INFORMATION THEORY

JOHN THICKSTUN

Introduction. These first half of these notes were inspired by a post on Terry Tao's blog[1]. In particular, I wanted to formalize his Example 7 from that blog post (it's worth reading his discussion; it is an great intuitive description of the argument). Basically everything up though the **Divergence** section of this document is an expansion upon this idea. I think this is also a pretty pedagogical way to introduce the various information functionals and prove their basic properties (contrast this with the standard approach of torturing these functionals with Jensen's inequality).

The second half of these notes (beginning with the **Topology**) discuss Sanov's theorem, and it's connection via duality to the Chernoff bound. The statement of Sanov's theorem and its proof via the method of types, given here for discrete random variables, are quite general; the statement holds unchanged for general probability measures and the proof also goes through essentially unchanged[2]. Duality only holds on convex subsets of the simplex; in this case there is a very elegant proof of a restricted Sanov-style result (Proposition 6[3]). In this convex setting, duality of the Sanov property and the Chernoff bound is established directly from the duality of KL-divergence (I-Projection) and the log-MGF (section **Duality**).

Many sources establish the KL version of the concentration bound by first deriving the Chernoff bound and then arguing by duality. This approach suggests that Chernoff is the "primal" version of the bound. The simple direct proof of Proposition 6 may provide some conceptual clarity; the KL version can be derived directly and has more right to the label "primal," in the sense that its optimization variables are the variables of interest. The final discussion of the Boltzmann's dice thought experiment show how the the state of the primal optimization variables at opt actually tells us something interesting about the state of a system in the event that concentration is violated (a "large deviation").

Entropy. Let $X \sim p$ be a discrete random variable on a finite state space $\mathcal{X}$ and define $X^{\otimes M} \equiv (X_1, \dots, X_M) \sim p^{\otimes M}$. By the weak law of large numbers, as $M \to \infty$,

$$-\frac{1}{M} \log p^{\otimes M}(X^{\otimes M}) = -\frac{1}{M} \sum_{i=1}^{M} \log p(X_i) \to_P - \mathop{\mathbb{E}}_{X \sim p} \log p(X) \equiv H(X).$$

[1] `https://terrytao.wordpress.com/2008/08/25/tricks-wiki-article-the-tensor-product-trick/`

[2] For a general exposition, see Csiszar's "A simple proof of Sanov's theorem."

[3] The proof is adapted from a fully general measure-theoretic version in Csiszar's "Sanov Property, Generalized I-Projection And a Conditional Limit Theorem," Theorem 1 (proof in Section 4)

This is the asymptotic equipartition principle: for large M, with high probability the likelihood of an observed sequence will be close to $e^{-MH(X)}$. Let $\epsilon > 0$ and define the typical set A_x to be the set of sequences $x \in \mathcal{X}^M$ such that

$$\left| -\frac{1}{M} \log p^{\otimes M}(x) - H(X) \right| \leq \epsilon.$$

Proposition 1. *For sufficiently large M, $(1-\epsilon)e^{M(H(X)-\epsilon)} \leq |A_x| \leq e^{MH(X)+\epsilon}$.*

Proof. For sufficiently large M, $\Pr(X^{\otimes M} \in A_x) > 1 - \epsilon$ and therefore

$$1 - \epsilon < \Pr(X^{\otimes M} \in A_x) = \sum_{x \in A_x} p^{\otimes M}(x) \leq \sum_{x \in A_x} e^{-M(H(X)-\epsilon)} = |A_x| e^{-M(H(X)-\epsilon)}.$$

This proves the lower bound. For the upper bound, observe that

$$1 \geq \sum_{x \in A_x} p^{\otimes M}(x) \geq \sum_{x \in A_x} e^{-M(H(X)+\epsilon)} = |A_x| e^{-M(H(X)+\epsilon)}. \qquad \square$$

For sufficiently large M, $X^{\otimes M}$ starts to behave like a uniform random variable on $e^{MH(X)}$ states. We immediately get an intuitive proof that $H(X) \leq \log |\mathcal{X}|$:

$$(1-\epsilon)e^{M(H(X)-\epsilon)} \leq |A_x| \leq |\mathcal{X}|^M = e^{M \log |\mathcal{X}|}.$$

Take M'th roots and letting $M \to \infty$, we have $e^{H(X)-\epsilon} \leq e^{M \log |\mathcal{X}|}$. This inequality holds for all $\epsilon > 0$, so we must have $H(X) \leq \log |\mathcal{X}|$. Likewise, for sufficiently large M, $|A_x| \geq 1$ and therefore $0 \leq H(X)$:

$$1 \leq |A_x| \leq e^{MH(X)+\epsilon}.$$

Joint Entropy. We can bound the joint entropy of a pair of random variables.

