Lec 4: Large deviation, hypothesis testing

Yanjun Han

Large deviation in finite alphabet: method of types

Suppose P is a pmf on $\mathcal{X}$, with $|\mathcal{X}| < \infty$. For $X_1, \cdots, X_n \overset{\text{i.i.d.}}{\sim} P$, what is the typical "type" of $(X_1, \cdots, X_n)$?

Def (type): For an "empirical distribution" Q on $\mathcal{X}$, let

$$T_Q^n = \left\{ (x_1, \cdots, x_n) \in \mathcal{X}^n : \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(x_i = x) = Q(x), \ \forall x \in \mathcal{X} \right\}.$$

(In other words, T_Q^n is the set of all length-n sequences with empirical distribution equal to Q)

Why types? Types encode all necessary information for $P(x^n)$:

Lemma 1. For $x^n \in T_Q^n$, then $P(x^n) = e^{-n(D_{KL}(Q \| P) + H(Q))}$.

Pf.

$$P(x^n) = \prod_{i=1}^{n} P(x_i) = \prod_{x \in \mathcal{X}} \prod_{i: x_i = x} P(x)$$

$$= \prod_{x \in \mathcal{X}} P(x)^{nQ(x)} \quad (\text{by defn. of } T_Q^n)$$

$$= \exp\left(n \sum_x Q(x) \log P(x)\right)$$

$$= \exp\left(-n(D_{KL}(Q \| P) + H(Q))\right)$$

∎

Another intriguing property is that, # of sequences in a given type is exponential in n, but # of different types is only polynomial in n:

Lemma 2. # of different type classes $= \binom{n + |\mathcal{X}| - 1}{|\mathcal{X}| - 1} \leq (n+1)^{|\mathcal{X}| - 1}$.

Pf. # of non-negative integer solutions to $\sum_{x \in \mathcal{X}} n_x = n$ is $\binom{n + |\mathcal{X}| - 1}{|\mathcal{X}| - 1}$. ∎

Lemma 3. $\frac{e^{nH(Q)}}{(n+1)^{|\mathcal{X}|-1}} \le |T_Q^n| \le e^{nH(Q)}$ (or $|T_Q^n| \doteq e^{nH(Q)}$ by ignoring polynomial factors)

Pf. (Upper bound) $1 \ge Q(X^n \in T_Q^n) \overset{\text{Lemma 1}}{=} |T_Q^n| e^{-nH(Q)}$.

(Lower bound) $1 = \sum_P Q(X^n \in T_P^n)$

$\le \sum_P Q(X^n \in T_Q^n)$ (mode of a multinomial$(n;Q)$ RV is nQ)

$\le (n+1)^{|\mathcal{X}|-1} \cdot |T_Q^n| \cdot e^{-nH(Q)}$. ∎

Corollary. $\frac{e^{-nD_{KL}(Q\|P)}}{(1+n)^{|\mathcal{X}|-1}} \le P(X^n \in T_Q^n) \le e^{-nD_{KL}(Q\|P)}$.

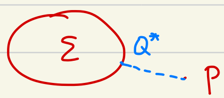
Pf. By Lemma 1 & 3. ∎

The above corollary, together with Lemma 2, leads to the following result known as Sanov's theorem.

Thm. Let $|\mathcal{X}| < \infty$, and $\hat{P}$ be the empirical distribution (type) of $X_1, \cdots, X_n \sim$ a strictly positive P. Let $\mathcal{E}$ be a closed set of distributions with an non-empty interior. Then

$$P(\hat{P} \in \mathcal{E}) = \exp\left(-n \min_{Q \in \mathcal{E}} D_{KL}(Q\|P) + o(n)\right).$$

Remark: The map $P \mapsto \arg\min_{Q \in \mathcal{E}} D_{KL}(Q\|P)$ is called the "information projection".



Pf. (Upper bound) $P(\hat{P} \in \mathcal{E}) = \sum_{Q \in \mathcal{E}} P(X^n \in T_Q^n) \le \sum_{Q \in \mathcal{E}} e^{-nD_{KL}(Q\|P)}$

$\le (n+1)^{|\mathcal{X}|-1} e^{-n \cdot \min_{Q \in \mathcal{E}} D_{KL}(Q\|P)}$,

(Lower bound) For any $Q \in \mathcal{E}$, $P(X^n \in T_Q^n) \ge \frac{1}{(n+1)^{|\mathcal{X}|-1}} e^{-nD_{KL}(Q\|P)}$.

Choose $Q \to Q^*$ and apply continuity of $Q \mapsto D_{KL}(Q\|P)$.

