COMS-E6998: Information Theory in Computer Science **Fall 2017**

Lecture 4: Mutual Information and KL Divergence

Lecturer: Omri Weinstein *Scribe: Benjamin Kuykendall*

Today we will try to become more comfortable with Mutual Information and KL Divergence. Specifically, we will give an application of KL Divergence to proving Chernoff Bounds, find an expression that relates the two measures to one another, consider a few different viewpoints of KL divergence, introduce statistical distance, and finally prove Pinsker's Inequality.

This material concludes our information theory "toolbox"; next week we will look at interaction. Later, these tools will return in interesting lower bounds in and applications to computer science.

4.1 Review

Before we begin, let us review some basic facts about Mutual Information and KL Divergence. All of these results were proven last lecture, but are repeated here before we make use of them.

Definition 4.1.1. *The **Mutual Information** between two random variables is*

$$I(X;Y) := H(X) - H(X|Y).$$

Fact 4.1.2. *Mutual information is symmetric*

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X).$$

Definition 4.1.3. *The **KL Divergence** between two distributions on a shared universe is*

$$\mathsf{D}\left(\frac{\mu}{\nu}\right) := \mathop{\mathbb{E}}_{x\sim\mu}\left[\lg\frac{\mu(x)}{\nu(x)}\right].$$

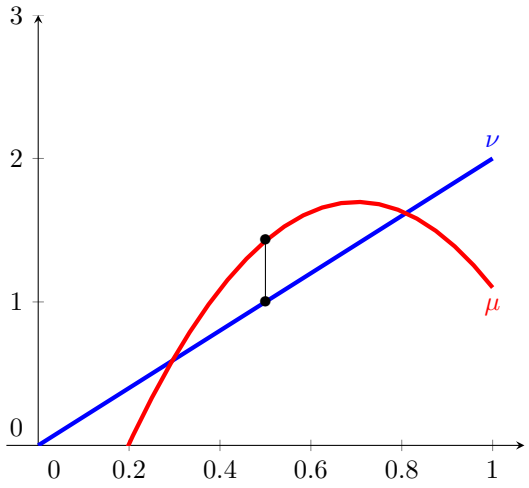

Figure 4.1: $\mathsf{D}(\mu||\nu)$ measures the expected (log of) the *ratio* between the histogram of μ and ν, the two lines illustrated. The point to measure the ratio (i.e., the expectation) is drawn according to μ.

Fact 4.1.4 (Gibb's inequality)**.** $\mathsf{D}(\mu||\nu) \geq 0$.

Fact 4.1.5 (Convexity of KL)**.** *Let* $\mu = \sum_i \alpha_i \mu_i$, $\nu = \sum_i \alpha_i \nu_i$. *Then*

$$\sum_i \alpha_i \mathsf{D}\left(\frac{\mu_i}{\nu_i}\right) \geq \mathsf{D}\left(\frac{\mu}{\nu}\right).$$

Fact 4.1.6 (Chain Rule for KL)**.**

$$\mathsf{D}\left(\frac{X_1, \ldots, X_n}{Y_1, \ldots, Y_n}\right) = \sum_i \mathop{\mathbb{E}}_{x_{<i}} \left[\mathsf{D}\left(\frac{X_i | x_{<i})}{Y_i | x_{<i}}\right)\right].$$

Fact 4.1.7 (Chain rule with independent denominator)**.** *If* $Y_1, \ldots, Y_n$ *are independent, then*

$$\mathsf{D}\left(\frac{X_1, \ldots, X_n}{Y_1, \ldots, Y_n}\right) \geq \sum_i \mathsf{D}\left(\frac{X_i}{Y_i}\right).$$

The next fact relates the two quantities. While MI is a measure on random variables, we can express this measure in terms of the KL divergence between the the following underlying distributions:

Lemma 4.1.8.

$$\mathsf{D}\left(\frac{\mu(x,y)}{\mu(x)\mu(y)}\right) = I(X;Y) = \mathop{\mathbb{E}}_x \left[\mathsf{D}\left(\frac{Y \mid X = x}{Y}\right)\right] = \mathop{\mathbb{E}}_y \left[\mathsf{D}\left(\frac{X \mid Y = y}{X}\right)\right].$$

That is, MI measures how "far" the joint distribution of the possibly correlated random variables x and y is from the product of its marginals. Thus $I(X;Y)$ is a measure of how far X and Y are from independence.