Proposition. $H(X,Y) \leq H(X) + H(Y)$.

Proof. Define the jointly typical set $J_{(x,y)}$ to be the set of sequences $(x,y) \in \mathcal{X}^M \times \mathcal{Y}^M$ such that $x \in A_x$, $y \in A_y$, and $(x,y) \in A_{(x,y)}$. By definition, $|J_{(x,y)}| \leq |A_x||A_y|$ and by the union bound, for sufficiently large M,

$$\Pr\left((X,Y)^{\otimes M} \in J_{(x,y)}\right) > 1 - \epsilon.$$

It follows by the same logic as the proposition 1 that

$$(1-\epsilon)e^{M(H(X,Y)-\epsilon)} \leq |J_{(x,y)}| \leq |A_x||A_y| \leq e^{M(H(X)+\epsilon)} e^{M(H(Y)+\epsilon)}.$$

Take M'th roots and let $M \to \infty$ to get

$$e^{H(X,Y)} \leq e^{H(X)} e^{H(Y)} = e^{H(X)+H(Y)}. \qquad \square$$

We can also give a lower bound $\max\{H(X), H(Y)\} \leq H(X,Y)$, but we need to introduce an additional idea before we can give a natural proof of this fact (see the next section on conditional entropy).

Define the mutual information $I(X;Y) \equiv H(X) + H(Y) - H(X,Y)$. Observe that $I(X;Y) \geq 0$ be the preceding proposition.

Proposition. *For sufficiently large M,*

$$(1-\epsilon)e^{-M(I(X;Y)-\epsilon)} \leq Pr\left((X^{\otimes M}, Y^{\otimes M}) \in A_{(x,y)}\right) \leq e^{-M(I(X;Y)+\epsilon)}.$$

Proof. For sufficiently large M,

$$\Pr\left((X^{\otimes M}, Y^{\otimes M}) \in A_{(x,y)}\right) = \sum_{(x,y)\in A_{(x,y)}} p^{\otimes M}(x)p^{\otimes M}(y)$$

$$\leq e^{MH(X,Y)+\epsilon/3}e^{-M(H(X)+\epsilon/3)}e^{-M(H(Y)+\epsilon/3)} = e^{-M(I(X;Y)+\epsilon)}.$$

And likewise for the lower bound. □

Conditional Entropy. Let $x \in A_x$ and define the conditional typical set

$$A_{y|x} \equiv \{y \in \mathcal{Y}^M : (x,y) \in J_{(x,y)}\}.$$

And further define the conditional entropy of Y given X,

$$H(Y|X) \equiv H(X,Y) - H(X).$$

Proposition. *For sufficiently large M,*

$$(1-\epsilon)e^{M(H(Y|X)-\epsilon)} \leq |A_{y|x}| \leq e^{M(H(Y|X)+\epsilon)}.$$

Proof. (Lower bound) By the union bound, for sufficiently large M,

$$\Pr(Y^{\otimes M} \in A_{y|x}) > 1-\epsilon.$$

It follows (for sufficiently large M) that

$$1-\epsilon < \Pr(Y^{\otimes M} \in A_{y|x}) = \sum_{y\in A_{y|x}} p^{\otimes M}(y|x) = \sum_{y\in A_{y|x}} \frac{p^{\otimes M}(x,y)}{p^{\otimes M}(x)}$$

$$\leq \sum_{y\in A_{y|x}} \frac{e^{-M(H(X,Y)+\epsilon/2)}}{e^{-M(H(X)-\epsilon/2)}} = |A_{y|x}|e^{-M(H(X,Y)-H(X)+\epsilon)}.$$

(Upper bound)

$$1 \geq \sum_{y\in A_{y|x}} p^{\otimes M}(y|x) \geq \sum_{y\in A_x} e^{-M(H(X,Y)-H(X)+\epsilon)} = |A_{y|x}|e^{-M(H(X,Y)-H(X)+\epsilon)}.$$ □

This is somewhat remarkable: the size of the conditional typical set concentrates at a rate that is independent of the element x that we condition on. As a corollary, for sufficiently large M, $|A_{y|x}| \geq 1$ and therefore $H(Y|X) \geq 0$. We also have $H(Y|X) \leq H(Y)$ from the observation that $|A_{y|x}| \leq |A_y|$ and the usual tensorization argument.

