ECE 598 - Fall 2025

Lecture 3: Properties of Information Measures

September 02, 2025

Instructor: Shubhanshu Shekhar

In this lecture, we continue our discussion of the properties of information measures. We begin by establishing the convexity/concavity properties of the three main information measures, and then state and prove Fano's inequality that will often be used in proving 'converse results' (or impossibility results). We end the lecture by discussing an important (functional) variational definition of relative entropy, called the Donsker-Varadhan representation.

1 Convexity/Concavity of Information Measures

For simplicity, we will focus on the case of discrete distributions. As we discussed in the last lecture, many results for relative entropy and mutual information can be generalized to arbitrary distributions via the Gelfand-Yaglom-Peres variational representation.

Theorem 1.1. *Suppose $\mathcal{X}$ and $\mathcal{Y}$ denote two finite alphabets of sizes m and n. Then, the following statements are true:*

1. *Suppose P_1, P_2, Q_1 and Q_2 distributions on the finite alphabet $\mathcal{X}$. For any $\lambda \in [0,1]$, define $P^\lambda = \lambda P_1 + \bar{\lambda} P_2$ and $Q^\lambda = \lambda Q_1 + \bar{\lambda} Q_2$ where $\bar{\lambda} = 1 - \lambda$. Then, we have*

$$D_{KL}(P^\lambda \,\|\, Q^\lambda) \leq \lambda D_{KL}(P_1 \,\|\, Q_1) + \bar{\lambda} D_{KL}(P_2 \,\|\, Q_2).$$

   That is D_{KL} is jointly convex in its arguments. Setting $P_1 = P_2$ or $Q_1 = Q_2$, this also implies that D_{KL} is convex in each of its arguments (with the other kept fixed).

2. *Suppose $X \sim P_X$ is a $\mathcal{X}$ valued random variable. Then, $H(P_X)$ is a concave function of P_X.*

3. *Suppose P_{XY} is a joint distribution on $\mathcal{X} \times \mathcal{Y}$. Then, $I(X;Y) \equiv I(P_{XY})$ is a convex function of $P_{Y|X}$ for a fixed input distribution P_X, and it is a concave function of P_X for a fixed channel $P_{Y|X}$.*

1.1 Proof of Theorem 1.1

A standard proof of the convexity of relative entropy is via an intermediate "log-sum-inequality". See Chapter 2 of Cover and Thomas for the details. Here we present a minor generalization of that approach using the notion of "perspective of a function" (Boyd and Vandenberghe, 2004, Sec. 3.2.6). The benefit is that the same argument will work for a more general class of information measures called f-divergences.

Lemma 1.2. *Let $\varphi : [0, \infty) \to [0, \infty]$ be a convex function. Then, the perspective of φ, denoted by $\varphi^\pi : [0, \infty) \times (0, \infty) \to [0, \infty]$ is defined as*

$$\varphi^\pi(z, t) = t\varphi(z/t)$$

is also convex. That is, the perspective operation preserves convexity.

Proof. The result follows directly from the convexity of φ. Let (z_1, t_1) and (z_2, t_2) denote two points in the domain of φ^π. For any $\lambda \in [0, 1]$, let (z_λ, t_λ) denote their convex combination. Then, we have

$$\begin{aligned}
\varphi^\pi(z_\lambda, t_\lambda) &= t_\lambda \varphi\left(\frac{\lambda z_1}{t_\lambda} + \frac{\bar{\lambda} z_2}{t_\lambda}\right) = t_\lambda \varphi\left(\frac{\lambda t_1}{t_\lambda}\frac{z_1}{t_1} + \frac{\bar{\lambda} t_2}{t_\lambda}\frac{z_2}{t_2}\right) \\
&\leq t_\lambda \left(\frac{\lambda t_1}{t_\lambda}\varphi(z_1/t_1) + \frac{\bar{\lambda} t_2}{t_\lambda}\varphi(z_2/t_2)\right) && \text{(convexity of } \varphi(\cdot)) \\
&= \lambda\varphi^\pi(z_1, t_1) + \bar{\lambda}\varphi^\pi(z_2, t_2) && \text{(definiton of } \varphi^\pi).
\end{aligned}$$