Proof of Lemma 4.1.8. By expressing $\mu(x,y)$ as $\mu(x)\mu(y|x)$ we immediately have

$$\begin{aligned}
\mathsf{D}\left(\frac{\mu(x,y)}{\mu(x)\mu(y)}\right) :&= \mathop{\mathbb{E}}_{x,y}\left[\lg \frac{\mu(x,y)}{\mu(x)\mu(y)}\right] \\
&= \mathop{\mathbb{E}}_{x,y}\left[\lg \frac{\mu(x)\mu(y|x)}{\mu(x)\mu(y)}\right] \\
&= \mathop{\mathbb{E}}_{y}\left[\mathsf{D}\left(\frac{X \mid Y = y}{X}\right)\right].
\end{aligned}$$

Canceling the $\mu(x)$ in the numerator and denominator gives

$$\begin{aligned}
\mathsf{D}\left(\frac{\mu(x,y)}{\mu(x)\mu(y)}\right) &= \mathop{\mathbb{E}}_{x,y}\left[\lg \frac{\mu(y|x)}{\mu(y)}\right] \\
&= \mathop{\mathbb{E}}_{x,y}\left[\lg \frac{1}{\mu(y)}\right] - \mathop{\mathbb{E}}_{x}\left[\mathop{\mathbb{E}}_{y} \lg \frac{1}{\mu(y|x)}\right] \\
&= \mathop{\mathbb{E}}_{y}\left[\lg \frac{1}{\mu(y)}\right] - \mathop{\mathbb{E}}_{x}\left[H(Y|X=x)\right] \\
&= H(Y) - H(Y|X) = I(X;Y).
\end{aligned}$$

□

With these facts in mind, we will now treat ourselves to a nice application.

4.2 Application: An easy proof of the Chernoff bound

The Chernoff bound says that a sum of i.i.d random variables has a mean that deviates from the expectation of the mean with exponentially small probability.

For simplicity, consider $X_1, \ldots, X_n$ to be i.i.d coin tosses with bias p. This bound is often presented as

$$\Pr\left[\frac{1}{n}\sum_{i=1}^{n} X_i \geq (1+\varepsilon)p\right] \leq e^{-\varepsilon^2 pn/2}.$$

The Chernoff bound is typically proved via Markov's inequality on the so-called moment-generating function of the sum. We now present a very simple proof which is a classical example not only of the elegance of information-theoretic language, but also that information theory captures the problem much more intuitively.

We begin with a simple lemma which will also be of use later in the course. It state that conditioning a distribution on a very likely event does not distort the distribution too much, in the sense of KL divergence.

Lemma 4.2.1. *Let X be a r.v and $\mathcal{W}$ be an event in the same probability space. Then*

$$\mathsf{D}\left(\frac{(X|\mathcal{W})}{X}\right) = \mathop{\mathbb{E}}_{p|\mathcal{W}}\left[\lg\frac{\mu(x|\mathcal{W})}{\mu(x)}\right] \leq \lg\frac{1}{\Pr[\mathcal{W}]}.$$

Proof of Lemma 4.2.1.

$$\begin{aligned}\mathsf{D}\left(\frac{(X|\mathcal{W})}{X}\right) &= \sum_x p(x|\mathcal{W})\lg\frac{p(x|\mathcal{W})}{p(x)} \\ &= \sum_x p(x|\mathcal{W})\lg\frac{\Pr(X=x \text{ and } \mathcal{W})}{p(x)\Pr[\mathcal{W}]}\end{aligned}$$

Bounding $\Pr[X = x \text{ and } \mathcal{W}] \leq p(x)$

$$\begin{aligned}&= \sum_x p(x|\mathcal{W})\lg\frac{p(x)}{p(x)\Pr[\mathcal{W}]} \\ &= \lg\frac{1}{p(w)}\sum_x p(x|\mathcal{W})\end{aligned}$$