Information projection, exponential tilting, and CGF

A corollary of Sanov's theorem is as follows.

Corollary.
$$\lim_{n\to\infty} \frac{1}{n} \log \frac{1}{P\left(\frac{1}{n}\sum_{i=1}^{n} X_i \geq \gamma\right)} = \min_{Q:\, \mathbb{E}_Q[X] \geq \gamma} D_{KL}(Q \| P).$$

If $\mathbb{E}_P[X] \geq \gamma$, then one can choose $Q = P$ and then RHS $= 0$. Can we find the minimizer Q^* if $\mathbb{E}_P[X] < \gamma$?

Def (exponential tilt) For $\lambda \in \mathbb{R}$, the exponential tilt of P along X is
$$P_\lambda(dx) = \exp(\lambda x - \psi(\lambda)) \cdot P(dx).$$
where $\psi(\lambda) = \log \mathbb{E}_P e^{\lambda X}$ is the cumulant generating function (CGF) of X.

(Note: the family of $\{P_\lambda\}$ is called an "exponential family" in statistics, where $\psi(\lambda)$ is called the "log partition function". In particular, $\mathbb{E}_{P_\lambda}[X] = \psi'(\lambda)$, and $\lambda \mapsto \psi(\lambda)$ is convex.)

Thm ("maximum entropy distribution")
If $\mathbb{E}_P[X] < \gamma$, and there exists $\lambda \in \mathbb{R}$ s.t. $\mathbb{E}_{P_\lambda}[X] = \gamma$. Then
$$\min_{Q:\, \mathbb{E}_Q[X] \geq \gamma} D_{KL}(Q \| P) \overset{①}{=} D_{KL}(P_\lambda \| P) \overset{②}{=} \lambda\gamma - \psi(\lambda) \overset{③}{=} \psi^*(\gamma).$$

where ψ^* is the convex conjugate of ψ.

Pf. Since $\mathbb{E}_P[X] = \psi'(0) < \gamma = \psi'(\lambda)$, by convexity of ψ we have $\lambda > 0$.

① + ②: If $\mathbb{E}_Q[X] \geq \gamma$, then

$$
\begin{aligned}
D_{KL}(Q \| P) &= \mathbb{E}_Q\left[\log \frac{Q}{P}\right] \\
&= \mathbb{E}_Q\left[\log \frac{Q}{P_\lambda} + \log \frac{P_\lambda}{P}\right] \\
&= D_{KL}(Q \| P_\lambda) + \mathbb{E}_Q[\lambda X - \psi(\lambda)] \\
&\overset{\substack{\mathbb{E}_Q[X] \geq \gamma \\ \text{and } \lambda \geq 0}}{\geq} \lambda\gamma - \psi(\lambda),
\end{aligned}
$$

and $D_{KL}(P_\lambda \| P) = \mathbb{E}_{P_\lambda}[\lambda X - \psi(\lambda)] = \lambda\gamma - \psi(\lambda)$.

③: By assumption, $\gamma = \mathbb{E}_{P_\lambda}[X] = \psi'(\lambda)$. Then

$$\psi^*(\gamma) = \sup_{\lambda^* \in \mathbb{R}} \lambda^*\gamma - \psi(\lambda^*) \leq \sup_{\lambda^* \in \mathbb{R}} \lambda^*\gamma - (\psi(\lambda) + (\lambda^* - \lambda)\psi'(\lambda)) = \lambda\gamma - \psi(\lambda)$$

by convexity of ψ. So $\psi^*(\gamma) = \lambda\gamma - \psi(\lambda)$. ∎

In other words, this result shows that the information projection yields an exponential tilt of P, and the value is given by the convex conjugate of the CGF of P.

Large deviation in general alphabets: Cramér's Thm.

Cramér's Thm. For i.i.d. $X_1, \cdots, X_n \sim P$ with $\mathbb{E}_P[X] < \gamma < \|X\|_\infty$, then

$$\lim_{n \to \infty} \frac{1}{n} \log \frac{1}{P(\frac{1}{n}\sum_{i=1}^n X_i > \gamma)} = \psi^*(\gamma) = \inf_{Q: \mathbb{E}_Q[X] > \gamma} D_{KL}(Q \| P)$$

where ψ^* is the convex conjugate of the CGF $\psi(\lambda) = \log \mathbb{E}_P e^{\lambda X}$.

Note: This generalizes our previous results to arbitrary alphabets. Also, we'll present two different proofs, one probabilistic and one information-theoretic, to arrive at the quantities $\psi^*(\gamma)$ and $\min_{Q: \mathbb{E}_Q[X] > \gamma} D_{KL}(Q \| P)$, respectively. These proofs will better illustrate the connections between different ideas.