□

Convexity of Relative Entropy. Consider the following instance of the above result: $\varphi(u) = u \log u$, $(z_1, t_1) = (p_1(x), q_1(x))$, $(z_2, t_2) = (p_2(x), q_2(x))$, and $(z_\lambda, t_\lambda) = (p_\lambda(x), q_\lambda(x))$. Then, by Lemma 1.2, we know that φ^π is convex, which implies

$$\begin{aligned}
p_\lambda(x) \log\left(\frac{p_\lambda(x)}{q_\lambda(x)}\right) &= \varphi^\pi(z_\lambda, t_\lambda) \leq \lambda\varphi^\pi(z_1, t_1) + \bar{\lambda}\varphi^\pi(z_2, t_2) \\
&= \lambda p_1(x) \log\left(\frac{p_1(x)}{q_1(x)}\right) + \bar{\lambda} p_2(x) \log\left(\frac{p_2(x)}{q_2(x)}\right).
\end{aligned}$$

Summing this inequality over all $x \in \mathcal{X}$ establishes the convexity of relative entropy.

Remark 1.3. As we will see in the next lecture, the same argument directly extends to a larger class of divergences called the f-divergence family, which includes measures like total variation, chi-squared, Hellinger metric, Jensen-Shannon divergence, etc.

Concavity of Entropy. This is a direct consequence of the convexity of relative entropy. In particular, note that the entropy of $X \sim P_X$ can be writen as

$$\begin{aligned}
H(P_X) = H(X) &= \sum_{x \in \mathcal{X}} p_X(x) \log\left(\frac{1}{p_X(x)}\right) = -\sum_{x \in \mathcal{X}} p_X(x) \log\left(\frac{p_X(x)}{1/|\mathcal{X}|}\right) + \log(|\mathcal{X}|) \\
&= \log(|\mathcal{X}|) - D_{\mathrm{KL}}(P_X \parallel P_U),
\end{aligned}$$

where P_U denotes the uniform distribution over the finite alphabet $\mathcal{X}$. Since we have already proved that relative entropy is convex in its second argument, this implies the concavity of entropy.

Convexity/Concavity of Mutual Information. Let us first consider the case of a fixed marginal P_X. Then, we have

$$I(X;Y) = D_{\mathrm{KL}}(P_{Y|X} \parallel P_Y|P_X) = D_{\mathrm{KL}}(P_{Y|X} \parallel \sum_x P_X(x)P_{Y|X=x} \mid P_X).$$

Since P_Y is a linear function of $P_{Y|X}$ and relative entropy is convex in both of its arguments, we can conclude that the above expression is convex in $P_{Y|X}$.

Next, we consider the case of a fixed $P_{Y|X}$. In this case, it is cleaner to work with the entropy definition:

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

Observe that $H(Y)$ is concave in P_Y, P_Y is a linear function of P_X, and $H(Y|X)$ is a linear function of P_X. Hence, $I(X;Y)$ is the sum of two terms: the first is concave $\circ$ linear in P_X and the second is linear in P_X. Together it implies that $I(X;Y)$ is concave in P_X, with $P_{Y|X}$ fixed.

2 Fano's inequality

Consider the following estimation problem: Let $W \in \mathcal{W}$ denote some unknown parameter (or message to be communicated), and let $Y \in \mathcal{Y}$ denote the observations with conditional distribution (or channel) $P_{Y|X}$. Suppose we construct an estimate of W, denoted by $\hat{W} \in \mathcal{W}$ based only on the observation Y. Let us denote the (possibly randomized) estimation procedure by $P_{\hat{W}|Y}$. We are interested in understanding how small the probability of error, $p_e = \mathbb{P}(\hat{W} \neq W)$ can be? Clearly, this answer should depend on how "informative" the observations Y are about the unknown parameter W. One way to formalize this is via Fano's inequality.

