12.1 Large-deviation exponents

Large deviations problems make statements about the tail probabilities of a sequence of distributions. We're interested in the speed of decay for probabilities such as $P\left[\frac{1}{n}\sum_{k=1}^{n} X_k \geq \gamma\right]$ for iid X_k.

In the last lecture we used Chernoff bound to obtain an upper bound on the exponent via the log-MGF and tilting. Next we use a different method to give a formula for the exponent as a convex optimization problem involving the KL divergence (information projection). Later in Section 12.3 we shall revisit the Chernoff bound after we have computed the value of the information projection.

Theorem 12.1. *Let $X^n \overset{i.i.d.}{\sim} P$. Then for any $\gamma \in \mathbb{R}$,*

$$\lim_{n\to\infty} \frac{1}{n} \log \frac{1}{P\left[\frac{1}{n}\sum_{k=1}^{n} X_k > \gamma\right]} = \inf_{Q:\mathbb{E}_Q[X]>\gamma} D(Q\|P) \tag{12.1}$$

$$\lim_{n\to\infty} \frac{1}{n} \log \frac{1}{P\left[\frac{1}{n}\sum_{k=1}^{n} X_k \geq \gamma\right]} = \inf_{Q:\mathbb{E}_Q[X]\geq\gamma} D(Q\|P) \tag{12.2}$$

Proof. We first prove (12.1). Set $P[E_n] = P\left[\frac{1}{n}\sum_{k=1}^{n} X_k > \gamma\right]$.

Lower Bound on $P[E_n]$: Fix a Q such that $\mathbb{E}_Q[X] > \gamma$. Let X^n be iid. Then by WLLN,

$$Q[E_n] = Q\left[\sum_{k=1}^{n} X_k > n\gamma\right] = 1 - o(1).$$

Now the data processing inequality gives

$$d(Q[E_n]\|P[E_n]) \leq D(Q_{X^n}\|P_{X^n}) = nD(Q\|P)$$

And a lower bound for the binary divergence is

$$d(Q[E_n]\|P[E_n]) \geq -h(Q[E_n]) + Q[E_n]\log\frac{1}{P[E_n]}$$

Combining the two bounds on $d(Q[E_n]\|P[E_n])$ gives

$$P[E_n] \geq \exp\left(\frac{-nD(Q\|P) - \log 2}{Q[E_n]}\right) \tag{12.3}$$

Optimizing over Q to give the best bound:

$$\limsup_{n\to\infty} \frac{1}{n} \log \frac{1}{P[E_n]} \leq \inf_{Q:\mathbb{E}_Q[X]>\gamma} D(Q\|P).$$

Upper Bound on $P[E_n]$: The key observation is that given any X and any event E, $P_X(E) > 0$ can be expressed via the divergence between the conditional and unconditional distribution as: $\log \frac{1}{P_X(E)} = D(P_{X|X\in E}\|P_X)$. Define $\tilde{P}_{X^n} = P_{X^n|\sum X_i > n\gamma}$, under which $\sum X_i > n\gamma$ holds a.s. Then

$$\log \frac{1}{P[E_n]} = D(\tilde{P}_{X^n}\|P_{X^n}) \geq \inf_{Q_{X^n}:\mathbb{E}_Q[\sum X_i]>n\gamma} D(Q_{X^n}\|P_{X^n}) \tag{12.4}$$

We know show that the last problem "single-letterizes", i.e. need to be solved only for $n = 1$. Consider the following two steps:

$$D(Q_{X^n}\|P_{X^n}) \geq \sum_{j=1}^{n} D(Q_{X_j}\|P) \tag{12.5}$$

$$\geq nD(\bar{Q}\|P)\,, \qquad \bar{Q} \triangleq \frac{1}{n}\sum_{j=1}^{n} Q_{X_j}\,, \tag{12.6}$$

where the first step follows from Corollary 2.1 after noticing that $P_{X^n} = P^n$, and the second step is by convexity of divergence Theorem 4.1. From this argument we conclude that

$$\inf_{Q_{X^n}:\mathbb{E}_Q[\sum X_i]>n\gamma} D(Q_{X^n}\|P_{X^n}) = n \cdot \inf_{Q:\mathbb{E}_Q[X]>\gamma} D(Q\|P) \tag{12.7}$$

$$\inf_{Q_{X^n}:\mathbb{E}_Q[\sum X_i]\geq n\gamma} D(Q_{X^n}\|P_{X^n}) = n \cdot \inf_{Q:\mathbb{E}_Q[X]\geq\gamma} D(Q\|P) \tag{12.8}$$

