \leq \frac{1}{2} \\ 1 - h_2(\alpha) & \frac{1}{2} < \alpha \leq 1 \\ \infty & 1 < \alpha \end{cases}
\end{aligned}$$

(a) follow from the fact that any binary distribution P can be written as a $\text{Ber}(p)$ distribution for some p.
(b) follows from the fact that $\langle \text{Ber}(p), f\rangle = (1-p)f(0) + pf(1) = 0 + p = p$.

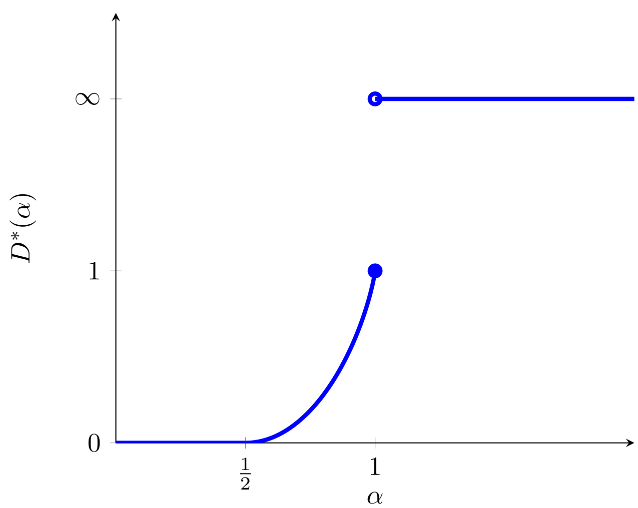

Figure 3: Plot of $D^*(\alpha)$, the exponential rate of decay, for the example of a $\text{Ber}(\frac{1}{2})$ source.

We note that this is consistent with our intuition from LLN:

- $\alpha \leq \frac{1}{2} \Rightarrow \Pr(\cdot) \to 1$ (exponential rate of decay is 0)
- $\frac{1}{2} < \alpha \leq 1 \Rightarrow \Pr(\cdot) \to 0$ (exponential rate of decay)
- $1 < \alpha \leq 1 \Rightarrow \Pr(\cdot) = 0$ (exponential rate of decay is ∞)

3.3 Proof of Sanov's Theorem

First we note

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} f(x_i) &= \frac{1}{n}\sum_{a\in\mathcal{X}} N(a|x^n)f(a) \\ &= \sum_{a\in\mathcal{X}} P_{x^n}(a)f(a) \quad (\text{ since } P_{x^n}(a) = \frac{N(a|x^n)}{n} \;) \\ &= \langle P_{x^n}, f\rangle\end{aligned}$$

Now since $Q(T(P)) = Q(\{x^n : P_{x^n} = P\}) = Pr(\{x^n : P_{x^n} = P\})$ we have

$$\Pr\left(\frac{1}{n}\sum_{i=1}^{n} f(X_i) \geq \alpha\right) = \sum_{P\in\mathbb{P}_n:\langle P,f\rangle\geq\alpha} Q(T(P))$$

Upper Bound:

$$\begin{aligned}\sum_{P\in\mathbb{P}_n:\langle P,f\rangle\geq\alpha} Q(T(P)) &\leq |\mathbb{P}_n| \max_{P\in\mathbb{P}_n:\langle P,f\rangle\geq\alpha} Q(T(P)) \\ &\leq (n+1)^{|\mathcal{X}|-1} \max_{P\in\mathbb{P}_n:\langle P,f\rangle\geq\alpha} 2^{-nD(P||Q)} \\ &= (n+1)^{|\mathcal{X}|-1} 2^{-n\min_{P\in\mathbb{P}_n:\langle P,f\rangle\geq\alpha} D(P||Q)} \\ &= (n+1)^{|\mathcal{X}|-1} 2^{-nD_n^*(\alpha)}\end{aligned}$$

Lower Bound:

$$\begin{aligned}\sum_{P\in\mathbb{P}_n:\langle P,f\rangle\geq\alpha} Q(T(P)) &\geq \max_{P\in\mathbb{P}_n:\langle P,f\rangle\geq\alpha} Q(T(P)) \\ &\geq \max_{P\in\mathbb{P}_n:\langle P,f\rangle\geq\alpha} \frac{1}{(n+1)^{|\mathcal{X}|-1}} 2^{-nD(P||Q)} \\ &= \frac{1}{(n+1)^{|\mathcal{X}|-1}} 2^{-nD_n^*(\alpha)}\end{aligned}$$

Q.E.D

3.4 A more general Sanov's Theorem

For X_i iid $\sim Q$ and $S \subset \mathcal{M}(\mathcal{X})$

$$Pr(\text{empirical distribution of } X^n \in \mathcal{S}) \doteq 2^{-n\min_{P\in\mathcal{S}} D(P||Q)}$$

Comment: This follows because, among the polynomially many terms in the expression for the probability (each of which decays exponentially with n), the largest term (one that is closest to Q) will dominate, and this term will be the one with the smallest exponent, i.e., $2^{-n\min_{P\in\mathcal{S}} D(P||Q)}$.