§ 11. Hypothesis testing asymptotics I

Setup:

$$\begin{aligned}
&H_0 : X^n \sim P_{X^n} \qquad H_1 : X^n \sim Q_{X^n} \\
&\text{test } P_{Z|X^n} : \mathcal{X}^n \to \{0,1\} \\
&\text{specification } 1-\alpha = \pi_{1|0} \qquad \beta = \pi_{0|1}
\end{aligned}$$

11.1 Stein's regime

$$\begin{aligned}
&1-\alpha = \pi_{1|0} \le \epsilon \\
&\beta = \pi_{0|1} \to 0 \quad \text{at the rate } 2^{-nV_\epsilon}
\end{aligned}$$

Note: interpretation of this specification, usually a "miss"(0|1) is much worse than a "false alarm" (1|0).

Definition 11.1 (ϵ-optimal exponent)**.** V_ϵ is called an *ϵ-optimal exponent in Stein's regime* if

$$\begin{aligned}
&V_\epsilon = \sup\{E : \exists n_0, \forall n \ge n_0, \exists P_{Z|X^n} \text{ s.t. } \alpha > 1-\epsilon, \beta < 2^{-nE}, \} \\
\Leftrightarrow\ &V_\epsilon = \liminf_{n\to\infty} \frac{1}{n} \log \frac{1}{\beta_{1-\epsilon}(P_{X^n}, Q_{X^n})}
\end{aligned}$$

where $\beta_\alpha(P,Q) = \min_{P_{Z|X}, P(Z=0)\ge\alpha} Q(Z=0)$.

Exercise: Check the equivalence.

Definition 11.2 (Stein's exponent)**.**

$$V = \lim_{\epsilon\to 0} V_\epsilon.$$

Theorem 11.1 (Stein's lemma)**.** *Let $P_{X^n} = P_X^n$ i.i.d. and $Q_{X^n} = Q_X^n$ i.i.d. Then*

$$V_\epsilon = D(P\|Q), \quad \forall \epsilon \in (0,1).$$

Consequently,

$$V = D(P\|Q).$$

Example: If it is required that $\alpha \ge 1-10^{-3}$, and $\beta \le 10^{-40}$, what's the number of samples needed? Stein's lemma provides a rule of thumb: $n \gtrsim -\frac{\log 10^{-40}}{D(P\|Q)}$.

Proof. Denote $F = \log \frac{dP}{dQ}$, and $F_n = \log \frac{dP_{X^n}}{dQ_{X^n}} = \sum_{i=1}^n \log \frac{dP}{dQ}(X_i)$ – iid sum.

Recall Neyman Pearson's lemma on optimal tests (likelihood ratio test): $\forall \tau$,

$$\alpha = P(F > \tau), \quad \beta = Q(F > \tau) \le e^{-\tau}$$

Also notice that by WLLN, under P, as $n \to \infty$,

$$\frac{1}{n} F_n = \frac{1}{n} \sum_{i=1}^n \log \frac{dP(X_i)}{dQ(X_i)} \xrightarrow{\mathbb{P}} \mathbb{E}_P \left[\log \frac{dP}{dQ} \right] = D(P\|Q). \tag{11.1}$$

Alternatively, under Q, we have

$$\frac{1}{n} F_n \xrightarrow{\mathbb{P}} \mathbb{E}_Q[\log \frac{dP}{dQ}] = -D(Q\|P) \tag{11.2}$$

1. Show $V_\epsilon \ge D(P\|Q) = D$.
   Pick $\tau = n(D - \delta)$, for some small $\delta > 0$. Then the optimal test achieves:

   $$\alpha = P(F_n > n(D - \delta)) \to 1, \text{ by (11.1)}$$
   $$\beta \le e^{-n(D-\delta)}$$

   then pick n large enough (depends on ϵ, δ) such that $\alpha \ge 1 - \epsilon$, we have the exponent $E = D - \delta$ achievable, $V_\epsilon \ge E$. Further let $\delta \to 0$, we have that $V_\epsilon \ge D$.