Types. Define the empirical distribution, or type, $\hat{p}_x$ of $x \in \mathcal{X}^n$ by

$$\hat{p}_x(a) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{x_i=a}.$$

It is not hard to see that the probability of a sequence x depends only on its type. We can explicitly compute this probability in terms of the cross-entropy,

$$H(p \parallel q) \equiv -\mathbb{E}_{X\sim p} \log q(X).$$

Proposition. *The probability of a sequence $x \in \mathcal{X}^M$ under measure $p^{\otimes M}$ is given by*

$$p^{\otimes M}(x) = e^{-MH(\hat{p}_x \parallel p)}.$$

Proof.

$$p^{\otimes M}(x) = \prod_{i=1}^{M} p(x_i) = \prod_{a\in\mathcal{X}} p(a)^{M\hat{p}_x(a)} = \prod_{a\in\mathcal{X}} e^{M\hat{p}_x(a)\log p(a)} = e^{M\sum_{a\in\mathcal{X}} \hat{p}_x(a)\log p(a)}. \qquad \square$$

Furthermore, because $1 \geq p^{\otimes M}(x)$, it follows by the usual tensorization argument that cross-entropy is non-negative (although some care is needed in this case to extend the result for empirical measures to all probability measures using a density argument: see section "Topology").

Type Classes. Type is an equivalence relation, so we may partition the set of length-n sequences by their type and define the type class T of a distribution q,

$$T(q) \equiv \{x \in \mathcal{X}^n : \hat{p}_x = q\}.$$

Let $\mathcal{P}_n$ denote the set of types of length-n sequences over an alphabet $\mathcal{X}$. Each type is uniquely indexed by $|\mathcal{X}|$ counts in the range from 0 to n, we can upper bound the number of types:

$$|\mathcal{P}_n| \leq (n+1)^{|\mathcal{X}|}.$$

Notably, because there are only polynomially many types but exponentially many sequences of length n, by the pigeonhole principle there must be at least one type class with exponentially many members. In fact, they almost all do. And the most populated type class contains all but an exponentially vanishing number of sequences.

Proposition. *For any $q \in \mathcal{P}_n$,*

$$e^{M(H(q)+o(1))} \leq |T(q)| \leq e^{MH(q)}.$$

Proof. For the upper bound, from the preceding proposition we have

$$1 \geq \sum_{x\in T(q)} q^{\otimes M}(x) = \sum_{x\in T(q)} e^{-MH(\hat{p}_x \parallel q)} = |T(q)|e^{-MH(q)}.$$

For the lower bound, observe that we can write down the exact size of $T(q)$ in terms of the multinomial coefficients:

$$|T(q)| = \binom{M}{Mq(a_1), Mq(a_2), \ldots, Mq(a_{|\mathcal{X}|})} = M! \left(\prod_{i=1}^{|\mathcal{X}|} (Mq(a_i))!\right)^{-1}.$$

Applying Stirling's approximation $\log n! = n \log n - n + O(\log n)$, we find that

$$\frac{1}{M} \log |T(q)| = H(q) + o(1). \qquad \square$$

Divergence. Let $\hat{p}_X$ denote the (random) type of $X^{\otimes M} \sim p^{\otimes M}$ and define the Kullback-Leibler divergence from q to p by

$$D(p \parallel q) \equiv \mathbb{E}_{X \sim p} \log \frac{p(X)}{q(X)}.$$

Proposition 2. *For any $q \in \mathcal{P}_n$,*

$$e^{-M(D(q \parallel p)+o(1))} \leq Pr(\hat{p}_X = q) \leq e^{-MD(q \parallel p)}.$$

Proof. First, observe that

$$\Pr(\hat{p}_X = q) = \sum_{x \in T(q)} p^{\otimes M}(x) = |T(q)| e^{-MH(q \parallel p)}.$$

The preceding proposition bounds $|T(q)|$; the result follows by the algebraic identity

$$H(q \parallel p) = H(p) + D(q \parallel p). \qquad \square$$

The usual tensorization argument establishes that KL-divergence is non-negative (again with an appeal to a density argument; see proposition 5 below).