Bounding $\sum_x p(x|\mathcal{W}) \leq 1$

$$\leq \log\frac{1}{\Pr[\mathcal{W}]}.$$ □

Proof of Chernoff bound. For simplicity set $p = 1/2$. The event of interest is $\mathcal{W} := \sum_{i=1}^{n} X_i \geq (\varepsilon + 1/2)n$. Apply the lemma to bound $\Pr[\mathcal{W}]$:

$$\mathsf{D}\left(\frac{(X_1,\ldots,X_n|\mathcal{W})}{X_1,\ldots,X_n}\right) \leq \lg\frac{1}{\Pr[\mathcal{W}]}.$$

On the other hand, since the X_i's are independent, by Fact 4.1.7 (with $Z_1, \ldots, Z_n := (X_1, \ldots, X_n|\mathcal{W})$)

$$\mathsf{D}\left(\frac{X,\ldots,X_n|\mathcal{W}}{X,\ldots,X_n}\right) \geq \sum_i \mathsf{D}\left(\frac{X_i|\mathcal{W}}{X_i}\right)$$

Now, even though the coin flips are no longer independent when conditioned on $\mathcal{W}$, marginally the $X_i|\mathcal{W}$ are distributed as Bernoulli random variables with some bias at least $q = (\varepsilon + 1/2)$. Then as for fixed p the value of $\mathsf{D}(q||p)$ is monotonically increasing in $|q - p|$ we can bound

$$\sum_i \mathsf{D}\left(\frac{X_i|\mathcal{W}}{X_i}\right) \geq n\mathsf{D}\left(\frac{\varepsilon + 1/2}{1/2}\right)$$

We discussed before that $\mathsf{D}(\varepsilon + 1/2||1/2)$ is close to $\Theta(1)\varepsilon^2$. To get the quantitative bound we need, use Pinsker's inequality (we will prove the inequality at the end of today's lecture, but with a worse constant).

$$\mathsf{D}\left(\frac{\varepsilon + 1/2}{1/2}\right) \geq \frac{\varepsilon^2}{4}$$

Chaining together the inequalities yields

$$\lg \frac{1}{\Pr[\mathcal{W}]} \geq \frac{\varepsilon^2}{4}$$

or

$$\Pr[\mathcal{W}] \leq e^{-\varepsilon^2(1/2)/2}$$ □

If we stop one step sooner, we get a bound in terms of Bernoulli KL divergence

$$\Pr\left[\frac{1}{n}\sum_{i=1}^{n} X_i \geq (1+\varepsilon)p\right] \leq e^{-n\mathsf{D}(p+\varepsilon||p)}.$$

This strengthened form of the Chernoff bound is known to be tight up to polynomial factors.

4.2.1 Another exercise

Before moving on, we give an exercise to practice transformations between MI and KL.

Example 4.2.2. *Let S be a uniform string $S \in_R \Sigma^n$, and J a random position $J \in_R [n]$. For $A = S[J]$ show that $I(S; A, J) = \lg \Sigma$.*

This exercise follows easily by considering $H(S) - H(S|A, J)$, but try repeating it by analyzing $\mathbb{E}_{A,J}[\mathsf{D}(S|A, J||S)]$.

4.3 Some Viewpoints on KL

We now give three different *operational* interpretations of KL divergence. These interpretations will be very useful to the applications we are going to see in the next month or so. In lecture, we focused more on intuition building than proof for these topics; proofs or pointers to proofs are included here.