2. Show $V_\epsilon \le D(P\|Q) = D$.

   a) (weak converse) $\forall (\alpha, \beta) \in \mathcal{R}(P_{X^n}, Q_{X^n})$, we have

   $$-h(\alpha) + \alpha \log \frac{1}{\beta} \le d(\alpha\|\beta) \le D(P_{X^n}\|Q_{X^n}) \tag{11.3}$$

   where the first inequality is due to

   $$d(\alpha\|\beta) = \alpha \log \frac{\alpha}{\beta} + \bar{\alpha} \log \frac{\bar{\alpha}}{\bar{\beta}} = -h(\alpha) + \alpha \log \frac{1}{\beta} + \underbrace{\bar{\alpha} \log \frac{1}{\bar{\beta}}}_{\ge 0 \text{ and } \approx 0 \text{ for small } \beta}$$

   and the second is due to the weak converse Theorem 10.4 proved in the last lecture (data processing inequality for divergence).
   $\forall$ achievable exponent $E < V_\epsilon$, by definition, there exists a sequence of tests $P_{Z|X^n}$ such that $\alpha_n \ge 1 - \epsilon$ and $\beta_n \le 2^{-nE}$. Plugging it in (11.3) and using $h \le \log 2$, we have

   $$-\log 2 + (1 - \epsilon) nE \le nD(P\|Q) \Rightarrow E \le \frac{D(P\|Q)}{1 - \epsilon} + \underbrace{\frac{\log 2}{n(1 - \epsilon)}}_{\to 0, \text{ as } n \to \infty}.$$

   Therefore

   $$V_\epsilon \le \frac{D(P\|Q)}{1 - \epsilon}$$

   Notice that this is weaker than what we hoped to prove, and this weak converse result is tight for $\epsilon \to 0$, i.e., for Stein's exponent we did have the desired result $V = \lim_{\epsilon \to 0} V_\epsilon \ge D(P\|Q)$.

b) (strong converse) In proving the weak converse, we only made use of the *expectation* of F_n in (11.3), we need to make use of the *entire distribution (CDF)* in order to obtain stronger results.

Recall the strong converse result which we showed in the last lecture:

$$\forall(\alpha,\beta)\in\mathcal{R}(P,Q),\forall\gamma,\quad \alpha-\gamma\beta\le P(F>\log\gamma)$$

Here, suppose there exists a sequence of tests $P_{Z|X_n}$ which achieve $\alpha_n\ge 1-\epsilon$ and $\beta_n\le 2^{-nE}$. Then

$$1-\epsilon-\gamma 2^{-nE}\le\alpha_n-\gamma\beta_n\le P_{X^n}[F_n>\log\gamma].$$

Pick $\log\gamma=n(D+\delta)$, by (11.1) the RHS goes to 0, and we have

$$\begin{aligned}
&1-\epsilon-2^{n(D+\delta)}2^{-nE}\le o(1)\\
\Rightarrow & D+\delta-E\ge\frac{1}{n}\log(1-\epsilon+o(1))\to 0\\
\Rightarrow & E\le D \text{ as } \delta\to 0\\
\Rightarrow & V_\epsilon\le D
\end{aligned}$$

□

Note: [Ergodic] Just like in last section of data compression. Ergodic assumptions on P_{X^n} and Q_{X^n} allow one to show that

$$V_\epsilon=\lim_{n\to\infty}\frac{1}{n}D(P_{X^n}\|Q_{X^n})$$

the counterpart of (11.3), which is the key for picking the appropriate τ, for ergodic sequence X^n is the Birkhoff-Khintchine convergence theorem.

Note: The theoretical importance of knowing the Stein's exponents is that:

$$\forall E\subset\mathcal{X}^n,\quad P_{X^n}[E]\ge 1-\epsilon\quad\Rightarrow Q_{X^n}[E]\ge 2^{-nV_\epsilon+o(n)}$$

Thus knowledge of Stein's exponent V_ϵ allows one to prove exponential bounds on probabilities of arbitrary sets, the technique is known as "change of measure".

11.2 Chernoff regime

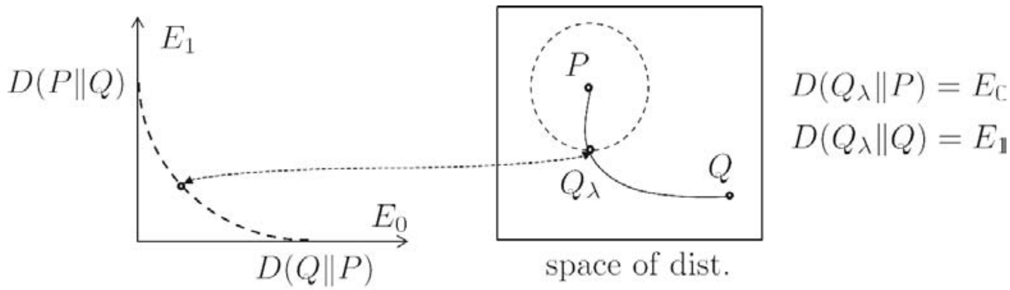
We are still considering i.i.d. sequence X^n, and binary hypothesis