Probabilistic proof.

(Lower bound) By Chernoff inequality,

$$P\left(\frac{1}{n}\sum_i X_i > \gamma\right) \le \inf_{\lambda \ge 0} e^{-\lambda n \gamma} \mathbb{E}_P\left[e^{\lambda \sum_i X_i}\right]$$

$$= \inf_{\lambda\ge 0} \exp(-n(\lambda\gamma - \psi(\lambda))) \underset{\substack{\uparrow \\ \text{this step uses } \mathbb{E}_P[X] < \gamma}}{=} \exp(-n\psi^*(\gamma)).$$

(Upper bound) Since $\mathbb{E}_P[X] < \gamma < \|X\|_\infty$, $\exists\ \lambda = \lambda(\varepsilon) > 0$ s.t. $\mathbb{E}_{P_\lambda}[X] = \gamma + \varepsilon$, where P_λ is the exponential tilt of P. By LLN,

$$P_\lambda\left(\frac{1}{n}\sum_i X_i \in (\gamma, \gamma+2\varepsilon)\right) = 1 - o(1) \quad \text{as} \quad n \to \infty.$$

At the same time, for $\frac{1}{n}\sum_i X_i \in (\gamma, \gamma + 2\varepsilon)$,

$$\frac{dP_\lambda}{dP}(x_1, \cdots, x_n) = \exp\left(\lambda \sum_i X_i - n\psi(\lambda)\right) \le \exp\left(n\left(\lambda(\gamma+2\varepsilon) - \psi(\lambda)\right)\right)$$

$$\Rightarrow P\left(\frac{1}{n}\sum_i X_i \in (\gamma, \gamma+2\varepsilon)\right) \ge (1 - o(1)) \exp\left(-n\left(\lambda(\gamma+2\varepsilon) - \psi(\lambda)\right)\right).$$

Choosing $\varepsilon \to 0^+$ completes the proof.

IT proof.

(Upper bound) Fix any Q with $\mathbb{E}_Q[X] > \gamma$. Then for $E_n = \{\frac{1}{n}\sum_i X_i > \gamma\}$,

$$Q(E_n) = 1 - o(1) \text{ by LLN.}$$

By Lec 2,
$$Q(E_n) \log \frac{Q(E_n)}{eP(E_n)} \le D_{KL}(Q_{X^n} \| P_{X^n}) = n D_{KL}(Q \| P)$$

$$\Rightarrow \frac{1}{n} \log \frac{1}{P(E_n)} \le \frac{D_{KL}(Q\|P)}{Q(E_n)} - \frac{\log(eQ(E_n))}{n} = (1 + o(1)) D_{KL}(Q\|P).$$

(Lower bound) Note $\widetilde{P}_{X^n} \triangleq P_{X^n | \frac{1}{n}\sum_i X_i > \gamma}$ has mean $> \gamma$, with

$$\frac{1}{n} \log \frac{1}{P(E_n)} = \frac{1}{n} D_{KL}(\widetilde{P}_{X^n} \| P_{X^n}).$$

We argue that $\frac{1}{n} D_{KL}(\widetilde{P}_{X^n} \| P_{X^n}) \ge \inf_{Q: \mathbb{E}_Q[X] > \gamma} D_{KL}(Q\|P)$. In fact,

$$D_{KL}(\widetilde{P}_{X^n} \| P_{X^n}) = \sum_{i=1}^n \mathbb{E}_{\widetilde{P}}\left[D_{KL}(\widetilde{P}_{X_i | X^{i-1}} \| P)\right]$$

$$\overset{\text{convexity}}{\ge} \sum_{i=1}^n D_{KL}\left(\mathbb{E}_{\widetilde{P}} \widetilde{P}_{X_i|X^{i-1}} \| P\right) \overset{\text{convexity}}{\ge} n D_{KL}\left(\frac{1}{n}\sum_i \widetilde{P}_{X_i} \| P\right),$$

where $\overline{P} := \frac{1}{n}\sum_i \widetilde{P}_{X_i}$ clearly satisfies $\mathbb{E}_{\overline{P}}[X] = \mathbb{E}_{\widetilde{P}}[\frac{1}{n}\sum_i X_i] > \gamma$. $\square$

Simple hypothesis testing.

$H_0: X \sim P$

$H_1: X \sim Q$

For a test $T = T(X) \in \{0,1\}$ (possibly randomized), define

$$\begin{cases} \alpha = P(T=0) & (1 - \text{Type I error}) \\ \beta = Q(T=0) & (\text{Type II error}) \end{cases}$$

Def. Let $R(P,Q)$ denote the set of all achievable points $(\alpha,\beta) \in [0,1]^2$ when T ranges over all possible sets.