4.3.1 KL as source coding loss

Consider two different Shannon-Fano codes over $[n]$: a code C_q designed for the distribution $q = (q_1, \ldots, q_n)$ and a code C_p for $p = (p_1, \ldots, p_n)$. We know that for elements distributed according to p we have the optimal expected length $H(p) \leq \mathbb{E}_{i\sim p}|C_p(i)| \leq H(p) + 1$, but what how much extra would we pay if we used the other code instead?

$$\begin{aligned}
\mathop{\mathbb{E}}_{i\sim p} |C_q(i)| &= \sum_i p_i \left\lceil \lg \frac{1}{q_i} \right\rceil \\
&\geq \sum_i p_i \lg \frac{1}{q_i} \\
&= \sum_i p_i \lg \frac{p_i}{p_i q_i} \\
&= \sum_i p_i \lg \frac{1}{p_i} + \sum_i p_i \lg \frac{p_i}{q_i} \\
&= H(p) + \mathsf{D}(p||q).
\end{aligned}$$

A similar calculation upper bounds L by $\leq H(p) + \mathsf{D}(p||q) + 1$. So up to $|c| \leq 2$ bits

$$\mathop{\mathbb{E}}_{i\sim p} |C_q(i)| - \mathop{\mathbb{E}}_{i\sim p} |C_p(i)| = \mathsf{D}(p||q) + c.$$

Because of this $\mathsf{D}(p||q)$ can be interpreted as precisely the expected asymptotic penalty paid when encoding an independent sequence of elements from p by using a code designed for q instead.

4.3.2 KL interpretation for rejection sampling

Let μ and ν be two distributions on U.

We would like to rigorously define the probability that a sequence of elements drawn from μ will "look like" it came from ν. In our proof of the Chernoff bound, we showed that in a very specific sense of "looking alike" (when $\mu = B(p)$ has a sample mean that is at least the mean of $\nu = B(q)$) this probability is $e^{-n\mathsf{D}(\mu||\nu)}$.

Another more natural definition of "looking alike" comes from rejection sampling. We have an oracle that outputs random samples according to ν, but we would like to use it to produce a random samples from μ. More formally we have:

Input A sequence $x_1, x_2, \ldots$ of i.i.d samples drawn according to ν.

Goal Output an index i^* such that x_{i^*} is distributed according to μ.

Again, the KL divergence almost exactly characterizes how well we can do.

Theorem 4.3.1. *There is a strategy for outputting i^* such that $\mathbb{E}[\lg i^*] \lesssim \mathsf{D}(\mu||\nu)$.*

This general result is stated without proof, though later in the course we will come back to it (when we need to use rejection sampling for compressing interactive protocols). An alternative operational property that hints at this same concept is more graphical:

Fact 4.3.2 (Informal)**.** *Let $K := \mathsf{D}(\mu||\nu)$. If one "rescales" the histogram of ν by a multiplicative factor of 2^K, then the histogram of μ is "almost entirely" lies under the histogram of $2^K\nu$. Or similarly: consider a random dart (x, y) uniformly on the rectangle $U \times [0, \max \nu(x)]$. Then*

$$\Pr[y < \mu(x) \mid y < \nu(x)] \approx 2^{-\mathsf{D}(\mu||\nu)}.$$

4.3.3 KL vs. Statistical Distance, and more Rejection Sampling

The standard distance measure which is often used in computer science applications is the *statistical distance* (a.k.a total-variation distance). In the discrete case, it is just the ℓ_1 norm of the difference between μ and ν when viewed as probability vectors.

Definition 4.3.3 (Statistical Distance)**.** *Two distributions μ and ν over U have* ***statistical distance***

$$\|\mu - \nu\|_1 := \frac{1}{2}\sum_{i \in U} |\mu(i) - \nu(i)|.$$

Here is another interpretation. Let A be the event that some i occurs such that $\mu(i) - \nu(i) > 0$. Then

$$\|\mu - \nu\|_1 = \frac{1}{2}\sum_{i \in A}(\mu(i) - \nu(i)) - \frac{1}{2}\sum_{i \notin A}(\mu(i) - \nu(i)),$$

and as $\sum_i \mu(i) = \sum_i \nu(i) = 1$ we have

$$= \frac{1}{2}\sum_{i \in A}(\mu(i) - \nu(i)) - \frac{1}{2}\sum_{i \notin A}(\mu(i) - \nu(i)) + \frac{1}{2}\sum_{i}(\mu(i) - \nu(i)) = \sum_{i \in A}(\mu(i) - \nu(i)).$$