$$H_0: X^n\sim P_X^n\qquad H_1: X^n\sim Q_X^n$$

But our objective in this section is to have both types of error probability to vanish exponentially fast simultaneously. We shall look at the following specification:

$$\begin{aligned}
1-\alpha=\pi_{1|0}\to 0 &\quad\text{at the rate } 2^{-nE_0}\\
\beta=\pi_{0|1}\to 0 &\quad\text{at the rate } 2^{-nE_1}
\end{aligned}$$

Apparently, E_0 (resp. E_1) can be made arbitrarily big at the price of making E_1 (resp. E_0) arbitrarily small. So the problem boils down to the optimal tradeoff, i.e., what's the achievable region of (E_0,E_1)? This problem is solved by [Hoeffding '65], [Blahut '74].



characterize the boundary of the achievable region of (E_0, E_1)

The optimal tests give the explict error probability:

$$\alpha_n = P\left[\frac{1}{n}F_n > \tau\right], \quad \beta_n = Q\left[\frac{1}{n}F_n > \tau\right]$$

and we are interested in the asymptotics when $n \to \infty$, in which scenario we know (11.1) and (11.2) occur.

Stein's regime corresponds to the corner points. Indeed, Theorem 11.1 tells us that when fixing $\alpha_n = 1 - \epsilon$, namely $E_0 = 0$, picking $\tau = D(P\|Q) - \delta$ $(\delta \to 0)$ gives the exponential convergence rate of β_n as $E_1 = D(P\|Q)$. Similarly, exchanging the role of P and Q, we can achieves the point $(E_0, E_1) = (D(Q\|P), 0)$. More generally, to achieve the optimal tradeoff between the two corner points, we need to introduce a powerful tool – Large Deviation Theory.

Note: Here is a roadmap of the upcoming 2 lectures:

1. basics of large deviation (ψ_X, ψ_X^*, tilted distribution P_λ)
2. information projection problem

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

3. use information projection to prove tight Chernoff bound

$$\mathbb{P}\left[\frac{1}{n}\sum_{k=1}^{n} X_k \geq \gamma\right] = 2^{-n\psi^*(\gamma)+o(n)}$$

4. apply the above large deviation theorem to (E_0, E_1) to get

$$(E_0(\theta) = \psi_P^*(\theta), \quad E_1(\theta) = \psi_P^*(\theta) - \theta) \quad \text{characterize the achievable boundary.}$$

11.3 Basics of Large deviation theory

Let X^n be an i.i.d. sequence and $X_i \sim P$. Large deviation focuses on the following inequality:

$$P\left[\sum_{i=1}^{n} X_i \geq n\gamma\right] = 2^{-nE(\gamma)+o(n)}$$

what is the rate function $E(\gamma) = -\lim_{n\to\infty} \frac{1}{n} \log P\left[\frac{\sum_{i=1}^n X_i}{n} \geq \gamma\right]$? (Chernoff's ineq.)

To motivate, let us recall the usual Chernoff bound: For iid X^n, for any $\lambda \geq 0$,

$$\begin{aligned}\mathbb{P}\left[\sum_{i=1}^n X_i \geq n\gamma\right] &= \mathbb{P}\left[\exp\left(\lambda\sum_{i=1}^n X_i\right) \geq \exp(n\lambda\gamma)\right] \\ &\overset{\text{Markov}}{\leq} \exp(-n\lambda\gamma)\mathbb{E}\left[\exp\left(\lambda\sum_{i=1}^n X_i\right)\right] \\ &= \exp\left\{-n\lambda\gamma + n\log\mathbb{E}\left[\exp(\lambda X)\right]\right\}.\end{aligned}$$

Optimizing over $\lambda \geq 0$ gives the *non-asymptotic* upper bound (concentration inequality) which holds for any n:

$$\mathbb{P}\left[\sum_{i=1}^n X_i \geq n\gamma\right] \leq \exp\Big\{-n\sup_{\lambda\geq 0}(\lambda\gamma - \underbrace{\log\mathbb{E}\left[\exp(\lambda X)\right]}_{\text{log MGF}})\Big\}.$$

Of course we still need to show the lower bound.