Basic properties.

① $R(P,Q)$ is convex (Pf: consider a randomized combination of two tests)

② $(\alpha,\alpha) \in R(P,Q)$ (Pf: consider $T \sim \text{Bern}(1-\alpha)$ independent of X)

③ $(\alpha,\beta) \in R(P,Q) \Leftrightarrow (1-\alpha, 1-\beta) \in R(P,Q)$ (Pf: replacing T by $1-T$)

④ Neyman-Pearson: likelihood ratio tests (LRT) attain the lower boundary of $R(P,Q)$, i.e., for

$$T^* = \begin{cases} 0 & \text{if } \log\frac{p(x)}{q(x)} > \tau, \\ \in\{0,1\} & \text{if } \log\frac{p(x)}{q(x)} = \tau, \quad (\text{randomized}) \\ 1 & \text{if } \log\frac{p(x)}{q(x)} < \tau, \end{cases}$$

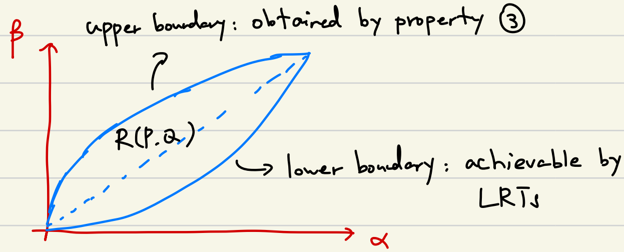
then for any other test T, $\alpha(T) \geq \alpha(T^*) \Rightarrow \beta(T) \geq \beta(T^*)$.

Pf. $\alpha(T) \geq \alpha(T^*) \Rightarrow \mathbb{E}_P[T - T^*] \leq 0$.

Since $\mathbb{E}_P[(\frac{dQ}{dP} - e^{-\tau})(T - T^*)] \leq 0$ (by distinguishing $\frac{dQ}{dP} \gtrless e^{-\tau}$)

we obtain $\mathbb{E}_P[\frac{dQ}{dP}(T - T^*)] \leq 0$, i.e., $\mathbb{E}_Q[T - T^*] \leq 0 \Rightarrow \beta(T) \geq \beta(T^*)$.

Example of $R(P,Q)$:



Asymptotics: Chernoff regime.

Consider $\begin{cases} H_0: & X^n \sim P^{\otimes n} \\ H_1: & X^n \sim Q^{\otimes n} \end{cases}$ with $n \to \infty$. What are all possible values of (E_0, E_1) s.t. $\exists\, T_n$ with $\begin{cases} 1-\alpha(T_n) \leq e^{-nE_0} \\ \beta(T_n) \leq e^{-nE_1} \end{cases}$ asymptotically?

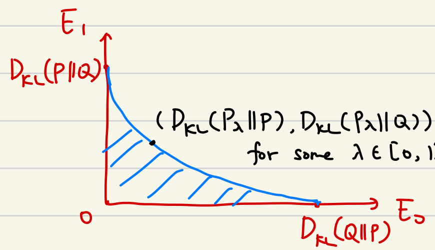
In other words, what are the best tradeoffs between (E_0, E_1), the error exponents on Type I & II errors?

Thm (E_0-E_1 tradeoff). Assume $P \ll Q$ and $Q \ll P$. The upper boundary of all achievable (E_0, E_1) pairs is given by

$$\begin{cases} E_0 = D_{KL}(P_\lambda \| P) \\ E_1 = D_{KL}(P_\lambda \| Q) \end{cases} \qquad \lambda \in [0,1]$$

where $P_\lambda \propto P^{1-\lambda} Q^{\lambda}$.

Illustration of (E_0, E_1):



Corollary.
$$\max_{\substack{(E_0,E_1)\\ \text{achievable}}} \min\{E_0, E_1\} = -\inf_{\lambda \in [0,1]} \log \int (dP)^{1-\lambda}(dQ)^{\lambda}.$$

Note: This quantity, denoted by $C(P,Q)$, is called the Chernoff information.

It can be shown that

choose $\lambda = \frac{1}{2}$

$$-\log\left(1-\tfrac{1}{2}H^2(P,Q)\right) \leq C(P,Q) \leq -2\log\left(1-\tfrac{1}{2}H^2(P,Q)\right).$$

$$\int P^{1-\lambda}Q^{\lambda} = \mathbb{E}_P\left[\left(\tfrac{Q}{P}\right)^{\lambda}\right] \geq \left(\mathbb{E}_P\sqrt{\tfrac{Q}{P}}\right)^{2\lambda}$$
$$\geq \left(\int \sqrt{PQ}\right)^2 \text{ if } \lambda \geq \tfrac{1}{2}$$

and symmetrically for $\lambda < \frac{1}{2}$.