Theorem 2.1. *Given the Markov chain $W \to Y \to \hat{W}$, with $W, \hat{W} \in \mathcal{W}$ and $Y \in \mathcal{Y}$ for some finite alphabets $\mathcal{W}, \mathcal{Y}$, let p_e denote the probability $\mathbb{P}(\hat{W} \neq W)$. Then, we have the following relation*

$$h_b(p_e) + p_e \log(|\mathcal{W}| - 1) \geq H(W|\hat{W}) \geq H(W|Y),$$

where $h_b(p) = -p \log p - \bar{p} \log(\bar{p})$ is the binary entropy function.

Proof. The classical proof of this result is by introducing an error indicator $E = \mathbf{1}_{\hat{W} \neq W}$, and observing that $E \sim \mathrm{Bernoulli}(p_e)$ by definition. Then, observe the following:

$$H(W, E|\hat{W}) = H(E|\hat{W}) + H(W|E, \hat{W}) = H(W|\hat{W}) + H(E|\hat{W}, W).$$

The blue term is easy to analyze: since $E = \mathbf{1}_{W \neq \hat{W}}$ (i.e., E is a deterministic function of $W, \hat{W}$), the conditional entropy $H(E|W, \hat{W}) = 0$. Thus, we have

$$H(W|\hat{W}) + H(E|\hat{W}, W) = H(W|\hat{W}). \tag{1}$$

For the red term, we can get an upper bound as follows:

$$\begin{aligned} H(E|\hat{W}) + H(W|E,\hat{W}) &\leq H(E) + H(W|E,\hat{W}) \\ &= h_b(p_e) + \mathbb{P}(E=1)H(W|E=1,\hat{W}) + \mathbb{P}(E=0)H(W|E=0,\hat{W}) \\ &= h_b(p_e) + p_e \underbrace{H(W|\hat{W},E=1)}_{\leq \log(|\mathcal{W}|-1)} + (1-p_e)\underbrace{H(W|E=0,\hat{W})}_{=0} \end{aligned} \tag{2}$$

The first inequality is due to the "conditioning reduces entropy" fact. In the last line, we use the fact that given $E = 1$ (error) and $\hat{W}$, the random variable W can take one of $|\mathcal{W}| - 1$ values. Hence, the conditional entropy $H(W|\hat{W}, E = 1)$ is upper bounded by $\log(|\mathcal{W}| - 1)$. Finally, given that $E = 0$ (no error) and $\hat{W}$, there is no uncertainty about W; hence $H(W|E = 0, \hat{W}) = 0$. Combining (1) and (2) we get the required

$$H(W|Y) \leq H(W|\hat{W}) \leq h_b(p_e) + p_e \log(|\mathcal{W}| - 1).$$

The first inequality is due to the DPI for entropy. This completes the proof. See also (Polyanskiy and Wu, 2025, Section 3.6) for an alternative proof technique. □

Interpretation. To interpret Fano's inequality, let us simplify the expression with $h_b(p_e) \leq 1$, and $\log(|\mathcal{W}| - 1) \leq \log(|\mathcal{W}|)$, to get

$$\inf_{P_{\hat{W}|W}} \mathbb{P}(\hat{W} \neq W) \geq \frac{H(W|Y) - 1}{\log(|\mathcal{W}|)}.$$

In other words, there is a fundamental lower bound on the error that can be achieved by any estimator that is characterized by $H(W|Y)$: the uncertainty about W given the observations Y. If there is a large amount of uncertainty about the parameter W after knowing Y, it is impossible to drive the probability of error below the fundamental limit implied by RHS above.

To simplify further, assume that $W \sim \text{Uniform}(\mathcal{W})$, or $H(W) = \log(|\mathcal{W}|)$, we get

$$\inf_{P_{\hat{W}|W}} \mathbb{P}(\hat{W} \neq W) \geq \frac{H(W) - (H(W) - H(W|Y)) - 1}{\log(|\mathcal{W}|)} = 1 - \frac{I(W;Y) + 1}{\log(|\mathcal{W}|)}.$$

In other words, the minimum probability of error achievable depends inversely on the mutual information between the unknown parameter W and the observation Y.