Let's first introduce the two key quantities: *log MGF* (also known as the *cumulant generating function*) $\psi_X(\lambda)$ and *tilted distribution* P_λ.

11.3.1 log MGF

Definition 11.3 (log MGF)**.**

$$\psi_X(\lambda) = \log(\mathbb{E}[\exp(\lambda X)]), \quad \lambda \in \mathbb{R}.$$

Per the usual convention, we will also denote $\psi_P(\lambda) = \psi_X(\lambda)$ if $X \sim P$.

Assumptions: In this section, we shall restrict to the distribution P_X such that

1. MGF exists, i.e., $\forall\lambda\in\mathbb{R}, \psi_X(\lambda) < \infty$,
2. $X \neq$const.

Example:

- Gaussian: $X \sim \mathcal{N}(0,1) \Rightarrow \psi_X(\lambda) = \frac{\lambda^2}{2}$.
- Example of R.V. such that $\psi_X(\lambda)$ does not exist: $X = Z^3$ with $Z \sim$ Gaussian. Then $\psi_X(\lambda) = \infty, \forall\lambda] \neq 0$.

Theorem 11.2 (Properties of ψ_X)**.**

1. *ψ_X is convex;*
2. *ψ_X is continuous;*
3. *ψ_X is infinitely differentiable and*

$$\psi_X'(\lambda) = \frac{\mathbb{E}[Xe^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]} = e^{-\psi_X(\lambda)}\mathbb{E}[Xe^{\lambda X}].$$

   In particular, $\psi_X(0) = 0, \psi_X'(0) = \mathbb{E}[X]$.

4. *If $a \leq X \leq b$ a.s., then $a \leq \psi_X' \leq b$;*

5. *Conversely, if*
$$A = \inf_{\lambda\in\mathbb{R}} \psi'_X(\lambda), \quad B = \sup_{\lambda\in\mathbb{R}} \psi'_X(\lambda),$$
then $A \le X \le B$ *a.s.;*

6. ψ_X *is strictly convex, and consequently,* ψ'_X *is strictly increasing.*

7. *Chernoff bound:*
$$P(X \ge \gamma) \le \exp(-\lambda\gamma + \psi_X(\lambda)), \quad \lambda \ge 0.$$

Remark 11.1. The slope of log MGF encodes the range of X. Indeed, 4) and 5) of Theorem 11.2 together show that the smallest closed interval containing the support of P_X equals (closure of) the range of ψ'_X. In other words, A and B coincide with the essential infimum and supremum (min and max of RV in the probabilistic sense) of X respectively,

$$\begin{aligned} A &= \operatorname{essinf} X \triangleq \sup\{a : X \ge a \text{ a.s.}\} \\ B &= \operatorname{esssup} X \triangleq \inf\{b : X \le b \text{ a.s.}\} \end{aligned}$$

Proof. Note: 1–4 can be proved right now. 7 is the usual Chernoff bound. The proof of 5–6 relies on Theorem 11.4, which can be skipped for now.

1. Fix $\theta \in (0,1)$. Recall Holder's inequality:
$$\mathbb{E}[|UV|] \le \|U\|_p \|V\|_q, \quad \text{for } p, q \ge 1, \frac{1}{p} + \frac{1}{q} = 1$$
where the L_p-norm of RV is defined by $\|U\|_p = (\mathbb{E}|U|^p)^{1/p}$. Applying to $\mathbb{E}[e^{(\theta\lambda_1+\bar\theta\lambda_2)X}]$ with $p = 1/\theta, q = 1/\bar\theta$, we get
$$\mathbb{E}[\exp((\lambda_1/p + \lambda_2/q)X)] \le \|\exp(\lambda_1 X/p)\|_p \|\exp(\lambda_2 X/q)\|_q = \mathbb{E}[\exp(\lambda_1 X)]^\theta \mathbb{E}[\exp(\lambda_2 X)]^{\bar\theta},$$
i.e., $e^{\psi_X(\theta\lambda_1+\bar\theta\lambda_2)} \le e^{\psi_X(\lambda_1)\theta} e^{\psi_X(\lambda_2)\bar\theta}$.

