Information Theory and Related Fileds

Lecture 4: Large Deviations Theory

Lei Yu

Nankai University

Online Short Course at Beijing Normal University

Outline

Outline

1. Background and Cramer's Theorem
2. Proof of Cramer's Theorem
3. Extensions

Recall: Interpretations of H and I

Fact: Under product distribution $P_S^{\otimes n}$, the smallest high-probability sets (e.g., typical sets) are of size roughly $2^{nH(S)}$.

Recall: Interpretations of H and I

Fact: Under product distribution $P_S^{\otimes n}$, the smallest high-probability sets (e.g., typical sets) are of size roughly $2^{nH(S)}$.

So, $H(S)$ is the optimal rate in the almost lossless source coding.

Recall: Interpretations of H and I

Fact: Under product distribution $P_S^{\otimes n}$, the smallest high-probability sets (e.g., typical sets) are of size roughly $2^{nH(S)}$.

So, $H(S)$ is the optimal rate in the almost lossless source coding.

Fact: In the approximate "sphere" covering/packing problem, the optimal number of small "spheres" is roughly $2^{nI(X;Y)} = \frac{2^{nH(Y)}}{2^{nH(Y|X)}}$ (i.e., the volume ratio).

Recall: Interpretations of H and I

Fact: Under product distribution $P_S^{\otimes n}$, the smallest high-probability sets (e.g., typical sets) are of size roughly $2^{nH(S)}$.

So, $H(S)$ is the optimal rate in the almost lossless source coding.

Fact: In the approximate "sphere" covering/packing problem, the optimal number of small "spheres" is roughly $2^{nI(X;Y)} = \frac{2^{nH(Y)}}{2^{nH(Y|X)}}$ (i.e., the volume ratio).

So, $I(X;Y)$ is used in characterizing the optimal rate in the lossy source coding or channel coding.

Recall: Interpretations of H and I

Fact: Under product distribution $P_S^{\otimes n}$, the smallest high-probability sets (e.g., typical sets) are of size roughly $2^{nH(S)}$.

So, $H(S)$ is the optimal rate in the almost lossless source coding.

Fact: In the approximate "sphere" covering/packing problem, the optimal number of small "spheres" is roughly $2^{nI(X;Y)} = \frac{2^{nH(Y)}}{2^{nH(Y|X)}}$ (i.e., the volume ratio).

So, $I(X;Y)$ is used in characterizing the optimal rate in the lossy source coding or channel coding.

Question: What is the intuitive interpretation of the relative entropy?

Recall: Interpretations of H and I

Fact: Under product distribution $P_S^{\otimes n}$, the smallest high-probability sets (e.g., typical sets) are of size roughly $2^{nH(S)}$.

So, $H(S)$ is the optimal rate in the almost lossless source coding.

Fact: In the approximate "sphere" covering/packing problem, the optimal number of small "spheres" is roughly $2^{nI(X;Y)} = \frac{2^{nH(Y)}}{2^{nH(Y|X)}}$ (i.e., the volume ratio).

So, $I(X;Y)$ is used in characterizing the optimal rate in the lossy source coding or channel coding.

Question: What is the intuitive interpretation of the relative entropy?

It is the rate function in the large deviations theory

What Is the Large Deviations Theory

Let $X_i \sim P$, $i = 1, 2, \ldots$ be i.i.d. real-valued random variables with mean μ.

What Is the Large Deviations Theory

Let $X_i \sim P$, $i = 1, 2, \ldots$ be i.i.d. real-valued random variables with mean μ.

Theorem ((Weak) law of large numbers (LLN))

For any $b > 0$,

$$\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geq b\right\} \to 0 \ as \ n \to \infty.$$

What Is the Large Deviations Theory

Let $X_i \sim P$, $i = 1, 2, \ldots$ be i.i.d. real-valued random variables with mean μ.

Theorem ((Weak) law of large numbers (LLN))

For any $b > 0$,

$$\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geq b\right\} \to 0 \ as\ n \to \infty.$$

The convergence above is usually exponentially fast.

What Is the Large Deviations Theory

Let $X_i \sim P$, $i = 1, 2, \ldots$ be i.i.d. real-valued random variables with mean μ.

Theorem ((Weak) law of large numbers (LLN))

For any $b > 0$,

$$\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geq b\right\} \to 0 \ as \ n \to \infty.$$

The convergence above is usually exponentially fast.

Characterizing the convergence rate? —— Large Deviations Theory

What Is the Large Deviations Theory

- WLOG, we study the convergence rate of

$$\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{n} X_i \geq \alpha\right\}$$

where α can be seen as $\mu + b$. By substitution $X_i \leftarrow -X_i$ and rechoose α, we obtain the probability of "$\leq$" part.

What Is the Large Deviations Theory

- WLOG, we study the convergence rate of

$$\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{n} X_i \geq \alpha\right\}$$

  where α can be seen as $\mu + b$. By substitution $X_i \leftarrow -X_i$ and rechoose α, we obtain the probability of "$\leq$" part.

- That is, we are going to characterize

$$\lim_{n\to\infty} -\frac{1}{n}\log \mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{n} X_i \geq \alpha\right\}$$

What Is the Large Deviations Theory

- WLOG, we study the convergence rate of

$$\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{n} X_i \geq \alpha\right\}$$

  where α can be seen as $\mu + b$. By substitution $X_i \leftarrow -X_i$ and rechoose α, we obtain the probability of "$\leq$" part.

- That is, we are going to characterize

$$\lim_{n\to\infty} -\frac{1}{n}\log \mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{n} X_i \geq \alpha\right\}$$

- For convenience, we use the notation $a_n \doteq b_n$ to denote $a_n = 2^{o(n)} b_n$ (i.e., they share the same exponential rate).

Binary Example

- Let $X_i \sim \mathrm{Bern}(p), i = 1, 2, \ldots$ Then, $S_n := \sum_{i=1}^{n} X_i$ follows the (n, p)-binomial distribution.