Proposition 3. *$D(\hat{p}_X \parallel p) \to 0$ in probability at an exponential rate:*

$$Pr(D(\hat{p}_X \parallel p) > \epsilon) \leq e^{-M\left(\epsilon - |\mathcal{X}| \frac{\log(M+1)}{M}\right)}.$$

Proof. Let $E = \{q \in \mathcal{P}_n : D(q \parallel p) > \epsilon\}$. By proposition 2,

$$\Pr(D(\hat{p}_X \parallel p) > \epsilon) = \sum_{q \in E} \Pr(\hat{p}_X = q) \leq \sum_{q \in E} e^{-MD(q \parallel p)} \leq (M+1)^{|\mathcal{X}|} e^{-M\epsilon}. \qquad \square$$

Because the convergence is exponential, the Borel-Cantelli lemma allows us to upgrade convergence in probability to almost sure convergence.

Topology. Before proceeding further, we must pay some attention to topologies on the simplex. Because the type class $\mathcal{P}_n$ consists of discrete points, we need a topology to relate these points to nearby points in the simplex. We have deferred several "density" arguments already to avoid topological technicalities. But we can go no further: in the next section we will discuss Sanov's theorem, which cannot be properly stated without reference to a topology.

Unless otherwise specified, we will interpret topological concepts with respect to the standard topology: the subspace topology on $\Delta \subset \mathbb{R}^{|\mathcal{X}|}$ inherited from the standard topology on $\mathbb{R}^{|\mathcal{X}|}$ induced by the inner product.

Proposition 4. *The set of all types $\bigcup_{n \geq 1} \mathcal{P}_n$ is dense in the simplex.*

Proof. Let $p \in \Delta$ and $\epsilon > 0$. Because the rationals are dense in the reals, for each $a \in \mathcal{X}$ we can find $r_a \in \mathbb{Q}$ such that $|r_a - p(a)| \leq \epsilon/|\mathcal{X}|$. Let n be a common denominator of r_a and let $x \in \mathcal{X}^n$ be a sequence where each item a occurs exactly nr_a times. Then $\hat{p}_x \in \mathcal{P}_n$ and

$$\|\hat{p}_x - p\|_1 = \sum_{a \in \mathcal{X}} |r_a - p(a)| \leq \epsilon. \qquad \square$$

Now we can complete our proof from earlier of the information inequality.

Proposition 5. *(Information Inequality) For all $p, q \in \Delta$, $D(q \parallel p) \geq 0$.*

Proof. Recall (proposition 2) that if $q_n \in \mathcal{P}_n$,

$$e^{-MD(q_n \parallel p)+o(1)} \leq \Pr(\hat{p}_X = q_n) \leq 1.$$

Therefore $D(q_n \parallel p) \geq 0$ for all $q \in \bigcup_{n \geq 1} \mathcal{P}_n$. Because $\bigcup_{n \geq 1} \mathcal{P}_n$ is dense in the simplex, we can find a sequence $q_n \to q \in \Delta$. Observe that KL-divergence is continuous and therefore

$$D(q \parallel p) = \lim_{n \to \infty} D(q_n \parallel p) \geq 0. \qquad \square$$

We will occasionally also consider the topology of KL-divergence induced by $D(\cdot \parallel q)$. Pinsker's inequality guarantees that any ϵ-ball in the topology of total variation contains an ϵ-ball in the topology of KL-divergence, and the equivalence of norms on finite-dimensional vector spaces establishes that the topology of total variation is exactly the standard topology. Therefore the KL topology is stronger (finer) than the standard topology. Note that these topologies are not identical: suppose p is a vertex of the simplex, then an ϵ-ball around p in the KL topology is just p itself.

Sanov's Theorem. Suppose $E \subset \Delta$ contains an open neighborhood of p in the KL topology, then by proposition 3, as $M \to \infty$,

$$\Pr(\hat{p}_X \in E) \to 1.$$

Likewise, if E does not contain such a neighborhood then the probability vanishes. Our intuition suggests that if E is far from p then this probability will vanish more quickly than if E is near to p. Sanov's theorem formalizes this intuition, quantifying the the notions of near and far in the sense of KL divergence.