In particular, (12.4) and (12.7) imply the required lower bound in (12.1).

Next we prove (12.2). First, notice that the lower bound argument (12.4) applies equally well, so that for each n we have

$$\frac{1}{n}\log \frac{1}{P\left[\frac{1}{n}\sum_{k=1}^{n} X_k \geq \gamma\right]} \geq \inf_{Q:\mathbb{E}_Q[X]\geq\gamma} D(Q\|P)\,.$$

To get a matching upper bound we consider two cases:

- Case I: $P[X > \gamma] = 0$. If $P[X \geq \gamma] = 0$, then both sides of (12.2) are $+\infty$. If $P[X = \gamma] > 0$, then $P[\sum X_k \geq n\gamma] = P[X_1 = \ldots = X_n = \gamma] = P[X = \gamma]^n$. For the right-hand side, since $D(Q\|P) < \infty \implies Q \ll P \implies Q(X \leq \gamma) = 1$, the only possibility for $\mathbb{E}_Q[X] \geq \gamma$ is that $Q(X = \gamma) = 1$, i.e., $Q = \delta_\gamma$. Then $\inf_{\mathbb{E}_Q[X]\geq\gamma} D(Q\|P) = \log \frac{1}{P(X=\gamma)}$.

- Case II: $P[X > \gamma] > 0$. Since $\mathbb{P}[\sum X_k \geq \gamma] \geq \mathbb{P}[\sum X_k > \gamma]$ from (12.1) we know that

$$\limsup_{n\to\infty} \frac{1}{n}\log \frac{1}{P\left[\frac{1}{n}\sum_{k=1}^{n} X_k \geq \gamma\right]} \leq \inf_{Q:\mathbb{E}_Q[X]>\gamma} D(Q\|P)\,.$$

  We next show that in this case

$$\inf_{Q:\mathbb{E}_Q[X]>\gamma} D(Q\|P) = \inf_{Q:\mathbb{E}_Q[X]\geq\gamma} D(Q\|P) \tag{12.9}$$

  Indeed, let $\tilde{P} = P_{X|X>\gamma}$ which is well defined since $P[X > \gamma] > 0$. For any Q such that $\mathbb{E}_Q[X] \geq \gamma$, set $\tilde{Q} = \bar{\epsilon}Q + \epsilon\tilde{P}$ satisfies $\mathbb{E}_{\tilde{Q}}[X] > \gamma$. Then by convexity, $D(Q\|P) \leq \bar{\epsilon}D(Q\|P) + \epsilon D(\tilde{P}\|P) = \bar{\epsilon}D(Q\|P) + \epsilon\log\frac{1}{P[X>\gamma]}$. Sending $\epsilon \to 0$, we conclude the proof of (12.9).

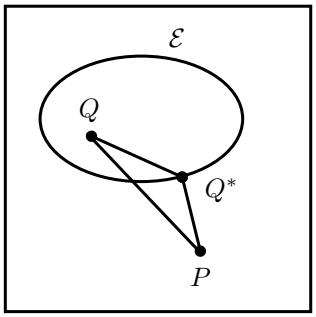
□

12.2 Information Projection

The results of Theorem 12.1 motivate us to study the following general **information projection problem:** Let $\mathcal{E}$ be a convex set of distributions on some abstract space Ω, then for the distribution P on Ω, we want

$$\inf_{Q \in \mathcal{E}} D(Q\|P)$$

Denote the minimizing distribution Q by Q^*. The next result shows that intuitively the "line" between P and optimal Q^* is "orthogonal" to $\mathcal{E}$.