3 Application: MAP Estimation and Reverse Fano's

Let us return to the estimation problem represented with the following Markov Chain:

$$\underbrace{W}_{\text{Unknown Parameter}} \rightarrow \underbrace{Y}_{\text{Observation}} \rightarrow \underbrace{\hat{W} \sim P_{\hat{W}|Y}}_{\text{Estimate}}.$$

Our goal is to find an estimator $P_{\hat{W}|Y}$ which minimizes the probability of error.

$$P^*_{\hat{W}|Y} = \underset{P_{\hat{W}|Y}}{\text{argmin}}\, \mathbb{P}\left(\hat{W} \neq W\right).$$

Proposition 3.1. *The optimal estimator $P^*_{\hat{W}|Y}$ is non-randomized (i.e., represented by a map $g^* : \mathcal{Y} \to \mathcal{W}$), and is defined as*

$$g^*(y) = \operatorname*{argmax}_{w \in \mathcal{W}} p_{W|Y}(w|y), \quad \textit{with} \quad \mathbb{P}(g^*(Y) \neq W) = \sum_{y \in \mathcal{Y}} p_Y(y)(1 - \max_{w \in \mathcal{W}} p_{W|Y}(w|y)).$$

This is also called the Maximum Aposteriori Probability (MAP) estimator.

Proof. We can prove this by using the Markov property to state the error in a suitable form. For any estimator $P_{\hat{W}|Y}$, we have

$$\begin{aligned}
\mathbb{P}(\hat{W} \neq W) &= \sum_{w,y,\hat{w}} P_{WY\hat{W}}(w, y, \hat{w}) \mathbf{1}_{w \neq \hat{w}} \\
&= \sum_{w,y,\hat{w}} p_Y(y) p_{W|Y}(w|y) p_{\hat{W}|Y}(\hat{w}|y) && \text{(since } W \perp \hat{W}|Y) \\
&= \sum_y p_Y(y) \sum_{\hat{w}} p_{\hat{w}|Y}(\hat{w}|y) \sum_{w \neq \hat{w}} p_{W|Y}(w|y).
\end{aligned}$$

Since $\sum_{w \neq \hat{w}} p_{W|Y}(w|y) = (1 - p_{W|Y}(\hat{w}|y))$, we get

$$\mathbb{P}(\hat{W} \neq W) = \sum_y p_Y(y) \sum_{\hat{w}} p_{\hat{w}|Y}(\hat{w}|y)(1 - p_{W|Y}(\hat{w}|y))$$

Let $w_y = \operatorname{argmax}_{w \in \mathcal{W}} p_{W|Y}(w|y)$ (breaking ties arbitrarily), and observe that for any conditional pmf $p_{\hat{W}|Y}(\cdot|y)$, we have

$$\sum_{\hat{w}} p_{\hat{w}|Y}(\hat{w}|y)(1 - p_{W|Y}(\hat{w}|y)) \geq 1 - p_{W|Y}(w_y|y),$$

with equality if $p_{\hat{W}|Y}(w_y|y) = 1$. Thus, $g^*(y) = w_y \in \operatorname{argmax}_{w \in \mathcal{W}} p_{W|Y}(w|y)$ is an estimator that achieves the minimum probability of error. □

The intuition behind the MAP estimator is simple: given the observation $Y = y$, we set our estimator value to w_y; the parameter value that is most likely to be true given $Y = y$. We can now use this estimator to state a "reverse Fano's' inequality: *just as Fano's inequality lower bounds the minimum probability of error in terms of $H(W|Y)$, reverse Fano's gives us an upper bound on $p^*_e = \mathbb{P}(g^*(Y) \neq W)$ in terms of a function involving $H(W|Y)$. In particular, it tells us that p^*_e cannot be too large if $H(W|Y)$ is small.*.

Proposition 3.2. *Let g^* denote the MAP estimator and let p^*_e denotes its probability of error. Then, we have*

$$p^*_e \leq 1 - 2^{-H(W|Y)}.$$

Proof. Note that we need to show that $H(W|Y) \geq \log(1/(1-p_e^*))$, where

$$p_e^* = \sum_y p_Y(y)(1 - \max_w p_{W|Y}(w|y)) = \sum_y p_Y(y)(1 - p(w_y|y)), \tag{3}$$

where recall that $w_y \in \operatorname{argmax}_{w \in \mathcal{W}} p_{W|Y}(w|y)$. Let us now look at the conditional entropy of W given $Y = y$:

$$\begin{aligned} H(W|Y=y) &= \sum_{w \in \mathcal{W}} p_{W|Y}(w|y) \log\left(1/p_{W|Y}(w|y)\right) \\ &\geq \sum_w p_{W|Y}(w|y) \log\left(\frac{1}{\max_w p_{W|Y}(w|y)}\right) = \log\left(\frac{1}{p_{W|Y}(w_y|y)}\right) \end{aligned}$$

