Def (3.1): If $X$ is a non-negative RV with $E(X) < \infty$, then the entropy of $X$ is

$$\text{Ent} \ X = E[\phi(X)] - \phi(E[X])$$

where $\phi(x) = x \log x$.

Properties:

• $\text{Ent} \ X = E[X \log X] - E[X] \log E[X] = E[X \log X] - E[X \log E[X]]$

  $$= E[X \log \frac{X}{E[X]}] \quad \text{(if } E[X] > 0)$$

• $\phi$ is convex, so by Jensen's, $\text{Ent} \ X \geq 0$ for all $X$.

• For any $X \geq 0, c > 0$, $\text{Ent} \ cX = E[cX \log (cX)] - E[(cX) \log E(cX)]$

  $$= c E[X \log X] + c E[X \log c] - c E[X \log E[X]] - c E[X \log c] = c \text{Ent} \ X.$$

• If $X \geq 0$ has $E[X] = 1$, we can use $X$ as a density:

  $$(Q(\mathcal{A}) = E[X 1_A]$. The relative entropy or Kullback-Leibler divergence $O(Q \text{ wrt } P)$ is

  $$D(Q \parallel P) = \text{Ent} \ X$$

Variational Characterization (Duality Formula): Let $X \geq 0$ satisfy $E[\phi(X)] < \infty$ (which implies $E[X] < \infty$). Then

$$\text{Ent} \ X = \sup \left\{ E[X^Y] : E[e^Y] = 1 \right\}$$

where $\sup$ is over $Y : \mathcal{Y} \to [-\infty, \infty]$, and $0 : (-\infty)$ and $e^{-\infty}$ are defined to be $0$.

$\text{Pf}$: Last time.
• **Exponential Hölder**: For \( X, Y \) with \( X \geq 0, E[X] < \infty \),
\[
E[XY] \leq c \left( E[X] + c E[X] \log E[e^Y/c] \right) \quad \text{for } c > 0.
\]

**Pf**: Take \( c = 1 \). Then assume \( X, e^Y \) integrable (else vacuous inequality).

Let \( Z = Y - \log E[e^Y] \). \( E[e^Z] = E\left[e^Y \cdot \frac{1}{E[e^Y]}\right] = 1 \). The variational characterization gives:
\[
\text{Ent } X \geq E[XY] - E[X] \log E[e^Y] \checkmark
\]

If \( c \neq 1 \), apply inequality to \( cX \) and \( Y/c \). \( \Box \)

• **Tensorization Inequality**: (\( \mathbb{R}^n, d, d \)) Let \( X_1, \ldots, X_n \) be independent and \( Z = f(X_1, \ldots, X_n) \) be nonnegative s.t. \( E[\phi(Z)] < \infty \). Then
\[
\text{Ent } Z \leq \sum_{i=1}^n E[\text{Ent}_i Z]
\]

where \( \text{Ent}_i Z \) is entropy of \( Z \) relative to \( X_i \) only (all other variables fixed).

**Pf**: Last time.
\[
\text{Ent}_i f(X_1, \ldots, X_n) = \text{Ent}(f(x_1, \ldots, x_i, \ldots, x_n))
\]
\[
= \text{Ent}\left[f(x_1, \ldots, x_n) \mid x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n\right]
\]
Page 3  Transportation Method:  Idea is to use variational characterization

of entropy which relates the cumulant generating function \( \Psi_2(t) = \log E[e^{tZ}] \)

(\(\Psi_2\))

to the relative entropy between given distribution \(P\) and another one \(Q\).

We estimate the "cost" of computing the expectation of \(Z\) relative to this
other distribution \(Q\) rather than relative to \(P\).

Gibbs Variational Principle:  Let \(Z\) be integrable and \(P\) take \(Z' = Z - E[Z]\),

\[
\forall t \in \mathbb{R}, \quad \Psi_2'(t) = \sup_{Q \ll P} \left[ t(E_Q[Z] - E[Z]) - D(Q || P) \right]
\]

(sup is over all \(Q\) which are absolutely continuous relative to \(P\).)