Binary Example

- Let $X_i \sim \mathrm{Bern}(p), i = 1, 2, \ldots$ Then, $S_n := \sum_{i=1}^n X_i$ follows the (n, p)-binomial distribution.
- Hence, the probability of interest is

$$\mathbb{P}\{S_n \geq n\alpha\} = \sum_{i \geq n\alpha} \binom{n}{i} p^i (1-p)^{n-i} \doteq \max_{i \geq n\alpha} \binom{n}{i} p^i (1-p)^{n-i}.$$

Binary Example

- Let $X_i \sim \mathrm{Bern}(p), i = 1, 2, \ldots$ Then, $S_n := \sum_{i=1}^n X_i$ follows the (n, p)-binomial distribution.
- Hence, the probability of interest is

$$\mathbb{P}\{S_n \geq n\alpha\} = \sum_{i \geq n\alpha} \binom{n}{i} p^i (1-p)^{n-i} \doteq \max_{i \geq n\alpha} \binom{n}{i} p^i (1-p)^{n-i}.$$

- Denoting $q := i/n$, we find that

$$\begin{aligned} \binom{n}{i} &= \frac{n!}{i!(n-i)!} \doteq 2^{nH_2(q)} \qquad (\text{Stirling formula } n! \doteq \left(\frac{n}{e}\right)^n), \\ p^i(1-p)^{n-i} &= 2^{n(q\log p + (1-q)\log(1-p))}. \end{aligned}$$

Binary Example

- Let $X_i \sim \mathrm{Bern}(p), i = 1, 2, \ldots$ Then, $S_n := \sum_{i=1}^n X_i$ follows the (n, p)-binomial distribution.
- Hence, the probability of interest is

$$\mathbb{P}\{S_n \geq n\alpha\} = \sum_{i \geq n\alpha} \binom{n}{i} p^i (1-p)^{n-i} \doteq \max_{i \geq n\alpha} \binom{n}{i} p^i (1-p)^{n-i}.$$

- Denoting $q := i/n$, we find that

$$\binom{n}{i} = \frac{n!}{i!(n-i)!} \doteq 2^{nH_2(q)} \qquad \text{(Stirling formula } n! \doteq \left(\frac{n}{e}\right)^n\text{)},$$
$$p^i(1-p)^{n-i} = 2^{n(q\log p + (1-q)\log(1-p))}.$$

- Combining all above yields

$$\mathbb{P}\{S_n \geq n\alpha\} \doteq \max_{q \geq \alpha} 2^{-nD_2(q\|p)},$$

where binary relative entropy function $D_2(q\|p) = q\log\frac{q}{p} + (1-q)\log\frac{1-q}{1-p}$.

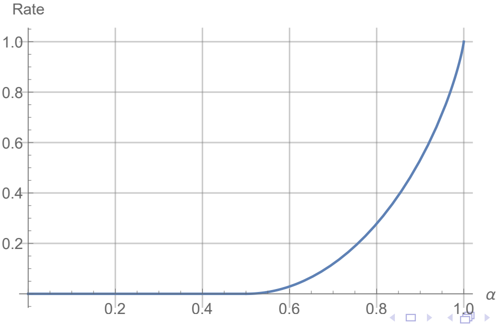
Binary Example (cont.)

So, the exponential rate for $\mathrm{Bern}(p)$ is

$$\min_{q\geq\alpha} D_2(q\|p) = \begin{cases} D_2(\alpha\|p) & \alpha > p \\ 0 & \alpha \leq p \end{cases}.$$

In particular, for $p = \frac{1}{2}$, the rate is

$$\min_{q\geq\alpha} D_2(q\|\frac{1}{2}) = \begin{cases} 1 - H_2(\alpha) & \alpha > \frac{1}{2} \\ 0 & \alpha \leq \frac{1}{2} \end{cases}$$

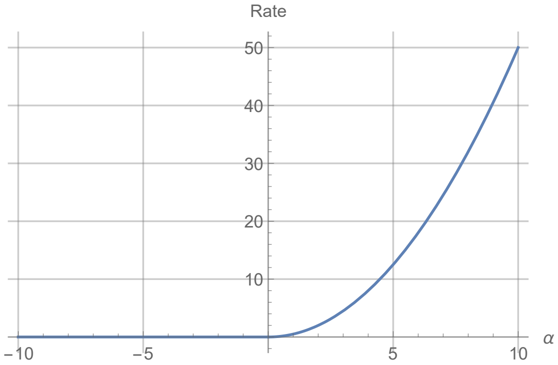


Gaussian Example

Let $X_i \sim \mathcal{N}(0,1), i = 1, 2, \ldots$ Then, $\bar{S}_n := \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i \sim \mathcal{N}(0,1)$. For $\alpha > 0$,

$$\mathbb{P}\left\{\bar{S}_n \geq \sqrt{n}\alpha\right\} = \int_{\sqrt{n}\alpha}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \mathrm{d}x \doteq e^{-n\alpha^2/2}$$

So, the exponential rate for $\mathcal{N}(0,1)$ is $\frac{\alpha^2}{2}$.



General Result

Extension to other distributions?

General Result

Extension to other distributions?

Theorem (Cramer's Theorem)

Let $X_i \sim P, i = 1, 2, \ldots$ *be i.i.d. real-valued random variables. For any* $\alpha \in \mathbb{R}$,

$$\lim_{n\to\infty} -\frac{1}{n} \log \mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{n} X_i \geq \alpha\right\} = \gamma_+(\alpha),$$

where

$$\gamma_+(\alpha) := \inf_{Q:\mathbb{E}_Q[X]\geq\alpha} D(Q\|P).$$

Extension to other distributions?

Theorem (Cramer's Theorem)

Let $X_i \sim P, i = 1, 2, \ldots$ be i.i.d. real-valued random variables. For any $\alpha \in \mathbb{R}$,

$$\lim_{n\to\infty} -\frac{1}{n} \log \mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{n} X_i \geq \alpha\right\} = \gamma_+(\alpha),$$

where

$$\gamma_+(\alpha) := \inf_{Q:\mathbb{E}_Q[X]\geq\alpha} D(Q\|P).$$

- For $\mathrm{Bern}(p)$, it is easily seen that $\gamma_+(\alpha) = \min_{q\geq\alpha} D_2(q\|p)$.