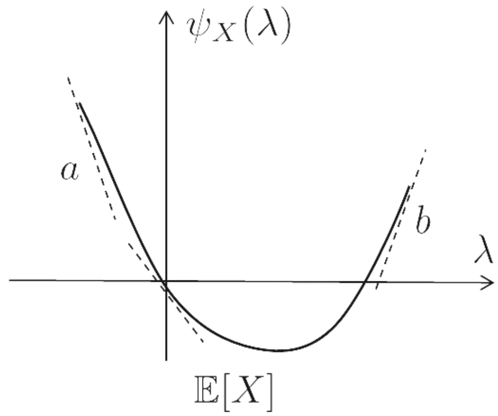
2. By our assumptions on X, domain of ψ_X is $\mathbb{R}$, and by the fact that convex function must be continuous on the interior of its domain, we have that ψ_X is continuous on $\mathbb{R}$.

3. Be careful when exchanging the order of differentiation and expectation.
Assume $\lambda > 0$ (similar for $\lambda \le 0$).
First, we show that $\mathbb{E}[|Xe^{\lambda X}|]$ exists. Since
$$\begin{aligned} &e^{|X|} \le e^X + e^{-X} \\ &|Xe^{\lambda X}| \le e^{|(\lambda+1)X|} \le e^{(\lambda+1)X} + e^{-(\lambda+1)X} \end{aligned}$$
by assumption on X, both of the summands are absolutely integrable in X. Therefore by dominated convergence theorem (DCT), $\mathbb{E}[|Xe^{\lambda X}|]$ exists and is continuous in λ.
Second, by the existence and continuity of $\mathbb{E}[|Xe^{\lambda X}|]$, $u \mapsto \mathbb{E}[|Xe^{uX}|]$ is integrable on $[0,\lambda]$, we can switch order of integration and differentiation as follows:
$$\begin{aligned} &e^{\psi_X(\lambda)} = \mathbb{E}[e^{\lambda X}] = \mathbb{E}\left[1 + \int_0^\lambda X e^{uX} du\right] \overset{\text{Fubini}}{=} 1 + \int_0^\lambda \mathbb{E}\left[X e^{uX}\right] du \\ \Rightarrow\ &\psi'_X(\lambda) e^{\psi_X(\lambda)} = \mathbb{E}[X e^{\lambda X}] \end{aligned}$$

thus $\psi'_X(\lambda) = e^{-\psi_X(\lambda)}\mathbb{E}[Xe^{\lambda X}]$ exists and is continuous in λ on $\mathbb{R}$.

Furthermore, using similar application of DCT we can extend to $\lambda \in \mathbb{C}$ and show that $\lambda \mapsto \mathbb{E}[e^{\lambda X}]$ is a holomorphic function. Thus it is infinitely differentiable.

4. $$a \le X \le b \Rightarrow \psi'_X(\lambda) = \frac{\mathbb{E}[Xe^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]} \in [a, b].$$



5. Suppose $P_X[X > B] > 0$ (for contradiction), then $P_X[X > B + 2\epsilon] > 0$ for some small $\epsilon > 0$. But then $P_\lambda[X \le B + \epsilon] \to 0$ for $\lambda \to \infty$ (see Theorem 11.4.3 below). On the other hand, we know from Theorem 11.4.2 that $\mathbb{E}_{P_\lambda}[X] = \psi'_X(\lambda) \le B$. This is not yet a contradiction, since P_λ might still have some very small mass at a very negative value. To show that this cannot happen, we first assume that $B - \epsilon > 0$ (otherwise just replace X with $X - 2B$). Next note that

$$\begin{aligned} B \ge \mathbb{E}_{P_\lambda}[X] &= \mathbb{E}_{P_\lambda}[X\mathbf{1}_{\{X<B-\epsilon\}}] + \mathbb{E}_{P_\lambda}[X\mathbf{1}_{\{B-\epsilon\le X\le B+\epsilon\}}] + \mathbb{E}_{P_\lambda}[X\mathbf{1}_{\{X>B+\epsilon\}}] \\ &\ge \mathbb{E}_{P_\lambda}[X\mathbf{1}_{\{X<B-\epsilon\}}] + \mathbb{E}_{P_\lambda}[X\mathbf{1}_{\{X>B+\epsilon\}}] \\ &\ge -\mathbb{E}_{P_\lambda}[|X|\mathbf{1}_{\{X<B-\epsilon\}}] + (B+\epsilon)\underbrace{P_\lambda[X > B+\epsilon]}_{\to 1} \end{aligned} \tag{11.4}$$