On simplifying this gives us $p_{W|Y}(w_y|y) \geq 2^{-H(W|Y=y)}$, or $1 - p_{W|Y}(w_y|y) \leq 1 - 2^{-H(W|Y=y)}$. Using this inequality with (3), we get

$$\begin{aligned} p_e^* &\leq \sum_{y \in \mathcal{Y}} p_Y(y)\left(1 - 2^{-H(W|Y=y)}\right) = 1 - \sum_y p_Y(y) 2^{-H(W|Y=y)} \\ &\overset{\text{Jensen's}}{\leq} 1 - 2^{-\sum_y p_Y(y) H(W|Y=y)} = 1 - 2^{-H(W|Y)}. \end{aligned}$$

This completes the proof. □

4 Donsker-Varadhan Variational Representation

A standard fact in convex analysis says that any convex function can be written as the supremum of a collection of linear or affine functions. We have already proved that the mapping $P \mapsto D_{\text{KL}}(P \parallel Q)$, for a fixed Q, is convex, and our next result discusses a representation of $D_{\text{KL}}(P \parallel Q)$ as a supremum of a particular class of affine functions of P.

Theorem 4.1 (Donsker-Varadhan Representation)**.** *Suppose P, Q are two distributions on an alphabet $\mathcal{X}$ (not necessarily finite), and let $\mathcal{C}_Q = \{f : \mathcal{X} \to \mathbb{R} \cup \{-\infty\} : \mathbb{E}_{X \sim Q}[e^f(X)] < \infty\}$. Then, we have the following:*

$$D_{KL}(P \parallel Q) = \sup_{f \in \mathcal{C}_Q} \mathbb{E}_P[f(X)] - \log \mathbb{E}_Q[e^f(X)].$$

Observe that for a fixed (f, Q), the term inside the supremum above is an affine function of P.

We will discuss the general idea of the proof, while not accounting for certain technicalities. For a more rigorous proof, see (Polyanskiy and Wu, 2025, Section 4.3).

Proof. For simplicity, we will assume that $D_{\text{KL}}(P \parallel Q) < \infty$, and P, Q have densities p, q with respect to some dominating measure μ. Let us define the "tilted measure" associated with Q, and denoted by Q_f, as a distribution with density $q_f(x) = q(x)e^{f(x) - \psi_{Q,f}}$, where $\psi_{Q,f} = \log E_Q[e^f(X)]$.

Then, observe that $\log(q_f/q) = f - \psi_{Q,f}$, which implies the following:

$$\begin{aligned}\mathbb{E}_P[f(X) - \psi_{Q,f}] &= \mathbb{E}_P\left[\log\left(\frac{q_f}{q}\right)\right] = \mathbb{E}_P\left[\log\left(\frac{p}{q}\right) - \log\left(\frac{p}{q_f}\right)\right] \\ &= D_{\mathrm{KL}}(P \parallel Q) - D_{\mathrm{KL}}(P \parallel Q_f) \leq D_{\mathrm{KL}}(P \parallel Q).\end{aligned}$$

Thus, we have proved that for any $f \in \mathcal{C}_Q$, we have $\mathbb{E}_P[f(X) - \psi_{Q,f}] \leq D_{\mathrm{KL}}(P \parallel Q)$, which implies that

$$\sup_{f \in \mathcal{C}_Q} \mathbb{E}_P[f(X)] - \log \mathbb{E}_Q\left[e^{f(X)}\right] \leq D_{\mathrm{KL}}(P \parallel Q). \tag{4}$$