Distributions on $\mathcal{X}$

Theorem 12.2. *Suppose* $\exists Q^* \in \mathcal{E}$ *such that* $D(Q^*\|P) = \min_{Q \in \mathcal{E}} D(Q\|P)$*, then* $\forall Q \in \mathcal{E}$

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

Proof. If $D(Q\|P) = \infty$, then we're done, so we can assume that $D(Q\|P) < \infty$, which also implies that $D(Q^*\|P) < \infty$. For $\theta \in [0,1]$, form the convex combination $Q^{(\theta)} = \bar{\theta}Q^* + \theta Q \in \mathcal{E}$. Since Q^* is the minimizer of $D(Q\|P)$, then[1]

$$0 \leq \frac{\partial}{\partial \theta}\Big|_{\theta=0} D(Q^{(\theta)}\|P) = D(Q\|P) - D(Q\|Q^*) - D(Q^*\|P)$$

and we're done. □

Remark: If we view the picture above in the Euclidean setting, the "triangle" formed by P, Q^* and Q (for Q^*, Q in a convex set, P outside the set) is always obtuse, and is a right triangle only when the convex set has a "flat face". In this sense, the divergence is similar to the squared Euclidean distance, and the above theorem is sometimes known as a "Pythagorean" theorem.

The interesting set of Q's that we will particularize to is the "half-space" of distributions $\mathcal{E} = \{Q : \mathbb{E}_Q[X] \geq \gamma\}$, where $X : \Omega \to \mathbb{R}$ is some fixed function. This is justified by relation (to be established) with the large deviation exponent in Theorem 12.1. First, we solve this I-projection problem explicitly.

Theorem 12.3. *Given distribution* P *on* Ω *and* $X : \Omega \to \mathbb{R}$ *let*

$$A = \inf \psi'_X = \operatorname{essinf} X = \sup\{a : X \geq a \ P\text{-a.s.}\} \tag{12.10}$$

$$B = \sup \psi'_X = \operatorname{esssup} X = \inf\{b : X \leq b \ P\text{-a.s.}\} \tag{12.11}$$

[1] This can be found by taking the derivative and matching terms (Exercise). Be careful with exchanging derivatives and integrals. Need to use dominated convergence theorem similar as in the "local behavior of divergence" in Proposition 4.1.

1. *The information projection problem over* $\mathcal{E} = \{Q : \mathbb{E}_Q[X] \geq \gamma\}$ *has solution*

$$\min_{Q:\mathbb{E}_Q[X]\geq\gamma} D(Q\|P) = \begin{cases} 0 & \gamma < \mathbb{E}_P[X] \\ \psi^*(\gamma) & \mathbb{E}_P[X] \leq \gamma < B \\ \log \frac{1}{P(X=B)} & \gamma = B \\ +\infty & \gamma > B \end{cases} \tag{12.12}$$

$$= \psi^*(\gamma) 1\{\gamma \geq \mathbb{E}_P[X]\} \tag{12.13}$$

2. *Whenever the minimum is finite, minimizing distribution is unique and equal to tilting of* P *along* X*, namely*[2]

$$dP_\lambda = \exp\{\lambda X - \psi(\lambda)\} \cdot dP \tag{12.14}$$

3. *For all* $\gamma \in [\mathbb{E}_P[X], B)$ *we have*

$$\min_{\mathbb{E}_Q[X]\geq\gamma} D(Q\|P) = \inf_{\mathbb{E}_Q[X]>\gamma} D(Q\|P) = \min_{\mathbb{E}_Q[X]=\gamma} D(Q\|P)\,.$$

Note: An alternative expression is

$$\min_{Q:\mathbb{E}_Q[X]\geq\gamma} = \sup_{\lambda\geq 0} \lambda\gamma - \psi_X(\lambda)\,.$$

Proof. First case: Take $Q = P$.