In fact, this choice of A maximizes $\mu(A) - \nu(A)$ so we give the alternate characterization

$$\|\mu - \nu\|_1 = \max_{A \subseteq U}(\mu(A) - \nu(A)).$$

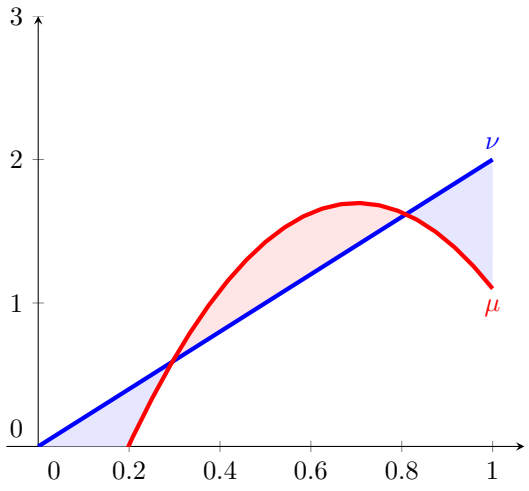

Figure 4.2: The shaded region is $\|\mu - \nu\|_1$.
The red corresponds to $\mu > \nu$, so it is the sum of $|\mu(i) - \nu(i)|$ under event A.
The blue corresponds to $\mu < \nu$ and the sum over event $\overline{A}$.

The reason that ℓ_1 distance is so natural in computer science applications is its algorithmic interpretation:

Claim 4.3.4. *If $\|\mu-\nu\|_1 \le \delta$, then there is an algorithm that, given a sample from either μ or ν, distinguishes correctly with probability $\ge 1-\delta$. More formally,*

$$\exists \text{ algorithm } A : [n] \mapsto \{0,1\} \quad \text{s.t} \quad \left| \mathop{\mathbb{E}}_{x\sim\mu}[A(x)] - \mathop{\mathbb{E}}_{x\sim\nu}[A(x)] \right| \ge \delta \iff \|\mu-\nu\|_1 = \Theta(\delta).$$

The algorithm is just the indicator of the event A we defined earlier. This can be viewed as a special case of the following rejection sampling game:

Problem 4.3.5 (Correlated sampling)**.** *Give Alice and Bob probability vectors μ and ν over a shared universe U. Allows them to access a string of "public randomness" but allow no communication. Alice must output an x distributed $x \sim \mu$. Bob can output any y. What value of $\Pr[x=y]$ can they achieve?*

For the binary case $\mu = B(p)$ and $\nu = B(q)$ we have the following protocol: interpret the public randomness as $x \in [0,1]$. Alice outputs $\mathbb{1}[x<p]$ and Bob outputs $\mathbb{1}[x<q]$. They fail with probability $\|B(p)-B(q)\|_1$.

Because of problems like this, we often want to use statistical distance in applications. However, sometimes analyzing this measure (or design algorithms that achieve it directly) is difficult (for example, since it does not have a chain rule), and we will see that often the "right" distance measure to analyze, especially when it comes to communication problems, is actually KL. Luckily, the following important inequality asserts that if two distributions are close in KL distance, then they are also close in statistical distance.

However, we know the two quantities behave a little differently. Take Bernoulli distributions $B(1/2)$ and $B(1/2+\epsilon)$. We know that

$$\mathsf{D}\left(\frac{1/2+\varepsilon}{\varepsilon}\right) = \Theta(\varepsilon^2)$$

but

$$\Big\| B(1/2+\varepsilon) - B(1/2) \Big\|_1 = |1/2+\varepsilon-1/2| = \varepsilon.$$

So the KL goes with the square of the distance. A version of this relationship exists in general.