This shows one direction of the required equality. For the other direction, condider the function $f = \log(p/q)$, and observe that $f \in \mathcal{C}_Q$, and

$$\mathbb{E}_P[f(X)] - \log E_Q[e^{f(X)}] = \mathbb{E}_P[\log(p/q)] - \log E_Q[p/q] = D_{\mathrm{KL}}(P \parallel Q) - \log 1 = D_{\mathrm{KL}}(P \parallel Q).$$

Hence, on taking supremum over all f, we get the other direction

$$\sup_{f \in \mathcal{C}_Q} \mathbb{E}_P[f(X)] - \log \mathbb{E}_Q[e^{f(X)}] \geq D_{\mathrm{KL}}(P \parallel Q). \tag{5}$$

Combining (4) and (5), we get the required equality. □

This result finds a large number of applications in probability, statistics, and machine learning. We discuss one such application next.

4.1 Application: Estimation of Relative Entropy and Density Ratio

Problem Statement. Let $X^n = (X_1, \ldots, X_n) \overset{i.i.d.}{\sim} P_X$ with density p, and $Y^m = (Y_1, \ldots, Y_m) \overset{i.i.d.}{\sim} Q_Y$ with density q (both w.r.t. the Lebesgue measure on $\mathbb{R}^d$) denote two sets of data points lying in $\mathcal{X} = \mathbb{R}^d$. For example, X^n could represent n natural images, and Y^m could be a set of images generated by an AI model. Our goal is to use this data to (i) estimate the relative entropy between P_X and Q_Y, and (ii) estimate the density-ratio function $\ell(\cdot) = p(\cdot)/q(\cdot)$.

Some naive approaches. A natural idea might be to construct the empirical distributions, $\widehat{P}_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ and $\widehat{Q}_m = \frac{1}{m}\sum_{j=1}^m \delta_{Y_j}$, and plug them into the standard expression of relative entropy. However, this approach will almost always fail as $\widehat{P}_n \not\ll \widehat{Q}_m$ for most data realizations. A more refined approach is to construct "smoothed" density estimators, and use them to evaluate the relative entropy. However, this method suffers from needing to evaluate a high-dimensional integral.

DV-based solution. The DV representation allows us to translate the estimation problem into an optimization problem, which can then be practically solved using gradient based solvers. Formally, let $\mathcal{F}_\Theta = \{f_\theta : \theta \in \Theta\}$ denote a class of functions parameterized by Θ (for example, $\mathcal{F}_\Theta$ could represent all neural networks with a particular architecture). Then, we know from the DV representation of relative entropy that

$$D_{\mathrm{KL}}(P_X \parallel Q_Y) \geq \sup_{f_\theta \in \mathcal{F}_\Theta} \mathbb{E}_{P_X}[f_\theta(X)] - \log\left(\mathbb{E}_{Q_Y}[e^{f_\theta(Y)}]\right).$$

Replacing the expectations with their empirical counterparts, we get the estimate for relative entropy

$$\widehat{D}_{n,m} = \max_{f_\theta \in \mathcal{F}_\Theta} \frac{1}{n} \sum_{i=1}^{n} f_\theta(X_i) - \frac{1}{m} \sum_{j=1}^{m} e^{f_\theta(Y_j)},$$

as well as for the density ratio

$$\widehat{\ell}_{n,m} = \operatorname*{argmax}_{f_\theta \in \mathcal{F}_\Theta} \frac{1}{n} \sum_{i=1}^{n} f_\theta(X_i) - \frac{1}{m} \sum_{j=1}^{m} e^{f_\theta(Y_j)}.$$

From a theoretical perspective, if we assume that the true density ratio p/q lies in the function class $\mathcal{F}_\Theta$ (or can be well approximated by it), and under certain capacity assumptions on $\mathcal{F}_\Theta$, we can show that $|\widehat{D}_{n,m} - D_{\mathrm{KL}}(P_X \parallel Q_Y)| \overset{n,m\to\infty}{\longrightarrow} 0$. In practice, by using a sufficiently large neural netowrk, we can ensure that the condition $p/q \in \mathcal{F}_\Theta$ is satisfied, and a solution can be obtained by running SGD on the empirical loss.

References

S. P. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Y. Polyanskiy and Y. Wu. *Information theory: From coding to learning*. Cambridge university press, 2025.