Fourth case: If $\mathbb{E}_Q[X] > B$, then $Q[X \geq B + \epsilon] > 0$ for some $\epsilon > 0$, but $P[X \geq B + \epsilon] = 0$, since $P(X \leq B) = 1$, by Theorem 11.2.5. Hence $Q \not\ll P \implies D(Q\|P) = \infty$.

Third case: If $P(X = B) = 0$, then $X < B$ a.s. under P, and $Q \not\ll P$ for any Q s.t. $\mathbb{E}_Q[X] \geq B$. Then the minimum is ∞. Now assume $P(X = B) > 0$. Since $D(Q\|P) < \infty \implies Q \ll P \implies Q(X \leq B) = 1$. Therefore the only possibility for $\mathbb{E}_Q[X] \geq B$ is that $Q(X = B) = 1$, i.e., $Q = \delta_B$. Then $D(Q\|P) = \log \frac{1}{P(X=B)}$.

Second case: Fix $\mathbb{E}_P[X] \leq \gamma < B$, and find the unique λ such that $\psi_X'(\lambda) = \gamma = \mathbb{E}_{P_\lambda}[X]$ where $dP_\lambda = \exp(\lambda X - \psi_X(\lambda))dP$. This corresponds to tilting P far enough to the right to increase its mean from $\mathbb{E}_P X$ to γ, in particular $\lambda \geq 0$. Moreover, $\psi_X^*(\gamma) = \lambda\gamma - \psi_X(\lambda)$. Take any Q such that $\mathbb{E}_Q[X] \geq \gamma$, then

$$D(Q\|P) = \mathbb{E}_Q\left[\log \frac{dQ dP_\lambda}{dP dP_\lambda}\right] \tag{12.15}$$

$$= D(Q\|P_\lambda) + \mathbb{E}_Q[\log \frac{dP_\lambda}{dP}] \tag{12.16}$$

$$= D(Q\|P_\lambda) + \mathbb{E}_Q[\lambda X - \psi_X(\lambda)] \tag{12.17}$$

$$\geq D(Q\|P_\lambda) + \lambda\gamma - \psi_X(\lambda) \tag{12.18}$$

$$= D(Q\|P_\lambda) + \psi_X^*(\gamma) \tag{12.19}$$

$$\geq \psi_X^*(\gamma), \tag{12.20}$$

where the last inequality holds with equality if and only if $Q = P_\lambda$. In addition, this shows the minimizer is unique, proving the second claim. Note that even in the corner case of $\gamma = B$ (assuming $P(X = B) > 0$) the minimizer is a point mass $Q = \delta_B$, which is also a tilted measure (P_∞), since $P_\lambda \to \delta_B$ as $\lambda \to \infty$, cf. Theorem 11.4.3.

[2]Note that unlike previous Lecture, here P and P_λ are measures on an abstract space Ω, not on a real line.

Another version of the solution, given by expression (12.13), follows from Theorem 11.3.

For the third claim, notice that there is nothing to prove for $\gamma < \mathbb{E}_P[X]$, while for $\gamma \geq \mathbb{E}_P[X]$ we have just shown

$$\psi_X^*(\gamma) = \min_{Q:\mathbb{E}_Q[X]\geq\gamma} D(Q\|P)$$

while from the next corollary we have

$$\inf_{Q:\mathbb{E}_Q[X]>\gamma} D(Q\|P) = \inf_{\gamma'>\gamma} \psi_X^*(\gamma')\,.$$

The final step is to notice that ψ_X^* is increasing and continuous by Theorem 11.3, and hence the right-hand side infimum equalis $\psi_X^*(\gamma)$. The case of $\min_{Q:\mathbb{E}_Q[X]=\gamma}$ is handled similarly. □

