Electrical Engineering 229A Lecture 26 Notes

Daniel Raban

November 30, 2021

1 Convex Dual of the Cumulant Generating Function and Sanov's Theorem

1.1 The cumulant generating function and convex duality

Suppose $X \in \mathbb{R}^d$ is a random variable.

Definition 1.1. The map $\theta \mapsto \mathbb{E}[e^{\theta^\top X}]$ with $\theta \in \mathbb{R}^d$ is called the **moment generating function**.

Definition 1.2. The map $\theta \mapsto \log \mathbb{E}[e^{\theta^\top X}]$ with $\theta \in \mathbb{R}^d$ is called the **cumulant generating function**.

If we differentiate the moment generating function with respect to θ and set $\theta = 0$, we get the moments of X. Likewise, doing the same to the cumulant generating function gives us the cumulants of X. One advantage of working with the cumulant generating function is that it is convex.

We have dealt with finite (and countable) random variables and some densities. For a finite random variable $X \in \mathscr{X}$ with $|\mathscr{X}| = d$,it is interesting to consider $Z \in \mathbb{R}^d$ where $Z = e_i$ with probability p_i (here, e_i is the i-th basis vector). Then

$$\log \mathbb{E}[e^{\theta^\top Z}] = \log \sum_{i=1}^d p_i e^{\theta_i}$$

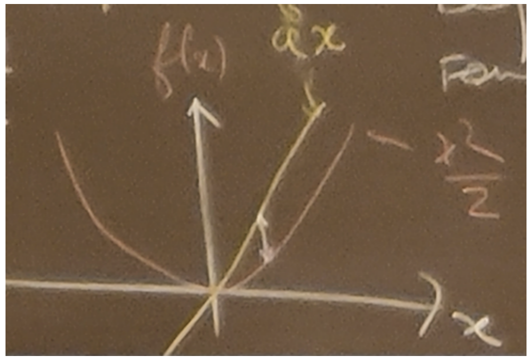
because $\theta^\top e_i = \theta_i$ for $i = 1, \dots, d$.

To any (extended real-valued) convex function there is a *dual*[1] convex function on $\mathbb{R}^d$.

Example 1.1. Let $d = 1$ and consider $f(x) = x^2/2$. Consider a line ax of slope ax and look at the height that separates the line from the function. Find the point at which this

[1] This is somtimes called Fenchel duality, Legendre duality, or Fenchel-Legendre duality.

height is the greatest to calculate the dual $\widehat{f}(a) := \sup_{x\in\mathbb{R}} ax - f(x)$.



Here, we can calculate $\widehat{f}(a) = a^2/2$. In a related sense to how the Gaussian is self-dual for the Fourier transform, this function is self-dual for the Frenchel-Legendre transform.

Example 1.2. Let $f(x) = e^x$. To find $\widehat{f}(a)$, since $f'(x) = a$ for x, if $a > 0$, this occurs if $x = \ln a$, and if $a \leq$, this is impossible. So we get

$$\begin{aligned}\widehat{f}(a) &= \sup_x (ax - e^x) \\ &= \begin{cases} a\ln a - a & a > 0 \\ 0 & a = 0 \\ \infty & a < 0. \end{cases}\end{aligned}$$

What if $d > 1$?

Definition 1.3. Suppose $\Phi : \mathbb{R}^d \mapsto \mathbb{R} \cup \{\infty\}$ is convex. Its **Fenchel-Lengendre dual** is

$$\widehat{\Phi}(a) := \sup_{x\in\mathbb{R}^d} a^\top x - \Phi(x)$$

for $a \in \mathbb{R}^d$.

Again,

$$\widehat{\Phi}(a) = a^\top x_a - \Phi(x_a),$$

where x_a is defined by $\nabla\Phi(x_a) = a$ (if x_a exists). It can be shown that

$$\Phi(x) = \sup_a x^\top a - \widehat{\Phi}(a).$$

To check this where Φ expresses all derivatives, write

$$\Phi(x) \geq x^\top a - \widehat{\Phi}(a) \quad \forall x, a \iff \widehat{\Phi}(a) \geq a^\top x - \Phi(x) \quad \forall x, a.$$

Proposition 1.1. *Let X take values in $\mathscr{X}$ with $|\mathscr{X}| = d$ and $p_i = \mathbb{P}(X = i)$. Let $Z = e_i$ iff $X = i$ (i.e. $P(Z = e_i) = p_i$ for $1 \le i \le d$). Then the Fenchel dual of $\Phi(\theta) = \ln \mathbb{E}[e^{\theta^\top Z}]$ is*

$$\widehat{\Phi}(a) = \begin{cases} D(a \,||\, p) & \text{if } a \text{ is a probability distribution} \\ \infty & \text{otherwise.} \end{cases}$$

Proof. Here,

$$\Phi_Z(\theta) = \ln \sum_{i=1}^d p_i e^{\theta_i},$$

so

$$\nabla \Phi_Z(\theta) = \begin{bmatrix} \frac{p_1 e^{\theta_1}}{\sum_{i=1}^d p_i e^{\theta_i}} \\ \vdots \end{bmatrix}.$$

This expresses only gradients that are probability distributions (means where $p_i \neq 0$). We have

$$\widehat{\Phi}_X(a) = a^\top p_a - \ln \sum_{i=1}^d p_i e^{\theta_{ai}},$$

where θ_a is defined in terms of a via $\nabla\Phi(\theta_a) = a$, i.e. $p_i e^{\theta_i}$ is proportional to a_i (i.e. $\theta_i = \ln \frac{a_i}{p_i} + \text{constant}$). The constant is $\log \sum_{i=1}^d p_i e^{(\theta_a)_i} = 0$.