Theorem. *(Sanov) Let $p \in \Delta$ with $X \sim p^{\otimes M}$. If $E \subset \Delta$ then*

$$-\inf_{q \in \operatorname{int}(E)} D(q \parallel p) \leq \liminf_{M \to \infty} \frac{1}{M} Pr(\hat{p}_X \in E) \leq \limsup_{M \to \infty} \frac{1}{M} Pr(\hat{p}_X \in E) \leq -\inf_{q \in \operatorname{cl}(E)} D(q \parallel p).$$

Furthermore, if E contains the closure of its interior then

$$\lim_{M \to \infty} \frac{1}{M} \log Pr(\hat{p}_X \in E) = -\inf_{q \in E} D(q \parallel p).$$

Proof. (Upper bound) By proposition 2,

$$\Pr(\hat{p}_X \in E) = \sum_{q \in E \cap \mathcal{P}_M} \Pr(\hat{p}_X = q) \leq \sum_{q \in E \cap \mathcal{P}_M} e^{-MD(q \parallel p)}$$

$$\leq (M+1)^{|\mathcal{X}|} \sup_{q \in E} e^{-MD(q \parallel p)} = (M+1)^{|\mathcal{X}|} e^{-M \inf_{q \in E} D(q \parallel p)}.$$

(Lower bound) Because $\text{int}(E)$ is open, because $\bigcup_{n \geq 1} \mathcal{P}_n$ is dense in Δ (proposition 4) we can find a sequence $q_n \in \text{int}(E) \cap \mathcal{P}_n$ such that $D(q_n \parallel p) \to \inf_{q \in \text{int}(E)} D(q \parallel p)$. By proposition 2,

$$\Pr(\hat{p}_X \in E) \geq \Pr(\hat{p}_X \in \text{int}(E)) \geq \Pr(\hat{p}_X = q_n) \geq e^{-M(D(q_n \parallel p) + o(1))}.$$

For the final claim, observe that if E contains the closure of its interior then the upper and lower bounds coincide and the limit exists. □

Observe that this bound is apparently independent of the size of E, which could seem unintuitive. But this is consistent with our earlier observations: recall (proposition 2) that the probability of any type q vanishes at an exponential rate governed by $D(q \parallel p)$. So if we measure the size of E by the measure of $E \cap \mathcal{P}_M$ under $p^{\otimes M}$, then we see that only the nearest elements of $E \cap \mathcal{P}_M$ contribute any significant probability mass. Indeed, this idea is the essence of the proof.

Also observe that this result is asymptotic. While opening up the proof gives us finite-sample rates, these rates have quite bad polynomial factors introduced by our loose analysis of type classes. For convex E, in the upper bound at least, it is possible to completely eliminate this factor and give a hard upper bound on the probability of observing an event in E. We will see a precise statement of this claim at the end of the next section.

Convexity. Define the information projection of p onto E by

$$q^* \in \arg\min_{q \in E} D(q \parallel p).$$

Because the probability of probability mass of $E \cap \mathcal{P}_M$ concentrates near minimizers q^*, we will be interested in solving this optimization problem to find the high-probability neighborhood of empirical distributions conditioned on the event $\hat{p}_X \in E$.

Proposition. *The KL divergence $D(q \parallel p)$ is smooth and strongly convex in $q : \mathcal{X} \to \mathbb{R}$.*

Proof. Calculus. □

Let $f : \mathcal{X} \to \mathbb{R}^n$ be some function (i.e. measurement, test statistic) of the states $\mathcal{X}$. And suppose the empirical mean of these measurements exceeds some threshold $\alpha \in \mathbb{R}^n$:

$$\frac{1}{M} \sum_{i=1}^{M} f(X_i) = \sum_{a \in \mathcal{X}} \hat{p}_x(a) f(a) \geq \alpha.$$

These are linear inequality constraints on the simplex, which define a polytope

$$E \equiv \left\{ q \in \Delta : \mathop{\mathbb{E}}_{X \sim q} f(X) \geq \alpha \right\}.$$

In particular, E is compact, convex, and it follows by the preceding proposition that the information projection has a well-defined, unique minimizer.