<u>Pf of corollary</u>. For $P_\lambda = \frac{P^{1-\lambda} Q^\lambda}{Z}$.

$$D_{KL}(P_\lambda \| P) = \mathbb{E}_{P_\lambda}[\log \frac{P_\lambda}{P}] = \mathbb{E}_{P_\lambda}[\lambda \log \frac{Q}{P} - \log Z]$$

$$D_{KL}(P_\lambda \| Q) = \mathbb{E}_{P_\lambda}[\log \frac{P_\lambda}{Q}] = \mathbb{E}_{P_\lambda}[(1-\lambda) \log \frac{P}{Q} - \log Z]$$

$$\Rightarrow D_{KL}(P_\lambda \| P) - D_{KL}(P_\lambda \| Q) = \mathbb{E}_{P_\lambda}[\log \frac{Q}{P}].$$

Let λ^* denote the minimizer of the convex function $\lambda \mapsto \log \int P^{1-\lambda} Q^\lambda$ on $[0,1]$,

then $0 = \frac{d}{d\lambda} \log \int P^{1-\lambda} Q^\lambda \Big|_{\lambda=\lambda^*} = \frac{1}{Z} \int P^{1-\lambda^*} Q^{\lambda^*} \log \frac{Q}{P} = \mathbb{E}_{P_{\lambda^*}}[\log \frac{Q}{P}]$.

For this λ^*, we have $D_{KL}(P_{\lambda^*} \| P) = D_{KL}(P_{\lambda^*} \| Q)$, and

$$D_{KL}(P_{\lambda^*} \| P) = -\log Z = -\log \int P^{1-\lambda^*} Q^{\lambda^*} = -\inf_{\lambda \in [0,1]} \log \int P^{1-\lambda} Q^\lambda. \quad \boxed{\text{M}}$$

Back to the (E_0, E_1) tradeoff:

<u>Achievability</u>: a sufficient statistic is $L \triangleq \frac{1}{n} \sum_{i=1}^n L_i \triangleq \frac{1}{n} \sum_{i=1}^n \log \frac{P(X_i)}{Q(X_i)}$,

so a natural test is $T_n = \mathbb{1}(L \le \gamma)$ for some threshold $\gamma \in \mathbb{R}$.

By large deviation:

$$\lim_{n\to\infty} \frac{1}{n} \log \frac{1}{P(L \le \gamma)} = \psi_P^*(\gamma) = D_{KL}(P^* \| P)$$

$$\lim_{n\to\infty} \frac{1}{n} \log \frac{1}{Q(L > \gamma)} = \psi_Q^*(\gamma) = D_{KL}(Q^* \| Q)$$

where $\psi_P(\lambda) = \log \mathbb{E}_P e^{\lambda L_1} = \log \int P^{1+\lambda} Q^{-\lambda}$ (similarly for ψ_Q), and

$$P^*(dx) = \exp\left(\lambda_P^* \log \frac{p(x)}{q(x)} - \psi_P(\lambda_P^*)\right) P(dx) \quad \text{with } \mathbb{E}_{P^*}[L_1] = \gamma.$$

$$Q^*(dx) = \exp\left(\lambda_Q^* \log \frac{p(x)}{q(x)} - \psi_Q(\lambda_Q^*)\right) Q(dx) \quad \text{with } \mathbb{E}_{Q^*}[L_1] = \gamma.$$

Since $P^* \propto p^{1+\lambda_P^*} q^{-\lambda_P^*}$ and $Q^* \propto p^{\lambda_Q^*} q^{1-\lambda_Q^*}$ belong to the family $(P_\lambda)_{\lambda \in [0,1]}$,

we conclude that $P^* = Q^* = P_{\lambda^*}$, where λ^* is the solution to $\mathbb{E}_{P_{\lambda^*}}[\log \frac{p(X)}{q(X)}] = \gamma$.