$$\begin{aligned} &= \sum_{i=1}^d a_i \ln \frac{a_i}{p_i} - \ln \left(\cancelto{0}{\sum_{i=1}^d p_i e^{\ln \frac{a_i}{p_i}}} \right) \\ &= D(a \,||\, p). \end{aligned}$$

□

1.2 Large deviations and Sanov's theorem

Roughly speaking, a basic large deviations theory result is of the form: If $Z_1, Z_2, \dots$ are iid $\mathbb{R}^d$-valued with $\log \mathbb{E}[e^{\theta^\top Z}]$ denoted $\Phi_Z(\theta)$ and $\mathbb{E}[Z_1] = 0 \in \mathbb{R}^d$, then for any open set $A \subseteq \mathbb{R}^d$,

$$\liminf_{n\to\infty} -\frac{1}{n} \log \mathbb{P}\left(\frac{Z_1 + \cdots + Z_n}{n} \in A \right) \le \inf_{z \in A} \widehat{\Phi}_Z(z).$$

Here is a special case.

If $X_1, X_2, \dots,$ are i.i.d. $\mathscr{X}$-valued with $\mathscr{X} = \{1, 2, \dots, d\}$ and $Z_1, Z_2, \dots$ are i.i.d. $\mathbb{R}^d$-valued creased from $X_1, X_2, \dots,$ then observe that $\frac{Z_1 + \cdots + Z_n}{n}$ is equivalent to the empirical distribution of $(X_1, \dots, X_n)$, i.e. $\frac{Z_1 + \cdots + Z_n}{n} = \sum_{i=1}^d \frac{N(i|x^n)}{n} e_i$. Let $P_{x^n} := (\frac{N(i|x^n)}{n}, i = 1, \dots, d)$. So for any open subset $A \subseteq$ simplex in $\mathbb{R}^d$,

$$\liminf_n -\frac{1}{n} \log \mathbb{P}(P_{X^n} \in A) \le \inf_{a \in A} D(a \,||\, p).$$

Recall that if $x^n = (x_1, \dots, x_n) \in \mathscr{X}^n$ with $\mathscr{X}$ finite and if $\mathcal{P}$ denotes the set of probability distributions on X, then $p_{x^n} \in \mathcal{P}$ denotes $(\frac{N(x|x^n)}{n}, x \in \mathscr{X})$ and $\mathcal{P}_n$ denotes the set of all such P_{x_n}. For an n**-type** $P \in \mathcal{P}_n$, the **typicality set for** P refers to $T(P) := \{x^n \in \mathscr{X}^n : P_{x^n} = P\}$. For $Q \in \mathcal{P}$,

$$\begin{aligned} Q(x^n) &= \prod_{i=1}^{n} q(x_i) \\ &= \prod_{x \in X} q(x)^{N(x|x^n)} \\ &= 2^{-n(H(P_{x^n}) + D(P_{x^n} || Q))}. \end{aligned}$$

We also proved that for $P \in \mathcal{P}_n$,

$$P^n(T(P)) \geq P^n(T(\widetilde{P})) \qquad \forall \widetilde{P} \in \mathcal{P}_n,$$

$|P_n| \leq (n+1)^{|\mathscr{X}|}$, and for $P \in \mathcal{P}_n$,

$$\frac{1}{(n+1)^{|\mathscr{X}|}} 2^{nH(P)} \leq |T(P)| \leq 2^{nH(P)}.$$

Theorem 1.1 (Sanov)**.** *Let $\mathscr{X}$ be finite, $X_1, X_2, \dots \overset{\text{iid}}{\sim} Q$, and $E \subseteq \mathcal{P}$. Assume that E is the closure of its interior. Then*

$$\lim_{n \to \infty} \frac{1}{n} \log Q^n(P_{X^n} \in E) = -D(P^* \,||\, Q),$$

where

$$P^* = \operatorname*{arg\,min}_{P \in E} D(P \,||\, Q).$$

Remark 1.1. Since E is closed and $D(\cdot \,||\, Q)$ is continuous, this argmin exists. P^* is called the ***I*-projection** of Q onto E.

Proof. For the upper bound,

$$\begin{aligned} Q^n(P_{X^n} \in E) &= Q^n(P_{X^n} \in E \cap \mathcal{P}_n) \\ &\leq (n+1)^{|\mathscr{X}|} 2^{-nD(P^* || Q)} \end{aligned}$$

For the lower bound, for any $P^{(n)} \in \mathcal{P}_n \cap E$,

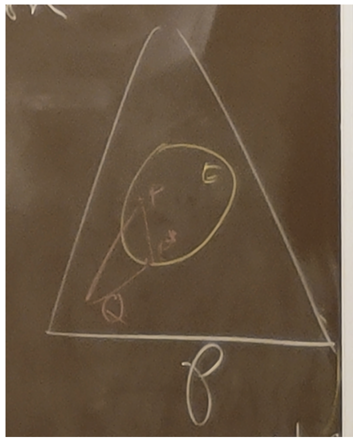
$$\begin{aligned} Q^n(P_{X^n} \in E) &\geq Q^n(T(P^{(n)})) \\ &\geq \frac{1}{(n+1)^{|\mathscr{X}|}} 2^{-nD(P^{(n)} || Q)}. \end{aligned}$$

Choose $P^{(n)} \to P^*$. ◻

Here is a nice observation about the I-projection of Q onto a *convex* set E.

Proposition 1.2. *For all $P \in E$,*

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



This tells us that we should think of $D(P \,||\, Q)$ as the *square* of a distance.

Proof. Consider the relative entropy $D(\lambda P + (1-\lambda)P^* \,||\, Q)$ for $\lambda \in [0,1]$. Differentiate in λ. It must be nonnegative. $\square$