Proposition 6. *Let $p \in \Delta$ with $X \sim p^{\otimes M}$. If $E \subset \Delta$ is convex then*

$$\frac{1}{M} \log Pr(\hat{p}_X \in E) \leq - \inf_{q \in E} D(q \parallel p).$$

Proof. Let $x \in \mathcal{X}^M$ and define the conditional distribution

$$\pi(x) \equiv \Pr\left(X = x | \hat{p}_X \in E\right) = \frac{p^{\otimes M}(x)}{\Pr(\hat{p}_X \in E)} \mathbb{1}_{\hat{p}_x \in E}.$$

Observe that

$$D(\pi \parallel p^{\otimes M}) = \sum_{x \in \mathcal{X}^M} \pi(x) \log \frac{\pi(x)}{p^{\otimes M}(x)}$$

$$= \sum_{\hat{p}_x \in E} \frac{p^{\otimes M}(x)}{\Pr(\hat{p}_X \in E)} \log \frac{1}{\Pr(\hat{p}_X \in E)} = - \log \Pr(\hat{p}_X \in E).$$

To prove the result, we must relate this divergence on the product space $\mathcal{X}^M$ to divergences on $\mathcal{X}$. Although π is not a product distribution, by symmetry it has identically distributed marginals $\tilde{\pi}$ and furthermore

$$\tilde{\pi}(a) = \sum_{x \in \mathcal{X}^M} \pi(x) \mathbb{1}_{x_i = a} = \frac{1}{M} \sum_{i=1}^{M} \sum_{x \in \mathcal{X}^M} \pi(x) \mathbb{1}_{x_i = a} = \sum_{\hat{p}_x \in E} \pi(x) \hat{p}_x(a).$$

Because E is convex, it follows that $\tilde{\pi} \in E$ and therefore

$$\inf_{q \in E} D(q \parallel p) \leq D(\tilde{\pi} \parallel p).$$

By algebraic manipulation,

$$M D(\tilde{\pi} \parallel p) = M \sum_{a \in \mathcal{X}} \tilde{\pi}(a) \log \frac{\tilde{\pi}(a)}{p(a)} = \sum_{a \in \mathcal{X}} \sum_{i=1}^{M} \sum_{x \in \mathcal{X}^M} \pi(x) \mathbb{1}_{x_i = a} \log \frac{\tilde{\pi}(a)}{p(a)}$$

$$= \sum_{x \in \mathcal{X}^M} \pi(x) \sum_{i=1}^{M} \log \frac{\tilde{\pi}(x_i)}{p(x_i)} = \sum_{x \in \mathcal{X}^M} \pi(x) \log \frac{\tilde{\pi}^{\otimes M}(x)}{p^{\otimes M}(x)} = D(\pi \parallel p^{\otimes M}) - D(\pi \parallel \tilde{\pi}^{\otimes M}).$$

We conclude

$$M \inf_{q \in E} D(q \parallel p) \leq M D(\tilde{\pi} \parallel p) \leq D(\pi \parallel p^{\otimes M}) = - \log \Pr\left(\hat{p}_X \in E\right). \qquad \square$$

Duality. We can compute this minimizer by introducing the Lagrangian

$$L(q, \lambda) \equiv D(q \parallel p) - \lambda^\top \left(\mathop{\mathbb{E}}_{X \sim q} f(X) - a \right).$$

By Sion's minimax theorem, we have strong duality:

$$\min_{q \in E} D(q \parallel p) = \min_{q \in \Delta} \sup_{\lambda \in \mathbb{R}^n_+} L(q, \lambda) = \sup_{\lambda \in \mathbb{R}^n_+} \min_{q \in \Delta} L(q, \lambda)$$

$$\sup_{\lambda \in \mathbb{R}^n_+} \left[\lambda^\top a + \min_{q \in \Delta} \left(D(q \parallel p) - \mathop{\mathbb{E}}_{X \sim q} \lambda^\top f(X) \right) \right].$$

We now turn our attention to the inner optimization.

Proposition. *The convex conjugate of $D(\,\cdot \parallel p)$ is given by $D^* : (\mathcal{X} \to \mathbb{R}) \to \mathbb{R}$,*

$$D^*(g) \equiv \max_{q \in \Delta} \left(\mathop{\mathbb{E}}_{X \sim q} g(X) - D(q \parallel p) \right) = \log \mathop{\mathbb{E}}_{X \sim p} e^{g(X)}.$$

Furthermore, this supremum is uniquely achieved at $q^* \in \Delta$ given by

$$q^*(a) = \frac{p(a) e^{g(a)}}{\sum_{a \in \mathcal{X}} p(a) e^{g(a)}}, \text{ for each } a \in \mathcal{X}.$$