Therefore, by choosing γ appropriately, this test asymptotically achieves all pairs

$$(E_0, E_1) = (D_{KL}(P_\lambda \| P), D_{KL}(P_\lambda \| Q)) \quad \text{for all } \lambda \in [0,1].$$

<u>Converse</u>. Suppose some test T_n asymptotically attains $\alpha(T_n) \geq 1 - e^{-nE_0}$
$\beta(T_n) \leq e^{-nE_1}$

<u>Weak converse</u> (by DPI): $D_{KL}(\mathrm{Bern}(\alpha) \| \mathrm{Bern}(\beta)) \leq n D_{KL}(P \| Q)$
$D_{KL}(\mathrm{Bern}(\beta) \| \mathrm{Bern}(\alpha)) \leq n D_{KL}(Q \| P)$

(They are insufficient to establish the tight (E_0, E_1) tradeoff!)

<u>Strong converse</u> (on the whole likelihood ratio): $\forall \gamma > 0$,

$$\alpha - \gamma\beta \leq P\left(\sum_{i=1}^{n} \log \frac{P}{Q}(X_i) > \log \gamma \right)$$
$$\beta - \frac{\alpha}{\gamma} \leq Q\left(\sum_{i=1}^{n} \log \frac{P}{Q}(X_i) < \log \gamma \right)$$

<u>Pf</u>. Let $L = \sum_{i=1}^{n} \log \frac{P}{Q}(X_i) = \log \frac{P^{\otimes n}}{Q^{\otimes n}}(X)$. Then

$$\begin{aligned} \alpha - \gamma\beta &= P^{\otimes n}(T_n = 0) - \gamma Q^{\otimes n}(T_n = 0) \\ &= \mathbb{E}_{Q^{\otimes n}}\left[(e^L - \gamma) 1(T_n = 0) \right] \\ &\leq \mathbb{E}_{Q^{\otimes n}}\left[(e^L - \gamma) 1(T_n = 0, L > \log \gamma) \right] \\ &\leq \mathbb{E}_{Q^{\otimes n}}\left[e^L 1(L > \log \gamma) \right] = P^{\otimes n}(L > \log \gamma). \end{aligned}$$

The second is similar. $\square$

(Compared with weak converse, the strong converse proposes to keep track of the whole behavior of L, and mimics the large deviation analysis in the achievability)

<u>Returning to the converse</u>: choose $\gamma = e^{n\theta}$, then

$$1 - e^{-nE_0} - e^{-n(E_1 - \theta)} \leq \alpha - \gamma\beta \leq P\left(\frac{1}{n} \sum_{i=1}^{n} \log \frac{P}{Q}(X_i) > \theta \right)$$
$$\Rightarrow \quad e^{-nE_0} + e^{-n(E_1 - \theta)} \geq P\left(\frac{1}{n} \sum_{i=1}^{n} \log \frac{P}{Q}(X_i) \leq \theta \right)$$
$$\Rightarrow \quad \min\{E_0, E_1 - \theta\} \leq \psi_P^*(\theta), \quad \forall \theta.$$

If $E_0 \geq D_{KL}(P_\lambda \| P) + \varepsilon$, $E_1 \geq D_{KL}(P_\lambda \| Q) + \varepsilon$, choose

$$\theta = D_{KL}(P_\lambda \| Q) - D_{KL}(P_\lambda \| P) = \mathbb{E}_{P_\lambda}\left[\log \frac{P}{Q} \right] \quad \text{(see previous page)}$$

then $\psi_P^*(\theta) = D_{KL}(P_\lambda \| P)$ (because λ is the solution to $\mathbb{E}_{P_\lambda}[\log \frac{P}{Q}] = \theta$)

$$\Rightarrow \min\{E_0, E_1 - \theta\} \geq \psi_P^*(\theta) + \varepsilon, \quad \text{a contradiction!}$$

Special topic: Stein's regime, strong converse for channel coding, finite blocklength

Stein's regime: $\begin{cases} H_0: & X \sim P^{\otimes n} \\ H_1: & X \sim Q^{\otimes n}. \end{cases}$

$\exists$ test T_n s.t. $\alpha(T_n) = 1-\varepsilon$ and $\beta(T_n) = e^{-nE}$

What's the largest possible value E_n^* of E?

From the Chernoff regime with $E_0 = 0$, we already know that

$$E_n^* = D_{KL}(P\|Q) + o(1). \quad \text{(Stein's lemma)}$$

Can we also get the next-order term?

> **Thm.** $$E_n^* = D_{KL}(P\|Q) - \sqrt{\frac{V(P\|Q)}{n}}\, \mathrm{erfc}^{-1}(\varepsilon) + o\left(\frac{1}{\sqrt{n}}\right),$$
>
> where $\mathrm{erfc}(z) = \mathbb{P}(N(0,1) > z) = \int_z^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx,$
>
> $V(P\|Q) = \mathrm{Var}_P\left(\log \frac{P}{Q}\right)$ (assumed to be $<\infty$)

Pf (achievability) Consider the test $T_n = 1\left(\frac{1}{n}\sum_{i=1}^n \log\frac{P}{Q}(X_i) \le \gamma\right)$.