therefore we will obtain a contradiction if we can show that $\mathbb{E}_{P_\lambda}[|X|\mathbf{1}_{\{X<B-\epsilon\}}] \to 0$ as $\lambda \to \infty$. To that end, notice that convexity of ψ_X implies that $\psi'_X \nearrow B$. Thus, for all $\lambda \ge \lambda_0$ we have $\psi'_X(\lambda) \ge B - \frac{\epsilon}{2}$. Thus, we have for all $\lambda \ge \lambda_0$

$$\psi_X(\lambda) \ge \psi_X(\lambda_0) + (\lambda - \lambda_0)(B - \frac{\epsilon}{2}) = c + \lambda(B - \frac{\epsilon}{2}), \tag{11.5}$$

for some constant c. Then,

$$\begin{aligned} \mathbb{E}_{P_\lambda}[|X|1\{X < B-\epsilon\}] &= \mathbb{E}[|X|e^{\lambda X - \psi_X(\lambda)}1\{X < B-\epsilon\}] && (11.6) \\ &\le \mathbb{E}[|X|e^{\lambda X - c - \lambda(B-\frac{\epsilon}{2})}1\{X < B-\epsilon\}] && (11.7) \\ &\le \mathbb{E}[|X|e^{\lambda(B-\epsilon) - c - \lambda(B-\frac{\epsilon}{2})}] && (11.8) \\ &= \mathbb{E}[|X|]e^{-\lambda\frac{\epsilon}{2} - c} \to 0 \quad \lambda \to \infty && (11.9) \end{aligned}$$

where the first inequality is from (11.5) and the second from $X < B - \epsilon$. Thus, the first term in (11.4) goes to 0 implying the desired contradiction.

6. Suppose ψ_X is not strictly convex. Since we know that ψ_X is convex, then ψ_X must be "flat" (affine) near some point, i.e., there exists a small neighborhood of some λ_0 such that $\psi_X(\lambda_0 + u) = \psi_X(\lambda_0) + ur$ for some $r \in \mathbb{R}$. Then $\psi_{P_\lambda}(u) = ur$ for all u in small neighborhood of zero, or equivalently $\mathbb{E}_{P_\lambda}[e^{u(X-r)}] = 1$ for u small. The following Lemma 11.1 implies $P_\lambda[X = r] = 1$, but then $P[X = r] = 1$, contradicting the assumption $X \neq \text{const}$.

□

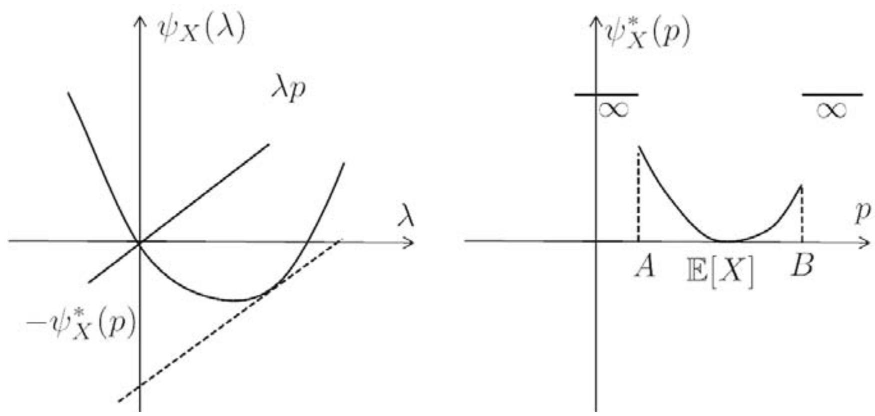
Lemma 11.1. $\mathbb{E}[e^{uS}] = 1$ *for all* $u \in (-\epsilon, \epsilon)$ *then* $S = 0$.