General Result

Extension to other distributions?

Theorem (Cramer's Theorem)

Let $X_i \sim P, i = 1, 2, \ldots$ be i.i.d. real-valued random variables. For any $\alpha \in \mathbb{R}$,

$$\lim_{n\to\infty} -\frac{1}{n} \log \mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{n} X_i \geq \alpha\right\} = \gamma_+(\alpha),$$

where

$$\gamma_+(\alpha) := \inf_{Q:\mathbb{E}_Q[X]\geq\alpha} D(Q\|P).$$

- For $\mathrm{Bern}(p)$, it is easily seen that $\gamma_+(\alpha) = \min_{q\geq\alpha} D_2(q\|p)$.
- For $\mathcal{N}(0,1)$, the optimal Q attaining $\gamma_+(\alpha)$ is $\mathcal{N}(\alpha,1)$ for $\alpha > 0$, and hence, $\gamma_+(\alpha) = D(\mathcal{N}(\alpha,1)\|\mathcal{N}(0,1)) = \frac{\alpha^2}{2}$.

General Result

Extension to other distributions?

Theorem (Cramer's Theorem)

Let $X_i \sim P, i = 1, 2, \ldots$ be i.i.d. real-valued random variables. For any $\alpha \in \mathbb{R}$,

$$\lim_{n\to\infty} -\frac{1}{n}\log \mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{n} X_i \geq \alpha\right\} = \gamma_+(\alpha),$$

where

$$\gamma_+(\alpha) := \inf_{Q:\mathbb{E}_Q[X]\geq\alpha} D(Q\|P).$$

- For $\mathrm{Bern}(p)$, it is easily seen that $\gamma_+(\alpha) = \min_{q\geq\alpha} D_2(q\|p)$.
- For $\mathcal{N}(0,1)$, the optimal Q attaining $\gamma_+(\alpha)$ is $\mathcal{N}(\alpha,1)$ for $\alpha > 0$, and hence, $\gamma_+(\alpha) = D(\mathcal{N}(\alpha,1)\|\mathcal{N}(0,1)) = \frac{\alpha^2}{2}$.
- This theorem gives an intuitive interpretation of relative entropy.

Remarks on $\gamma_+(\alpha)$

$$\gamma_+(\alpha) := \inf_{Q:\mathbb{E}_Q[X]\geq\alpha} D(Q\|P).$$

- The optimization above is known as the information projection problem (convex optimization problem).

Remarks on $\gamma_+(\alpha)$

$$\gamma_+(\alpha) := \inf_{Q:\mathbb{E}_Q[X]\geq\alpha} D(Q\|P).$$

- The optimization above is known as the information projection problem (convex optimization problem).
- For discrete P on finite alphabet, there always exists a unique Q^* attaining $\gamma_+(\alpha)$, and moreover, it is of form

$$Q^*(x) = \frac{P(x)e^{\lambda x}}{\mathbb{E}_P[e^{\lambda X}]} \text{ for some } \lambda \geq 0.$$

Remarks on $\gamma_+(\alpha)$

$$\gamma_+(\alpha) := \inf_{Q:\mathbb{E}_Q[X]\geq\alpha} D(Q\|P).$$

- The optimization above is known as the information projection problem (convex optimization problem).
- For discrete P on finite alphabet, there always exists a unique Q^* attaining $\gamma_+(\alpha)$, and moreover, it is of form

$$Q^*(x) = \frac{P(x)e^{\lambda x}}{\mathbb{E}_P[e^{\lambda X}]} \text{ for some } \lambda \geq 0.$$

- The rate function $\gamma_+(\alpha)$ admits the dual formula:

$$\gamma_+(\alpha) = \sup_{\lambda\geq 0} \lambda\alpha - \ln \mathbb{E}_P[e^{\lambda X}].$$

Geometry of Information Projection

- Interesting geometric property ("Pythagorean" theorem): For any $R \in A := \left\{Q : \mathbb{E}_Q [X] \geq \alpha\right\}$,

$$D(R\|P) \geq D(R\|Q^*) + D(Q^*\|P),$$

with equality iff R is on the tangent line.

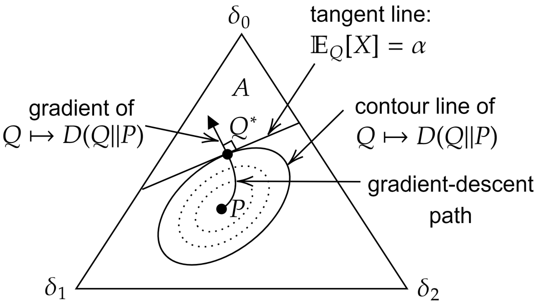
Geometry of Information Projection

- Interesting geometric property ("Pythagorean" theorem): For any $R \in A := \left\{Q : \mathbb{E}_Q[X] \geq \alpha\right\}$,

$$D(R\|P) \geq D(R\|Q^*) + D(Q^*\|P),$$

  with equality iff R is on the tangent line.

- Example: $\mathcal{X} = \{0, 1, 2\}$, any distribution on $\mathcal{X}$ is of form $Q = p_0\delta_0 + p_1\delta_1 + p_2\delta_2$ with nonnegative p_i whose sum is 1.