Proof. Observe that the objective is a continuous function, optimized over a compact set, so it attains its supremum; i.e. a maximizer q^* exists. Rewriting our problem as a minimization, we have

$$D^*(g) = -\min_{q \in \Delta} \left(D(q \parallel p) - \mathop{\mathbb{E}}_{X \sim q} g(X) \right)$$

Because KL divergence is strongly convex and expectation is linear, this objective is strongly convex; therefore the minimizer q^* is unique. Consider the Lagrangian

$$L(q, \nu) \equiv \sum_{a \in \mathcal{X}} q(a) \log \frac{q(a)}{p(a)} - \sum_{a \in \mathcal{X}} q(a) g(a) - \nu \left(\sum_{a \in \mathcal{X}} q(a) - 1 \right).$$

The primal problem is clearly feasible and the simplex constraint is linear, i.e. Slater's condition holds. Therefore the duality gap is zero and moreover, the dual optimal value is attained by some solution ν^*:

$$D^*(g) = \inf_{q : \mathcal{X} \to \mathbb{R}} \sup_{\nu \in \mathbb{R}} L(q, \nu) = \sup_{\nu \in \mathbb{R}} \inf_{q : \mathcal{X} \to \mathbb{R}} L(q, \nu) = \inf_{q : \mathcal{X} \to \mathbb{R}} L(q, \nu^*).$$

The (unique) minimizer of L over q is given by the first-order optimality conditions

$$0 = \frac{\partial}{\partial q(a)} L(q, \nu) = 1 + \log q(a) - \log p(a) - g(a) - \nu, \text{ for each } a \in \mathcal{X}.$$

Therefore, for fixed ν, $q_\nu \equiv \arg\min_{q : \mathcal{X} \to \mathbb{R}} L(q, \nu)$ is given by

$$q_\nu(a) = p(a) e^{g(a) + \nu - 1}, \text{ for each } a \in \mathcal{X}.$$

Observe that because $D(q \parallel p)$ is strongly convex for all $q : \mathcal{X} \to \mathbb{R}$, a minimizer q_{ν^*} of $L(q, \nu^*)$ exists and is unique; it remains to show that $q^* = q_{\nu^*}$. By strong duality,

$$D^*(g) = L(q_{\nu^*}, \nu^*) = \inf_{q : \mathcal{X} \to \mathbb{R}} L(q, \nu^*) \leq L(q^*, \nu^*) = D^*(g).$$

Therefore $L(q_{\nu^*}, \nu^*) = L(q^*, \nu^*)$ and, because the solution q_{ν^*} is unique, $q^* = q_{\nu^*}$. It follows that $\sum_{a \in \mathcal{X}} q_\nu(a) = 1$ and consequently,

$$e^{1 - \nu} = \sum_{a \in \mathcal{X}} p(a) e^{g(a)}.$$

We conclude that

$$q^*(a) = \frac{p(a)e^{g(a)}}{\sum_{a\in\mathcal{X}} p(a)e^{g(a)}}, \text{ for each } a \in \mathcal{X}.$$

The optimal value is obtained by substitution (for algebra, it is convenient to work in terms of the parameterization q_ν):

$$D^*(g) = D(q^* \parallel p) - \mathop{\mathbb{E}}_{X\sim q^*} g(X)$$

$$= \sum_{a\in\mathcal{X}} q_\nu(a)(g(a)+\nu-1) - \sum_{a\in\mathcal{X}} q_\nu(a)g(a) = \nu - 1 = -\log \mathop{\mathbb{E}}_{X\sim p} e^{g(a)}. \qquad \square$$

As a corollary we see that $D^*(\lambda^\top f) = \mu(\lambda)$, where μ is the cumulant generating function of $f(X)$, and from proposition 6 we recover the familiar Chernoff bound

$$-\frac{1}{M}\log \Pr\left(\frac{1}{M}\sum_{i=1}^M f(X_i) \geq \alpha\right) \leq \inf_{q\in E} D(q \parallel p) = \sup_{\lambda\in\mathbb{R}^n_+} \left[\lambda^\top\alpha - \mu(\lambda)\right].$$

Moreover, the proposition gives us a solution to this optimization:

$$q_\lambda(a) = \frac{p(a)e^{\lambda^\top f(a)}}{\sum_{a\in\mathcal{X}} p(a)e^{\lambda^\top f(a)}}.$$