By CLT, $$\frac{1}{\sqrt{n}} \sum_{i=1}^n \left(\log\frac{P}{Q}(X_i) - D_{KL}(P\|Q)\right) \xrightarrow[\text{under } P]{d} N(0, V(P\|Q)),$$

so $\gamma = nD_{KL}(P\|Q) - \sqrt{nV(P\|Q)}\,\mathrm{erfc}^{-1}(\varepsilon)$ yields $\alpha(T_n) \longrightarrow 1-\varepsilon$ as $n\to\infty$.

For $\beta(T_n)$: $$Q\left(\frac{1}{n}\sum_{i=1}^n \log\frac{P}{Q}(X_i) > \gamma\right) \le e^{-n\gamma}\, \mathbb{E}_Q\left[e^{\sum_i \log\frac{P}{Q}(X_i)}\right] = e^{-n\gamma}.$$

(converse) If $E_n \ge D_{KL}(P\|Q) + \frac{C}{\sqrt{n}}$, then strong converse yields

$$1-\varepsilon-o(1) = \alpha - e^{n\left(D_{KL}(P\|Q)+\frac{C-\delta}{\sqrt{n}}\right)}\beta \le P\left(\frac{1}{n}\sum_i \log\frac{P}{Q}(X_i) > D_{KL}(P\|Q) + \frac{C-\delta}{\sqrt{n}}\right)$$

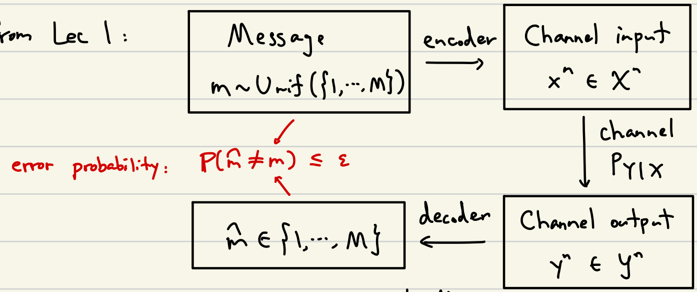
$$\xrightarrow[\text{CLT}]{n\to\infty} \mathrm{erfc}\left(\frac{C-\delta}{\sqrt{V(P\|Q)}}\right)$$

$$\Longrightarrow C \le -\sqrt{V(P\|Q)}\,\mathrm{erfc}^{-1}(\varepsilon) + \delta.$$ ∎

Note: If one uses Berry-Esseen bounds, then under moment conditions, the $o\left(\frac{1}{\sqrt{n}}\right)$ factor can be improved to $O\left(\frac{\log n}{n}\right)$.

Strong converse for channel coding.

Recall from Lec 1:



Communications aim to minimize $R = \frac{\log M}{n}$. In Lec 1, we use Fano's inequality (i.e. DPI for KL) to prove the weak converse $R \le (1+o(1))C$ if $\varepsilon = o(1)$, with

$$C = \max_{P_X} I(X;Y) = \max_{P_X} I(P_X; P_{Y|X}).$$

What happens if $\varepsilon = 0.01$, or even $\varepsilon = 0.999$?

> Thm (strong converse). For any fixed $\varepsilon < 1$,
> $$R \le (1+o(1))C.$$

Remark: This means that the communication problem has a "sharp" threshold on the error probability. When $R < 0.999C$, then asymptotically one cannot achieve a success probability of 10^{-8}; when $R > 1.001C$, then asymptotically one can suddenly achieve a success probability of $1-10^{-8}$.

Pf. The communication problem is not binary hypothesis testing; instead, it is a recovery problem (i.e. recover the message m from Y^n). However, a useful idea is to reduce a recovery problem to detection: if one can distinguish between different inputs (recovery), then one can also distinguish from the case where the input and output are independent. This idea is also frequently used in statistical problems.