Proof. Expand in Taylor series around $u = 0$ to obtain $E[S] = 0$, $E[S^2] = 0$. Alternatively, we can extend the argument we gave for differentiating $\psi_X(\lambda)$ to show that the function $z \mapsto \mathbb{E}[e^{zS}]$ is holomorphic on the entire complex plane[1]. Thus by uniqueness, $\mathbb{E}[e^{uS}] = 1$ for all u. □

Definition 11.4 (Rate function)**.** The rate function $\psi_X^* : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ is given by the *Legendre-Fenchel transform* of the log MGF:

$$\psi_X^*(\gamma) = \sup_{\lambda \in \mathbb{R}} \lambda\gamma - \psi_X(\lambda) \tag{11.10}$$

Note: The maximization (11.10) is a nice convex optimization problem since ψ_X is strictly convex, so we are maximizing a strictly concave function. So we can find the maximum by taking the derivative and finding the stationary point. In fact, ψ_X^* is the *dual* of ψ_X in the sense of convex analysis.



Theorem 11.3 (Properties of ψ_X^*)**.**

1. *Let* $A = \operatorname{essinf} X$ *and* $B = \operatorname{esssup} X$. *Then*

$$\psi_X^*(\gamma) = \begin{cases} \lambda\gamma - \psi_X(\lambda) \text{ for some } \lambda \text{ s.t. } \gamma = \psi_X'(\lambda), & A < \gamma < B \\ \log \frac{1}{P(X=\gamma)} & \gamma = A \text{ or } B \\ +\infty, & \gamma < A \text{ or } \gamma > B \end{cases}$$

2. ψ_X^* *is strictly convex and strictly positive except* $\psi_X^*(\mathbb{E}[X]) = 0$.

3. ψ_X^* *is decreasing when* $\gamma \in (A, \mathbb{E}[X])$, *and increasing when* $\gamma \in [\mathbb{E}[X], B)$

[1]More precisely, if we only know that $\mathbb{E}[e^{\lambda S}]$ is finite for $|\lambda| \le 1$ then the function $z \mapsto \mathbb{E}[e^{zS}]$ is holomorphic in the vertical strip $\{z : |\text{Re}z| < 1\}$.

Proof. By Theorem 11.2.4, since $A \le X \le B$ a.s., we have $A \le \psi'_X \le B$. When $\gamma \in (A, B)$, the strictly concave function $\lambda \mapsto \lambda\gamma - \psi_X(\lambda)$ has a single stationary point which achieves the unique maximum. When $\gamma > B$ (resp. $< A$), $\lambda \mapsto \lambda\gamma - \psi_X(\lambda)$ increases (resp. decreases) without bounds. When $\gamma = B$, since $X \le B$ a.s., we have

$$\begin{aligned}\psi^*_X(B) &= \sup_{\lambda\in\mathbb{R}} \lambda B - \log(\mathbb{E}[\exp(\lambda X)]) = -\log \inf_{\lambda\in\mathbb{R}} \mathbb{E}[\exp(\lambda(X-B))] \\ &= -\log \lim_{\lambda\to\infty} \mathbb{E}[\exp(\lambda(X-B))] = -\log P(X=B),\end{aligned}$$

by monotone convergence theorem.

By Theorem 11.2.6, since ψ_X is strictly convex, the derivative of ψ_X and ψ^*_X are inverse to each other. Hence ψ^*_X is strictly convex. Since $\psi_X(0) = 0$, we have $\psi^*_X(\gamma) \ge 0$. Moreover, $\psi^*_X(\mathbb{E}[X]) = 0$ follows from $\mathbb{E}[X] = \psi'_X(0)$. □

11.3.2 Tilted distribution

As early as in Lecture 3, we have already introduced *tilting* in the proof of Donsker-Varadhan's variational characterization of divergence (Theorem 3.6). Let us formally define it now.

Definition 11.5 (Tilting)**.** Given $X \sim P$, the tilted measure P_λ is defined by

$$P_\lambda(dx) = \frac{e^{\lambda x}}{\mathbb{E}[e^{\lambda X}]} P(dx) = e^{\lambda x - \psi_X(\lambda)} P(dx) \tag{11.11}$$

In other words, if P has a pdf p, then the pdf of P_λ is given by $p_\lambda(x) = e^{\lambda x - \psi_X(\lambda)} p(x)$.