Outline

Recall: Properties of D

- Chain rule:

$$D(P_{X^n}\|Q_{X^n})=\sum_{i=1}^{n}D(P_{X_i|X^{i-1}}\|Q_{X_i|X^{i-1}}|P_{X^{i-1}})$$

Recall: Properties of D

- Chain rule:

$$D(P_{X^n}\|Q_{X^n}) = \sum_{i=1}^{n} D(P_{X_i|X^{i-1}}\|Q_{X_i|X^{i-1}}|P_{X^{i-1}})$$

- Nonnegativity:

$$D(P\|Q) \geq 0 \text{ with equality iff } P = Q$$

Recall: Properties of D

- Chain rule:

$$D(P_{X^n}\|Q_{X^n})=\sum_{i=1}^{n} D(P_{X_i|X^{i-1}}\|Q_{X_i|X^{i-1}}|P_{X^{i-1}})$$

- Nonnegativity:

$$D(P\|Q)\geq 0 \text{ with equality iff } P=Q$$

- Consequences of chain rule and nonnegativity:

$$D(P_{X^n}\|Q_X^{\otimes n})\geq \sum_{i=1}^{n} D(P_{X_i}\|Q_X) \text{ with equality iff } P_{X^n}=\Pi P_{X_i} \qquad \text{(superadditivity)}$$

$$D(P_{XY}\|Q_{XY}) \geq \max\{D(P_Y\|Q_Y)\, D(P_X\|Q_X)\}$$

$$(P,Q)\mapsto D(P\|Q) \text{ is convex.}$$

Proof of "$\geq$"

- We first prove the "$\geq$" part by using properties of relative entropy.

Proof of "$\geq$"

- We first prove the "$\geq$" part by using properties of relative entropy.
- We denote $A_n := \left\{x^n : \frac{1}{n}\sum_{i=1}^n x_i \geq \alpha\right\}$. Then, the probability of interest is $P^{\otimes n}(A_n)$.

Proof of "≥"

- We first prove the "≥" part by using properties of relative entropy.
- We denote $A_n := \left\{x^n : \frac{1}{n}\sum_{i=1}^n x_i \geq \alpha\right\}$. Then, the probability of interest is $P^{\otimes n}(A_n)$.
- Define an auxiliary distribution $Q_{X^n} := P^{\otimes n}(\cdot|A_n)$ (conditional distribution given A_n).

Proof of "$\geq$"

- We first prove the "$\geq$" part by using properties of relative entropy.
- We denote $A_n := \left\{x^n : \frac{1}{n}\sum_{i=1}^n x_i \geq \alpha\right\}$. Then, the probability of interest is $P^{\otimes n}(A_n)$.
- Define an auxiliary distribution $Q_{X^n} := P^{\otimes n}(\cdot|A_n)$ (conditional distribution given A_n).
- Fact: By symmetry, the marginals Q_{X_i} are the same for $i \in [n]$, denoted by $\bar{Q}$.

Proof of "$\geq$"

- We first prove the "$\geq$" part by using properties of relative entropy.
- We denote $A_n := \left\{x^n : \frac{1}{n}\sum_{i=1}^n x_i \geq \alpha\right\}$. Then, the probability of interest is $P^{\otimes n}(A_n)$.
- Define an auxiliary distribution $Q_{X^n} := P^{\otimes n}(\cdot|A_n)$ (conditional distribution given A_n).
- Fact: By symmetry, the marginals Q_{X_i} are the same for $i \in [n]$, denoted by $\bar{Q}$.
- Conclusion 1:

$$
\begin{aligned}
-\frac{1}{n}\log P^{\otimes n}(A_n) &= \frac{1}{n}D(Q_{X^n}\|P^{\otimes n}) && \frac{Q_{X^n}(x^n)}{P^{\otimes n}(x^n)} = \frac{1}{P^{\otimes n}(A_n)}, \forall x^n \in A_n \\
&\geq \frac{1}{n}\sum_{i=1}^n D(Q_{X_i}\|P) && \text{superadditivity} \\
&= D(\bar{Q}\|P) && Q_{X_i} \text{ are the same}
\end{aligned}
$$

Proof of "≥" (cont.)

- By definition of Q_{X^n}, $Q_{X^n}(A_n) = 1$, i.e.,

$$\frac{1}{n}\sum_{i=1}^{n} X_i \geq \alpha \qquad (Q_{X^n}\text{-a.s.})$$

Proof of "≥" (cont.)

- By definition of Q_{X^n}, $Q_{X^n}(A_n) = 1$, i.e.,

$$\frac{1}{n}\sum_{i=1}^{n} X_i \geq \alpha \qquad (Q_{X^n}\text{-a.s.})$$

- Conclusion 2:

$$\begin{aligned}
\alpha &\leq \mathbb{E}_{Q_{X^n}}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] && \text{take expectation for the inequality above}\\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{Q_{X_i}}[X_i] && \text{swap } \sum \text{ and } \mathbb{E}\\
&= \mathbb{E}_{\bar{Q}}[X] && Q_{X_i} \text{ are the same}
\end{aligned}$$

Proof of "≥" (cont.)

- By definition of Q_{X^n}, $Q_{X^n}(A_n) = 1$, i.e.,

$$\frac{1}{n}\sum_{i=1}^{n} X_i \geq \alpha \qquad (Q_{X^n}\text{-a.s.})$$

- Conclusion 2:

$$\begin{aligned}
\alpha &\leq \mathbb{E}_{Q_{X^n}}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] && \text{take expectation for the inequality above} \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{Q_{X_i}}\left[X_i\right] && \text{swap } \sum \text{ and } \mathbb{E} \\
&= \mathbb{E}_{\bar{Q}}\left[X\right] && Q_{X_i} \text{ are the same}
\end{aligned}$$

- Combining Conclusions 1 and 2:

$$-\frac{1}{n}\log P^{\otimes n}(A_n) \overset{\text{Con 1}}{\geq} D(\bar{Q}\|P) \overset{\text{Con 2}}{\geq} \inf_{Q:\mathbb{E}_Q[X]\geq\alpha} D(Q\|P) = \gamma_+(\alpha).$$

Proof of "$\leq$"

We now turn to proving the "$\leq$" part.

Proof of "$\leq$"

We now turn to proving the "$\leq$" part.

Fact

$$-\log P^{\otimes n}(A_n) = \inf_{Q_{X^n}:Q_{X^n}(A_n)=1} D(Q_{X^n}\|P^{\otimes n}), \tag{1}$$

where the optimal Q_{X^n} attaining the infimum is $P^{\otimes n}(\cdot|A_n)$.