Corollary 12.1. *$\forall Q$ with $\mathbb{E}_Q[X] \in (A, B)$, there exists a unique $\lambda \in \mathbb{R}$ such that the tilted distribution P_λ satisfies*

$$\mathbb{E}_{P_\lambda}[X] = \mathbb{E}_Q[X]$$
$$D(P_\lambda\|P) \leq D(Q\|P)$$

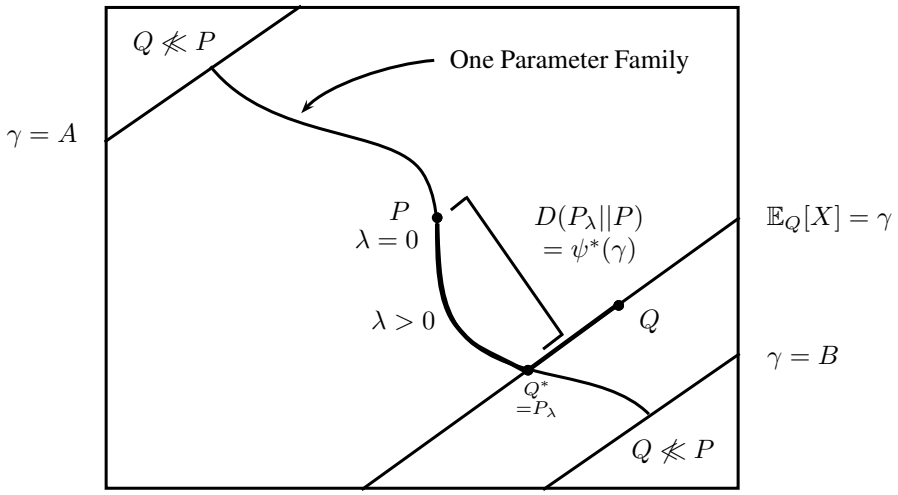
and furthermore the gap in the last inequality equals $D(Q\|P_\lambda) = D(Q\|P) - D(P_\lambda\|P)$.

Proof. Same as in the proof of Theorem 12.3, find the unique λ s.t. $\mathbb{E}_{P_\lambda}[X] = \psi_X'(\lambda) = \mathbb{E}_Q[X]$. Then $D(P_\lambda\|P) = \psi_X^*(\mathbb{E}_Q[X]) = \lambda\mathbb{E}_Q[X] - \psi_X(\lambda)$. Repeat the steps (12.15)-(12.20) obtaining $D(Q\|P) = D(Q\|P_\lambda) + D(P_\lambda\|P)$. □

Remark: For any Q, this allows us to find a tilted measure P_λ that has the same mean yet smaller (or equal) divergence.

12.3 Interpretation of Information Projection

The following picture describes many properties of information projections.



- Each set $\{Q : \mathbb{E}_Q[X] = \gamma\}$ corresponds to a slice. As γ varies from A to B, the curves fill the entire space minus the corner regions.

- When $\gamma < A$ or $\gamma > B$, $Q \not\ll P$.
- As γ varies, the P_λ's trace out a curve via $\psi^*(\gamma) = D(P_\lambda \| P)$. This set of distributions is called a *one parameter family*, or *exponential family*.

Key Point: The one parameter family curve intersects each γ-slice $\mathcal{E} = \{Q : \mathbb{E}_Q[X] = \gamma\}$ "orthogonally" at the minimizing $Q^* \in \mathcal{E}$, and the distance from P to Q^* is given by $\psi^*(\lambda)$. To see this, note that applying Theorem 12.2 to the convex set $\mathcal{E}$ gives us $D(Q\|P) \geq D(Q\|Q^*) + D(Q^*\|P)$. Now thanks to Corollary 12.1, we in fact have *equality* $D(Q\|P) = D(Q\|Q^*) + D(Q^*\|P)$ and $Q^* = P_\lambda$ for some tilted measure.