Optimization. Consider the dual problem

$$g(\lambda) \equiv \sup_{\lambda\in\mathbb{R}^N_+} L(q_\lambda, \lambda).$$

Because $D(q \parallel p)$ is smooth in q, g is strongly concave. It follows g has a unique maximizer λ^*, and by strong duality

$$\min_{q\in E} D(q \parallel p) = \min_{q\in\Delta} L(q,\lambda^*) \leq L(q^*,\lambda^*) \leq L(q^*,0) = \min_{q\in E} D(q \parallel p).$$

We conclude that $L(q_{\lambda^*},\lambda^*) = L(q^*,\lambda^*)$ and $q^* = q_{\lambda^*}$. We can explicitly find q^* given λ^* by projected gradient descent on the dual objective

$$\lambda^* = \arg\min_{\lambda\in\mathbb{R}^n_+} g(\lambda) = \arg\min_{\lambda\in\mathbb{R}^n_+} \left[\lambda^\top\alpha - \mu(\lambda)\right].$$

Boltzmann's Dice. As an example, take X to be a random variable on a six-element state space ($|\mathcal{X}| = 6$) which we will interpret as the state of a fair die, i.e. $X \sim \text{Uniform}(6)$. Let $f(a_i) = i$ for $i = 1,\dots,6$, the number written on face a_i of the die. Suppose we observe many (independent) successive rolls of this die. As $M\to\infty$, by the law of large numbers the empirical mean $S(X)$ of the die rolls will converge to its expectation:

$$\lim_{M\to\infty} S(X) = \lim_{M\to\infty} \frac{1}{M}\sum_{i=1}^M f(X_i) \to_P \mathbb{E}X = 3.5.$$

But for a fixed, finite number of rolls, the empirical mean will have a random distribution (which converges to a normal according to the central limit theorem at a rate bounded by

Berry-Esseen-type results). With high probability, the empirical distribution of rolls will converge to the uniform distribution (with rates controlled by concentration arguments).

Suppose we witness a large deviation. For example, after some large finite number of rolls, we observe an empirical mean no less than 4. In principle we could precisely compute the probability of this event by summing the probabilities of all outcomes that satisfy this condition; i.e.

$$\Pr\left(S(X) \geq 4\right) = \sum_{x \in \mathcal{X}} p(x) \mathbb{1}_{S(x) \geq 4}.$$

But this sum grows exponentially in M and quickly becomes intractable; bounding the sum is more practical. Recall Hoeffding's bound on the moment generating function of a random variable bounded on an interval of length c:

$$\mathbb{E} e^{\lambda X} \leq e^{\lambda^2 c^2/8}.$$

Plugging this into the Chernoff bound and optimizing over λ gives us

$$\Pr\left(S(X) \geq \mathbb{E}X + t\right) \leq e^{-2M^2 t^2/c^2} = e^{-M^2/50}.$$

Sanov's theorem allows us to say something stronger. Let $E \subset \Delta$ be the set of distributions such that $\mathbb{E}_{X \sim q} X \geq 4$ iff $q \in E$. For sufficiently large M,

$$\frac{1}{M} \log \Pr(S(X) \geq 4) \leq -D(q^* \parallel p).$$

Recall that q^* is the information projection of p onto E. The probability of observing an empirical distribution bounded away from q^* drops of exponentially as $M \to \infty$ (see the discussion following Sanov's theorem) so with high probability the empirical distribution of die rolls lies in a small neighborhood of q^*. In the preceding section, we saw that q^* has the form

$$q^*(j) = \frac{p(a_j) e^{\lambda^* f(a_j)}}{\sum_{a_i \in \mathcal{X}} p(a_i) e^{\lambda^* f(a_i)}} = \frac{e^{j\lambda^*}}{\sum_{i=1}^6 e^{i\lambda^*}}$$

Because E does not contain the data generating distribution p, λ^* must be strictly greater than zero (by complementary slackness, q^* lies on the boundary of E; i.e. $\mathbb{E}_{X \sim q} X = 4$). Therefore λ^* is the global minimizer of the dual objective and we can compute it numerically by (unconditional) gradient descent. In particular, $\lambda^* = 0.2519$ and

$$q^* = q_{\lambda^*} \approx (0.1031, 0.1227, 0.1461, 0.1740, 0.2072, 0.2468).$$