Proof of "$\leq$"

We now turn to proving the "$\leq$" part.

Fact

$$-\log P^{\otimes n}(A_n) = \inf_{Q_{X^n}:Q_{X^n}(A_n)=1} D(Q_{X^n} \| P^{\otimes n}), \tag{1}$$

where the optimal Q_{X^n} attaining the infimum is $P^{\otimes n}(\cdot|A_n)$.

- This fact follows since for any Q_{X^n} concentrated on A_n,

$$D(Q_{X^n} \| P^{\otimes n}) = D(Q_{X^n} \| P^{\otimes n}(\cdot|A_n)) - \log P^{\otimes n}(A_n) \geq -\log P^{\otimes n}(A_n).$$

Proof of "$\leq$"

We now turn to proving the "$\leq$" part.

Fact

$$-\log P^{\otimes n}(A_n) = \inf_{Q_{X^n}:Q_{X^n}(A_n)=1} D(Q_{X^n}\|P^{\otimes n}), \tag{1}$$

where the optimal Q_{X^n} attaining the infimum is $P^{\otimes n}(\cdot|A_n)$.

- This fact follows since for any Q_{X^n} concentrated on A_n,

$$D(Q_{X^n}\|P^{\otimes n}) = D(Q_{X^n}\|P^{\otimes n}(\cdot|A_n)) - \log P^{\otimes n}(A_n) \geq -\log P^{\otimes n}(A_n).$$

- By this fact, to upper bound LHS of (1), it suffices to construct a feasible solution to RHS of (1).

Proof of "$\leq$" (cont.)

Constructing feasible solution:

- Let Q be such that $\mathbb{E}_Q[X] > \alpha$

Proof of "$\leq$" (cont.)

Constructing feasible solution:

- Let Q be such that $\mathbb{E}_Q[X] > \alpha$
- Denote auxiliary probability measures $R_{X^n} := Q^{\otimes n}(\cdot|A_n)$ and $\bar{R}_{X^n} := Q^{\otimes n}(\cdot|A_n^c)$.

Proof of "$\leq$" (cont.)

Constructing feasible solution:

- Let Q be such that $\mathbb{E}_Q[X] > \alpha$
- Denote auxiliary probability measures $R_{X^n} := Q^{\otimes n}(\cdot|A_n)$ and $\bar{R}_{X^n} := Q^{\otimes n}(\cdot|A_n^c)$.
- Then, R_{X^n} is a feasible solution to the infimization in (1).

Proof of "$\leq$" (cont.)

Constructing feasible solution:

- Let Q be such that $\mathbb{E}_Q[X] > \alpha$
- Denote auxiliary probability measures $R_{X^n} := Q^{\otimes n}(\cdot|A_n)$ and $\bar{R}_{X^n} := Q^{\otimes n}(\cdot|A_n^c)$.
- Then, R_{X^n} is a feasible solution to the infimization in (1).
- Conclusion 1:

$$-\log P^{\otimes n}(A_n) \leq D(R_{X^n} \| P^{\otimes n})$$

Proof of "$\leq$" (cont.)

Computing the limit of $\frac{1}{n}D(R_{X^n}\|P^{\otimes n})$:

- Conclusion 2: By LLN, $p_n := Q^{\otimes n}(A_n) \to 1$ (Recall $A_n := \left\{x^n : \frac{1}{n}\sum_{i=1}^{n} x_i \geq \alpha\right\}$)

Proof of "$\leq$" (cont.)

Computing the limit of $\frac{1}{n}D(R_{X^n}\|P^{\otimes n})$:

- Conclusion 2: By LLN, $p_n := Q^{\otimes n}(A_n) \to 1$ (Recall $A_n := \left\{x^n : \frac{1}{n}\sum_{i=1}^n x_i \geq \alpha\right\}$)
- Conclusion 3:
$$D(R_{X^n}\|P^{\otimes n}) \leq \frac{nD(Q\|P)+1}{p_n}.$$

Proof of "$\leq$" (cont.)

Computing the limit of $\frac{1}{n}D(R_{X^n}\|P^{\otimes n})$:

- Conclusion 2: By LLN, $p_n := Q^{\otimes n}(A_n) \to 1$ (Recall $A_n := \left\{x^n : \frac{1}{n}\sum_{i=1}^n x_i \geq \alpha\right\}$)
- Conclusion 3:
$$D(R_{X^n}\|P^{\otimes n}) \leq \frac{nD(Q\|P)+1}{p_n}.$$
- Conclusion 3 follows since

$$\begin{aligned} nD(Q\|P) &= D(Q^{\otimes n}\|P^{\otimes n}) \\ &= p_n D(R_{X^n}\|P^{\otimes n}) + (1-p_n)D(\bar{R}_{X^n}\|P^{\otimes n}) - H_2(p_n) \quad \text{by definition} \\ &\geq p_n D(R_{X^n}\|P^{\otimes n}) - 1. \end{aligned}$$

Proof of "≤" (cont.)

- Combining Conclusions 1-3, and letting $n \to \infty$, we obtain

$$\begin{aligned}\limsup_{n\to\infty} -\frac{1}{n}\log P^{\otimes n}(A_n) &\overset{\text{Con 1\&3}}{\leq} \limsup_{n\to\infty} \frac{nD(Q\|P)+1}{np_n}\\ &\overset{\text{Con 2}}{=} D(Q\|P).\end{aligned}$$

Proof of "$\leq$" (cont.)