Consider two scenarios (i.e. joint distributions on $m, X^n, Y^n, \hat{m}$):

$$H_0: \quad P_{m,X^n,Y^n,\hat{m}} = \frac{1}{M} P_{X^n|m} P_{Y^n|X^n} P_{\hat{m}|Y^n}$$

$$H_1: \quad Q_{m,X^n,Y^n,\hat{m}} = \frac{1}{M} P_{X^n|m} Q_Y^{\otimes n} P_{\hat{m}|Y^n} \quad (\text{i.e. } (m, X^n) \perp\!\!\!\perp (Y^n, \hat{m}))$$

Then $\begin{cases} P(m=\hat{m}) \ge 1-\varepsilon \\ Q(m=\hat{m}) = \frac{1}{M} \end{cases}$, and the likelihood ratio is

$$\frac{P_{m,X^n,Y^n,\hat{m}}}{Q_{m,X^n,Y^n,\hat{m}}} = \frac{P_{Y^n|X^n}}{Q_Y^{\otimes n}} = \prod_{i=1}^{n} \frac{P_{Y_i|X_i}}{Q_{Y_i}}$$

Therefore, by strong converse,

$$1-\varepsilon - \frac{\gamma}{M} \le P\left(\sum_i \log \frac{P_{Y_i|X_i}}{Q_{Y_i}} > \log \gamma\right)$$

<u>A technical difficulty</u>: P_{X^n} is often <u>not</u> a product distribution

<u>Solution</u>: When $|\mathcal{X}| < \infty$, can WLOG assume that all codewords X^n have the same type P_0. In fact, since there are $\le (n+1)^{|\mathcal{X}|-1}$ types, one can find a type that changes the error probability to $\varepsilon + o(1)$ while with a rate change at most $O(\frac{\log n}{n})$.

When X^n has type P_0 a.s., choose $Q_Y = \sum_x P_0(x) P_{Y|X=x}$. Then

$$\mathbb{E}\left[\sum_i \log \frac{P_{Y_i|X_i}}{Q_{Y_i}}\right] = n I(P_0; P_{Y|X}) \le nC$$

$$\mathrm{Var}\left(\sum_i \log \frac{P_{Y_i|X_i}}{Q_{Y_i}}\right) = n \mathbb{E}_{P_0}\left[\mathrm{Var}\left(\log \frac{P_{Y|X}}{Q_Y} \,\Big|\, X\right)\right] \le n \mathrm{Var}\left(\log \frac{P_{Y|X}}{Q_Y}\right) \underset{\uparrow}{=} O(n)$$

Exercise: $\sum_x p(x) \log^2 p(x) \le 2 \log^2 |\mathcal{X}|$

Now choosing $\gamma = \frac{1-\varepsilon}{2} M$ in the strong converse, Chebyshev's inequality yields

$$\log \gamma \le nC + O(\sqrt{n}) \implies R = \frac{\log M}{n} \le C + O\left(\frac{1}{\sqrt{n}}\right). \quad \square$$

Converse for finite blocklength

Is there a next-order upper bound on R?

Thm. Suppose that the capacity-achieving distribution P_X^* is unique, and $|\mathcal{X}|, |\mathcal{Y}| < \infty$. Under regularity conditions,

$$R \le C - \sqrt{\frac{V}{n}}\, \mathrm{erfc}^{-1}(\varepsilon) + o\left(\frac{1}{\sqrt{n}}\right),$$

with $V = \mathbb{E}_{P_X^*}\left[\mathrm{Var}\left(\log \frac{P_{Y|X}}{P_Y^*}\right)\right]$.

Pf sketch. Using the previous analysis, and due to the uniqueness of P_X^*, we only need to deal with the input type $P_0 \approx P_X^*$. Then the result follows from Stein's regime as long as we can show

$$\mathbb{E}_{P_X^*}\left[\mathrm{Var}\left(\log \frac{P_{Y|X}}{P_Y^*} \,\Big|\, X\right)\right] = \mathbb{E}_{P_X^*}\left[\mathrm{Var}\left(\log \frac{P_{Y|X}}{P_Y}\right)\right] = V.$$

This follows from the following lemma. $\boxed{\text{B}}$

Lemma. Any capacity-achieving input P_X^* satisfies

$$D_{KL}(P_{Y|X=x} \| P_Y^*) \le C, \quad \forall x \in \mathcal{X}$$

$$D_{KL}(P_{Y|X=x} \| P_Y^*) = C, \quad \forall x \in \mathrm{supp}(P_X^*).$$

Pf. $$0 \ge \lim_{\varepsilon \to 0^+} \frac{I(P_X^* + \varepsilon(P_X - P_X^*); P_{Y|X}) - I(P_X^*; P_{Y|X})}{\varepsilon} = (\mathbb{E}_{P_X} - \mathbb{E}_{P_X^*})[D_{KL}(P_{Y|X} \| P_Y^*)].$$

Choosing $P_X = \delta_x$ gives the first claim. The second claim follows from

$$C = \mathbb{E}_{P_X^*}[D_{KL}(P_{Y|X} \| P_Y^*)] \le C,$$

so the equality must hold for $x \in \mathrm{supp}(P_X^*)$. $\blacksquare$