Theorem 4.3.6 (Pinsker's Inequality)**.**

$$\mathsf{D}\left(\frac{\mu}{\nu}\right) \ge \frac{1}{2}\|\mu-\nu\|_1^2.$$

We will prove a weakened version

$$\mathsf{D}\left(\frac{\mu}{\nu}\right) \ge \frac{1}{2\ln 2}\|\mu-\nu\|_1^2.$$

Proof of Pinsker's Inequality for binary case. As usual, we first prove the inequality for binary random variable, and then show how to easily generalize it to the general one. For $B(p)$ and $B(q)$ random variables, we can give write explicity the KL and distance.

$$f(p,q) := \mathsf{D}\left(\frac{p}{q}\right) - \frac{1}{2\ln 2}\|\mu-\nu\|_1^2 = p\lg\left(\frac{p}{q}\right) + (1-p)\lg\left(\frac{1-p}{1-q}\right) - \frac{1}{2\ln 2}(2(p-q)^2).$$

Then it remains to show that $f(p,q) \ge 0$ for all $p,q \in [0,1]$.

Assume without loss of generality $p \ge q$ (the other case is symmetric).

This is an excercise in calculus. One can verify that the function is monotone decreasing in q since

$$\frac{\partial f}{\partial q} = -\frac{p-q}{\ln 2}\left(\frac{1}{q(1-q)} - 4\right) \leq 0.$$

Since $f = 0$ when $q = p$, it follows that $f(p, q) \leq 0$ whenever $p \geq q$. □

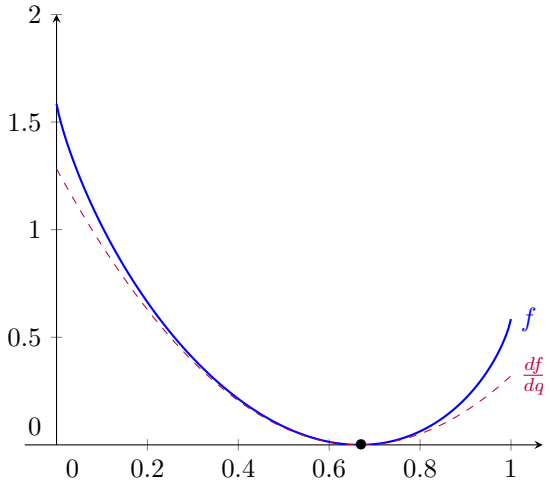

Figure 4.3: A plot of $f = \mathsf{D}(p||q)$ fixing $p = 2/3$ and varying $q \in [0, 1]$ and its derivative. This shows the minimum which occurs when $q = p = 2/3$.

Proof of Pinsker's Inequality. To generalize to arbitrary random variables μ and ν, define

$$A := \{x \mid \mu(x) > \nu(x)\}.$$

Consider the indicator random variables for A, under μ and ν. It has a distribution we call the *projection* of μ or ν to A, which is a binary distribution $\mu_A = B(\mu(A))$ or $\nu_A = B(\nu(A))$.

Compute the statistical distance

$$\begin{aligned}
\|\mu_A - \nu_A\|_1 :&= |\sum_{x\in A}\mu(x) - \sum_{x\in A}\nu(x)| - |\sum_{x\notin A}\mu(x) - \sum_{x\notin A}\nu(x)| \\
&= \sum_{x\in A}\mu(x) - \sum_{x\in A}\nu(x) + \sum_{x\notin A}\nu(x) - \sum_{x\notin A}\mu(x) && \text{(by definition of } A\text{)} \\
&= \sum_{x\in A}|\mu(x) - \nu(x)| + \sum_{x\notin A}|\mu(x) - \nu(x)| := \|\mu - \nu\|_1.
\end{aligned}$$

On the other hand, by the chain rule (and non-negativity of KL) it follows that

$$\mathsf{D}\left(\frac{\mu}{\nu}\right) \geq \mathsf{D}\left(\frac{\mu_T}{\nu_T}\right)$$

Hence we conclude by the inequality for the binary case

$$\mathsf{D}\left(\frac{\mu}{\nu}\right) \geq \mathsf{D}\left(\frac{\mu_T}{\nu_T}\right) \geq \frac{1}{2\ln 2}\|\mu_T - \nu_T\|_1^2 = \frac{1}{2\ln 2}\|\mu - \nu\|_1^2.$$

□

References