- Combining Conclusions 1-3, and letting $n \to \infty$, we obtain

$$\limsup_{n\to\infty} -\frac{1}{n} \log P^{\otimes n}(A_n) \overset{\text{Con 1\&3}}{\leq} \limsup_{n\to\infty} \frac{nD(Q\|P)+1}{np_n} \overset{\text{Con 2}}{=} D(Q\|P).$$

- Since Q satisfying $\mathbb{E}_Q[X] > \alpha$ is arbitrary, we can choose an optimal Q:

$$\limsup_{n\to\infty} -\frac{1}{n} \log P^{\otimes n}(A_n) \leq \inf_{Q:\mathbb{E}_Q[X]>\alpha} D(Q\|P) \overset{\text{continuity}}{=} \gamma_+(\alpha)$$

Remarks

The proof given here is called the entropy method (or information-theoretic method), typically consisting of:

Remarks

The proof given here is called the entropy method (or information-theoretic method), typically consisting of:

1. First, introduce auxiliary probability measures, e.g.,

$$Q_{X^n} := P^{\otimes n}(\cdot|A_n)$$

Remarks

The proof given here is called the entropy method (or information-theoretic method), typically consisting of:

1. First, introduce auxiliary probability measures, e.g.,

$$Q_{X^n} := P^{\otimes n}(\cdot|A_n)$$

2. Then, express the problem in terms of relative entropies of these auxiliary probability measures, e.g.,

$$-\log P^{\otimes n}(A_n) = D(Q_{X^n}\|P^{\otimes n})$$

Remarks

The proof given here is called the entropy method (or information-theoretic method), typically consisting of:

1. First, introduce auxiliary probability measures, e.g.,

$$Q_{X^n} := P^{\otimes n}(\cdot|A_n)$$

2. Then, express the problem in terms of relative entropies of these auxiliary probability measures, e.g.,

$$-\log P^{\otimes n}(A_n) = D(Q_{X^n}\|P^{\otimes n})$$

3. Lastly, derive bounds by using properties of relative entropies.

Remarks

The proof given here is called the entropy method (or information-theoretic method), typically consisting of:

1. First, introduce auxiliary probability measures, e.g.,

$$Q_{X^n} := P^{\otimes n}(\cdot|A_n)$$

2. Then, express the problem in terms of relative entropies of these auxiliary probability measures, e.g.,

$$-\log P^{\otimes n}(A_n) = D(Q_{X^n}\|P^{\otimes n})$$

3. Lastly, derive bounds by using properties of relative entropies.

In contrast, common proofs of Cramer's theorem (e.g., [Dembo–Zeitouni's book]) are from the dual view (by using the Chernoff bound).

More Intuitions

Let Q^* attain $\gamma_+(\alpha) = \inf_{Q:\mathbb{E}_Q[X]\geq\alpha} D(Q\|P)$.

More Intuitions

Let Q^* attain $\gamma_+(\alpha) = \inf_{Q:\mathbb{E}_Q[X]\geq\alpha} D(Q\|P)$.

Our proof of "$\geq$" part starts with $Q_{X^n} := P^{\otimes n}(\cdot|A_n)$ but ends with Q^*

More Intuitions

Let Q^* attain $\gamma_+(\alpha) = \inf_{Q:\mathbb{E}_Q[X]\geq\alpha} D(Q\|P)$.

Our proof of "$\geq$" part starts with $Q_{X^n} := P^{\otimes n}(\cdot|A_n)$ but ends with Q^*

What is the relationship between Q_{X^n} and Q^*?

More Intuitions

Let Q^* attain $\gamma_+(\alpha) = \inf_{Q:\mathbb{E}_Q[X]\geq\alpha} D(Q\|P)$.

Our proof of "$\geq$" part starts with $Q_{X^n} := P^{\otimes n}(\cdot|A_n)$ but ends with Q^*

What is the relationship between Q_{X^n} and Q^*?

In fact, the marginals of Q_{X^n} satisfy $D(Q_{X_i}\|P) \to D(Q^*\|P)$ as $n \to \infty$, which, by "Pythagorean" theorem, further implies $D(Q_{X_i}\|Q^*) \to 0$ (known as Gibbs conditioning principle)

More Intuitions

Let Q^* attain $\gamma_+(\alpha) = \inf_{Q:\mathbb{E}_Q[X]\geq\alpha} D(Q\|P)$.

Our proof of "$\geq$" part starts with $Q_{X^n} := P^{\otimes n}(\cdot|A_n)$ but ends with Q^*

What is the relationship between Q_{X^n} and Q^*?

In fact, the marginals of Q_{X^n} satisfy $D(Q_{X_i}\|P) \to D(Q^*\|P)$ as $n \to \infty$, which, by "Pythagorean" theorem, further implies $D(Q_{X_i}\|Q^*) \to 0$ (known as Gibbs conditioning principle)

Similarly, in the proof of "$\leq$" part, roughly speaking, the optimal auxiliary measure is $R_{X^n} := Q^{*\otimes n}(\cdot|A_n)$ which satisfies $D(R_{X_i}\|Q^*) \to 0$

Outline

Extension to Any Open/Closed Sets

Let $\bar{S}_n := \frac{1}{n}\sum_{i=1}^n X_i$. Define

$$\gamma(\alpha) := \inf_{Q:\mathbb{E}_Q[X]=\alpha} D(Q\|P) \ = \sup_{\lambda\in\mathbb{R}} \lambda\alpha - \log \mathbb{E}_P[e^{\lambda X}] \text{ (dual formula)}.$$

Extension to Any Open/Closed Sets

Let $\bar{S}_n := \frac{1}{n}\sum_{i=1}^{n} X_i$. Define

$$\gamma(\alpha) := \inf_{Q:\mathbb{E}_Q[X]=\alpha} D(Q\|P) = \sup_{\lambda\in\mathbb{R}} \lambda\alpha - \log \mathbb{E}_P[e^{\lambda X}] \text{ (dual formula)}.$$

The simple version of Cramer's theorem can be easily extended to:

Theorem (Cramer's Theorem (General Version))

(a) For any closed set $F \subseteq \mathbb{R}$,

$$\liminf_{n\to\infty} -\frac{1}{n}\log \mathbb{P}(\bar{S}_n \in F) \geq \inf_{x\in F} \gamma(x).$$

(b) For any open set $G \subseteq \mathbb{R}$,

$$\limsup_{n\to\infty} -\frac{1}{n}\log \mathbb{P}(\bar{S}_n \in G) \leq \inf_{x\in G} \gamma(x).$$

In this case, the laws of $\bar{S}_n$ are said to satisfy the large deviations principle (LDP) with the rate function γ.