Chernoff bound revisited: The proof of upper bound in Theorem 12.1 is via the definition of information projection. Theorem 12.3 shows that the value of the information projection coincides with the rate function (conjugate of log-MGF). This shows the optimality of the Chernoff bound (recall Theorem 11.2.7). Indeed, we directly verify this for completeness: For all $\lambda \geq 0$,

$$P\left[\sum_{k=1}^{n} X_k \geq n\gamma\right] \leq e^{-n\gamma\lambda}(\mathbb{E}_P[e^{\lambda X}])^n = e^{-n(\lambda\gamma - \psi_X(\lambda))}$$

where we used iid X_k's in the expectation. Optimizing over $\lambda \geq 0$ to get the best upper bound:

$$\sup_{\lambda \geq 0} \lambda\gamma - \psi_X(\lambda) = \sup_{\lambda \in \mathbb{R}} \lambda\gamma - \psi_X(\lambda) = \psi_X^*(\gamma)$$

where the first equality follows since $\gamma \geq \mathbb{E}_P[X]$, therefore $\lambda \mapsto \lambda\gamma - \psi_X(\lambda)$ is increasing when $\lambda \leq 0$.

Remark: The Chernoff bound is tight precisely because, from information projection, the lower bound showed that the best change of measure is to change to the tilted measure P_λ.

12.4 Generalization: Sanov's theorem

Theorem 12.4 (Sanov's Theorem)**.** *Consider observing n samples $X_1, \ldots, X_n \sim$ iid P. Let $\hat{P}$ be the empirical distribution, i.e., $\hat{P} = \frac{1}{n}\sum_{j=1}^{n} \delta_{X_j}$. Let $\mathcal{E}$ be a convex set of distributions. Then under regularity conditions on $\mathcal{E}$ and P we have*

$$\mathbb{P}[\hat{P} \in \mathcal{E}] = e^{-n \min_{Q \in \mathcal{E}} D(Q\|P) + o(n)}$$

Note: Examples of regularity conditions: space $\mathcal{X}$ is finite and $\mathcal{E}$ is closed with non-empty interior; space $\mathcal{X}$ is Polish and the set $\mathcal{E}$ is weakly closed and has non-empty interior.

Proof sketch. The lower bound is proved as in Theorem 12.1: Just take an arbitrary $Q \in \mathcal{E}$ and apply a suitable version of WLLN to conclude $Q^n[\hat{P} \in \mathcal{E}] = 1 + o(1)$.

For the upper bound we can again adapt the proof from Theorem 12.1. Alternatively, we can write the convex set $\mathcal{E}$ as an intersection of half spaces. Then we've already solved the problem for half-spaces $\{Q : \mathbb{E}_Q[X] \geq \gamma\}$. The general case follows by the following consequence of Theorem 12.2: if Q^* is projection of P onto $\mathcal{E}_1$ and Q^{**} is projection of Q^* on $\mathcal{E}_2$, then Q^{**} is also projection of P onto $\mathcal{E}_1 \cap \mathcal{E}_2$:

$$D(Q^{**}\|P) = \min_{Q \in \mathcal{E}_1 \cap \mathcal{E}_2} D(Q\|P) \Leftarrow \begin{cases} D(Q^*\|P) = \min_{Q \in \mathcal{E}_1} D(Q\|P) \\ D(Q^{**}\|Q^*) = \min_{Q \in \mathcal{E}_2} D(Q\|Q^*) \end{cases}$$

(Repeated projection property)

Indeed, by first tilting from P to Q^* we find

$$P[\hat{P} \in \mathcal{E}_1 \cap \mathcal{E}_2] \leq 2^{-nD(Q^*\|P)} Q^*[\hat{P} \in \mathcal{E}_1 \cap \mathcal{E}_2] \tag{12.21}$$

$$\leq 2^{-nD(Q^*\|P)} Q^*[\hat{P} \in \mathcal{E}_2] \tag{12.22}$$

and from here proceed by tilting from Q^* to Q^{**} and note that $D(Q^*\|P) + D(Q^{**}\|Q^*) = D(Q^{**}\|P)$. □

Remark: Sanov's theorem tells us the probability that, after observing n iid samples of a distribution, our empirical distribution is still far away from the true distribution, is exponentially small.

MIT OpenCourseWare
https://ocw.mit.edu

6.441 Information Theory

Spring 2016

For information about citing these materials or our Terms of Use, visit: https://ocw.mit.edu/terms.