Note: The set of distributions $\{P_\lambda : \lambda \in \mathbb{R}\}$ parametrized by λ is called a *standard exponential family*, a very useful model in statistics. See [Bro86, p. 13].
Example:

- *Gaussian*: $P = \mathcal{N}(0,1)$ with density $p(x) = \frac{1}{\sqrt{2\pi}}\exp(-x^2/2)$. Then P_λ has density $\frac{\exp(\lambda x)}{\exp(\lambda^2/2)}\frac{1}{\sqrt{2\pi}}\exp(-x^2/2) = \frac{1}{\sqrt{2\pi}}\exp(-(x-\lambda)^2/2)$. Hence $P_\lambda = \mathcal{N}(\lambda, 1)$.
- *Binary*: P is uniform on $\{\pm 1\}$. Then $P_\lambda(1) = \frac{e^\lambda}{e^\lambda + e^{-\lambda}}$ which puts more (resp. less) mass on 1 if $\lambda > 0$ (resp. < 0). Moreover, $P_\lambda \xrightarrow{D} \delta_1$ if $\lambda \to \infty$ or δ_{-1} if $\lambda \to -\infty$.
- *Uniform*: P is uniform on $[0,1]$. Then P_λ is also supported on $[0,1]$ with pdf $p_\lambda(x) = \frac{\lambda\exp(\lambda x)}{e^\lambda - 1}$. Therefore as λ increases, P_λ becomes increasingly concentrated near 1, and $P_\lambda \to \delta_1$ as $\lambda \to \infty$. Similarly, $P_\lambda \to \delta_0$ as $\lambda \to -\infty$.

So we see that P_λ shifts the mean of P to the right (resp. left) when $\lambda > 0$ (resp. < 0). Indeed, this is a general property of tilting.

Theorem 11.4 (Properties of P_λ)**.**

1. *Log MGF:*
$$\psi_{P_\lambda}(u) = \psi_X(\lambda + u) - \psi_X(\lambda)$$
2. *Tilting trades mean for divergence:*
$$\mathbb{E}_{P_\lambda}[X] = \psi'_X(\lambda) \gtrless \mathbb{E}_P[X] \text{ if } \lambda \gtrless 0. \tag{11.12}$$
$$D(P_\lambda \| P) = \psi^*_X(\psi'_X(\lambda)) = \psi^*_X(\mathbb{E}_{P_\lambda}[X]). \tag{11.13}$$

3.

$$P(X > b) > 0 \Rightarrow \forall \epsilon > 0, P_\lambda(X \le b - \epsilon) \to 0 \ \textit{as} \ \lambda \to \infty$$
$$P(X < a) > 0 \Rightarrow \forall \epsilon > 0, P_\lambda(X \ge a + \epsilon) \to 0 \ \textit{as} \ \lambda \to -\infty$$

Therefore if $X_\lambda \sim P_\lambda$*, then* $X_\lambda \xrightarrow{\text{D}} \operatorname{essinf} X = A$ *as* $\lambda \to -\infty$ *and* $X_\lambda \xrightarrow{\text{D}} \operatorname{esssup} X = B$ *as* $\lambda \to \infty$*.*

Proof. 1. By definition. (DIY)

2. $\mathbb{E}_{P_\lambda}[X] = \frac{\mathbb{E}[X \exp(\lambda X)]}{\mathbb{E}[\exp(\lambda X)]} = \psi'_X(\lambda)$, which is strictly increasing in λ, with $\psi'_X(0) = \mathbb{E}_P[X]$.

   $D(P_\lambda \| P) = \mathbb{E}_{P_\lambda} \log \frac{dP_\lambda}{dP} = \mathbb{E}_{P_\lambda} \log \frac{\exp(\lambda X)}{\mathbb{E}[\exp(\lambda X)]} = \lambda \mathbb{E}_{P_\lambda}[X] - \psi_X(\lambda) = \lambda \psi'_X(\lambda) - \psi_X(\lambda) = \psi^*_X(\psi'_X(\lambda))$, where the last equality follows from Theorem 11.3.1.

3.

$$\begin{aligned}
P_\lambda(X \le b - \epsilon) &= \mathbb{E}_P[e^{\lambda X - \psi_X(\lambda)} \mathbf{1}[X \le b - \epsilon]] \\
&\le \mathbb{E}_P[e^{\lambda(b-\epsilon) - \psi_X(\lambda)} \mathbf{1}[X \le b - \epsilon]] \\
&\le e^{-\lambda\epsilon} e^{\lambda b - \psi_X(\lambda)} \\
&\le \frac{e^{-\lambda\epsilon}}{P[X > b]} \to 0 \text{ as } \lambda \to \infty
\end{aligned}$$

   where the last inequality is due to the usual Chernoff bound (Theorem 11.2.7): $P[X > b] \le \exp(-\lambda b + \psi_X(\lambda))$.

□