Extension to Strong Version

Theorem (Cramer's Theorem (Strong Version))

Let $\alpha > 0$. *Suppose that* Q^* *attain* $\gamma_+(\alpha)$. *Then,*

$$\mathbb{P}\left\{\sum_{i=1}^{n} X_i \geq n\alpha\right\} \sim \frac{c}{\sqrt{2\pi n \mathrm{V}(Q^*\|P)}} e^{-nD(Q^*\|P)},$$

where $c = 1$ *if* $X_i \sim P$ *are non-lattice, and* $c = \frac{\lambda^* d}{1-e^{-\lambda^* d}}$ *if* $X_i \sim P$ *are lattice with maximal step* d *and* $0 < \mathbb{P}(X_i = \alpha) < 1$.

Extension to Strong Version

Theorem (Cramer's Theorem (Strong Version))

Let $\alpha > 0$. Suppose that Q^* attain $\gamma_+(\alpha)$. Then,

$$\mathbb{P}\left\{\sum_{i=1}^{n} X_i \geq n\alpha\right\} \sim \frac{c}{\sqrt{2\pi n \mathrm{V}(Q^*\|P)}} e^{-nD(Q^*\|P)},$$

where $c = 1$ if $X_i \sim P$ are non-lattice, and $c = \frac{\lambda^* d}{1-e^{-\lambda^* d}}$ if $X_i \sim P$ are lattice with maximal step d and $0 < \mathbb{P}(X_i = \alpha) < 1$.

- The relative varentropy (relative dispersion) is $\mathrm{V}(Q\|P) = \mathrm{Var}_Q\left[\log \frac{\mathrm{d}Q}{\mathrm{d}P}(X)\right]$.

Extension to Strong Version

Theorem (Cramer's Theorem (Strong Version))

Let $\alpha > 0$. Suppose that Q^* attain $\gamma_+(\alpha)$. Then,

$$\mathbb{P}\left\{\sum_{i=1}^{n} X_i \geq n\alpha\right\} \sim \frac{c}{\sqrt{2\pi n \mathrm{V}(Q^*\|P)}} e^{-nD(Q^*\|P)},$$

where $c = 1$ if $X_i \sim P$ are non-lattice, and $c = \frac{\lambda^* d}{1-e^{-\lambda^* d}}$ if $X_i \sim P$ are lattice with maximal step d and $0 < \mathbb{P}(X_i = \alpha) < 1$.

- The relative varentropy (relative dispersion) is $\mathrm{V}(Q\|P) = \mathrm{Var}_Q\left[\log \frac{\mathrm{d}Q}{\mathrm{d}P}(X)\right]$.
- X is lattice, if for some x_0, d, the random variable $d^{-1}(X - x_0)$ is (a.s.) an integer number, and maximal step d is the largest number with this property.

Extension to Strong Version

Theorem (Cramer's Theorem (Strong Version))

Let $\alpha > 0$. Suppose that Q^* attain $\gamma_+(\alpha)$. Then,

$$\mathbb{P}\left\{\sum_{i=1}^{n} X_i \geq n\alpha\right\} \sim \frac{c}{\sqrt{2\pi n \mathrm{V}(Q^*\|P)}} e^{-nD(Q^*\|P)},$$

where $c = 1$ if $X_i \sim P$ are non-lattice, and $c = \frac{\lambda^* d}{1-e^{-\lambda^* d}}$ if $X_i \sim P$ are lattice with maximal step d and $0 < \mathbb{P}(X_i = \alpha) < 1$.

- The relative varentropy (relative dispersion) is $\mathrm{V}(Q\|P) = \mathrm{Var}_Q\left[\log \frac{\mathrm{d}Q}{\mathrm{d}P}(X)\right]$.
- X is lattice, if for some x_0, d, the random variable $d^{-1}(X - x_0)$ is (a.s.) an integer number, and maximal step d is the largest number with this property.
- E.g., lattice $X \in \{1, 2, 3, \dots\}$, and maximal step $d = 1$ if $P(1), P(2) > 0$; non-lattice: $P_X(\frac{1}{n}) > 0, \forall n$

Extension to LDP for Empirical Measures

Assume $\mathcal{X}$ is a finite set.

Extension to LDP for Empirical Measures

Assume $\mathcal{X}$ is a finite set.

The empirical measure L_{X^n} of a random vector X^n is given by

$$\mathsf{L}_{X^n}(a) = \frac{\#\text{ of } a \text{ in } X^n}{n}, \ \forall a \in \mathcal{X}.$$

Extension to LDP for Empirical Measures

Assume $\mathcal{X}$ is a finite set.

The empirical measure L_{X^n} of a random vector X^n is given by

$$\mathsf{L}_{X^n}(a) = \frac{\#\text{ of } a \text{ in } X^n}{n}, \ \forall a \in \mathcal{X}.$$

Denote $\mathcal{P}(\mathcal{X})$ as the set of probability measures on $\mathcal{X}$, i.e.,

$$\mathcal{P}(\mathcal{X}) = \{P : P(x) \geq 0, \sum_{x \in \mathcal{X}} P(x) = 1\}$$

Extension to LDP for Empirical Measures

Theorem (Sanov's Theorem)

The empirical measures L_{X^n} satisfy the LDP in $\mathcal{P}(\mathcal{X})$ with rate function $D(\cdot\|P)$. That is, (a) For any closed set $F \subseteq \mathcal{P}(\mathcal{X})$,

$$\liminf_{n\to\infty} -\frac{1}{n}\log\mathbb{P}(\mathsf{L}_{X^n}\in F) \geq \inf_{Q\in F} D(Q\|P).$$

(b) For any open set $G \subseteq \mathcal{P}(\mathcal{X})$,

$$\limsup_{n\to\infty} -\frac{1}{n}\log\mathbb{P}(\mathsf{L}_{X^n}\in G) \leq \inf_{Q\in G} D(Q\|P).$$

Extension to LDP for Empirical Measures

Theorem (Sanov's Theorem)

The empirical measures L_{X^n} *satisfy the LDP in* $\mathcal{P}(\mathcal{X})$ *with rate function* $D(\cdot\|P)$*. That is, (a) For any closed set* $F \subseteq \mathcal{P}(\mathcal{X})$*,*

$$\liminf_{n\to\infty} -\frac{1}{n}\log \mathbb{P}(\mathsf{L}_{X^n} \in F) \geq \inf_{Q\in F} D(Q\|P).$$

(b) For any open set $G \subseteq \mathcal{P}(\mathcal{X})$*,*

$$\limsup_{n\to\infty} -\frac{1}{n}\log \mathbb{P}(\mathsf{L}_{X^n} \in G) \leq \inf_{Q\in G} D(Q\|P).$$

The proof is similar to that of Cramer's Theorem (Simple Version), just replacing $A_n := \left\{\frac{1}{n}\sum_{i=1}^n X_i \geq \alpha\right\}$ with $A_n := \{\mathsf{L}_{X^n} \in F\}$ or $\{\mathsf{L}_{X^n} \in G\}$

Extension to LDP for Empirical Measures

Theorem (Sanov's Theorem)

The empirical measures L_{X^n} satisfy the LDP in $\mathcal{P}(\mathcal{X})$ with rate function $D(\cdot\|P)$. That is, (a) For any closed set $F \subseteq \mathcal{P}(\mathcal{X})$,

$$\liminf_{n\to\infty} -\frac{1}{n}\log \mathbb{P}(\mathsf{L}_{X^n} \in F) \geq \inf_{Q\in F} D(Q\|P).$$

(b) For any open set $G \subseteq \mathcal{P}(\mathcal{X})$,

$$\limsup_{n\to\infty} -\frac{1}{n}\log \mathbb{P}(\mathsf{L}_{X^n} \in G) \leq \inf_{Q\in G} D(Q\|P).$$

The proof is similar to that of Cramer's Theorem (Simple Version), just replacing $A_n := \left\{\frac{1}{n}\sum_{i=1}^n X_i \geq \alpha\right\}$ with $A_n := \{\mathsf{L}_{X^n} \in F\}$ or $\{\mathsf{L}_{X^n} \in G\}$

Sanov's theorem can be generalized to Polish spaces $\mathcal{X}$ with the weak topologies (including Euclidean and countable spaces).

Recover Cramer From Sanov

The empirical mean is determined by the empirical measure:

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \sum_{a\in\mathcal{X}} \frac{\#\text{ of } a \text{ in } X^n}{n}\, a = \mathbb{E}_{\mathsf{L}_{X^n}}[X]$$

Recover Cramer From Sanov

The empirical mean is determined by the empirical measure:

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \sum_{a\in\mathcal{X}} \frac{\#\text{ of } a \text{ in } X^n}{n}\, a = \mathbb{E}_{\mathsf{L}_{X^n}}[X]$$

So, $\left\{\frac{1}{n}\sum_{i=1}^{n} X_i \geq \alpha\right\} = \left\{\mathbb{E}_{\mathsf{L}_{X^n}}[X] \geq \alpha\right\}$.

Recover Cramer From Sanov

The empirical mean is determined by the empirical measure:

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \sum_{a\in\mathcal{X}} \frac{\#\text{ of } a \text{ in } X^n}{n} a = \mathbb{E}_{\mathsf{L}_{X^n}}[X]$$

So, $\left\{\frac{1}{n}\sum_{i=1}^{n} X_i \geq \alpha\right\} = \left\{\mathbb{E}_{\mathsf{L}_{X^n}}[X] \geq \alpha\right\}$.

By Sanov's theorem,

$$\lim_{n\to\infty} -\frac{1}{n}\log \mathbb{P}\left\{\mathbb{E}_{\mathsf{L}_{X^n}}[X] \geq \alpha\right\} = \inf_{Q:\mathbb{E}_Q[X]\geq\alpha} D(Q\|P) = \gamma_+(\alpha)$$

Various Deviations

Let X_i, $i = 1, 2, \ldots$ be i.i.d. real-valued r.v.'s with mean zero. Define the χ^2-divergence

$$\chi^2(Q\|P) := \sum_x P(x)\left(1 - \frac{Q(x)}{P(x)}\right)^2$$

	Event A_n	Asymptotics of $\mathbb{P}(A_n)$
Central Limit Theorem (Small Deviations)	$\sum_{i=1}^n X_i \geq \sqrt{n}\alpha$	Constant, $1 - \Phi(\frac{\alpha}{\sqrt{\mathrm{Var}(X)}})$
Moderate Deviations	$\sum_{i=1}^n X_i \geq n^\beta \alpha$, with $\alpha > 0$, $\frac{1}{2} < \beta < 1$	Subexponential convergence, $e^{-n^{2\beta-1}(\hat{\gamma}_+(\alpha)+o(1))}$ with rate $\hat{\gamma}_+(\alpha) = \inf_{Q:\mathbb{E}_Q[X]\geq\alpha} \frac{1}{2}\chi^2(Q\|P)$
Large Deviations	$\sum_{i=1}^n X_i \geq n\alpha$, $\alpha > 0$	Exponential convergence, $e^{-n(\gamma_+(\alpha)+o(1))}$ with rate $\gamma_+(\alpha) = \inf_{Q:\mathbb{E}_Q[X]\geq\alpha} D(Q\|P)$

References

1. A. Dembo and O. Zeitouni. Large Deviations Techniques and Applications. Springer, 2nd edition, 1998.
2. I. Csiszár. I-divergence geometry of probability distributions and minimization problems. The Annals of Probability, pages 146–158, 1975.
3. I. Csiszár. Sanov property, generalized I-projection and a conditional limit theorem. The Annals of Probability, 12:768–793, 08 1984.
4. L. Yu. The entropy method in large deviation theory, arXiv:2210.13121, 2022

Thank you!