Proof:  (2) Assume \(\Psi_2'(t) < 0, t > 0\) since \(Z'\) has mean 0

(else vacuous)

\[
\log E[e^{tZ}] \geq \log e^{tE[Z]} = \log e^{t \cdot 0} = 0
\]

Let \(Q\) be absolutely cont. wrt \(P\) and let \(Y = \frac{dQ}{dP} \) (Radon-Nikodym derivative)

Let \(U = tZ' - \Psi_2'(t)\),

\[
E[e^U] = E[e^{tZ'}] = \frac{E[e^{Z'}]}{e^{\Psi_2'(t)}} = 1
\]

By the duality formula,

\[
D(Q || P) = \text{Ent} Y + E[Y] = tE[Z'] - \Psi_2'(t)E[Y] = t(E_Q[Z] - E[Z]) - \Psi_2'(t), \quad \text{so} \quad \Psi_2'(t) \geq t(E_Q[Z] - E[Z]) - D(Q || P).
\]
\( \leq \) Set \( a = \sup_{Q \ll P} \left[ t(E_{Q} Z - E Z) - D(Q \| P) \right] \). If \( a = \infty \), by previous direction, \( \psi_2(t) = \infty \) and done. Else \( a < \infty \). Note also \( a > 0 \) since 0 is attained by choosing \( P = Q \).

Define \( U = t \xi^{'} - a \). Our goal is \( E[e^{U}] \leq 1 \). Then \( Ee^{t\xi'} \leq e^{a} \) and \( \psi_2(t) \leq a \).

Claim: This follows from (\( \ast \)) \( E[UY] \leq \text{Ent} Y \) for all \( Y \geq 0, E[Y] < \infty \).

Assuming (\( \ast \)), Set \( Y_n = \min \{ U_n, n \} \), \( Y_n = e^{U_n} / E[e^{U_n}] \). \( E[e^{Y_n}] = 1 \), \( \forall n \Rightarrow \)

\[
\frac{E[ue^{Un}]}{E[e^{Un}]} = E[UY_n] \leq \text{Ent} Y_n = E[Y_n \log Y_n] - E[Y_n] \log \frac{E[Y_n]}{E[1]} = E[Y_n \log Y_n]
\]

\[
= E\left[ \frac{e^{Un}}{E[e^{Un}]} \log \frac{e^{Un}}{E[e^{Un}]} \right] = \frac{1}{E[e^{Un}]} \left[ e^{Un} U_n - e^{Un} \log E[e^{Un}] \right]
\]

So \( E[ue^{Un}] \leq E[e^{Un} U_n - e^{Un} \log E[e^{Un}]] \)

\[
E[e^{Un}] \log E[e^{Un}] \leq E[e^{Un} U_n - U e^{Un}]
\]

\[
\log E[e^{Un}] \leq \frac{E[e^{Un}(U_n - U)]}{E[e^{Un}]} \leq 0 \quad \text{because} \quad e^{Un} > 0 \quad \text{and} \quad U_n - U \leq 0
\]

\( E[e^{Un}] \leq 1 \). By MCT, \( E[e^{Un}] \leq 1 \).

Done except proof of (\( \ast \)).
Proof of (b): Let \( Y \geq 0 \) w/ \( E[Y] = 1 \). Define \( Q \) by \( \frac{dQ}{dP} = Y \). Then

\[
E[UY] = E[(Y - a) \frac{dQ}{dP}] = E(Q \mathbb{1}_Z - E_Z) - a
\]

\[
= E(Q \mathbb{1}_Z - E_Z) - D(Q \parallel P) - a + \text{Ent} + Y
\]

\[
\leq a - a + \text{Ent} + Y = \text{Ent} + Y. \quad \checkmark
\]

If \( E[Y] = 0 \), then \( Y = 0 \) a.s. \( \psi(M) \) is trivial.

If \( E[Y] > 0 \) but not 1, apply argument to \( \frac{Y}{E[Y]} \):

\[
E(\psi(U \cdot \frac{Y}{E[Y]})) \leq \text{Ent} \frac{Y}{E[Y]}
\]

so \( E[UY] \leq \text{Ent} + Y. \quad \checkmark
\]

A Transportation Lemma: Let \( \mathbb{Z} \) be R.V., \( E[\mathbb{Z}] = 0 \), then

\[
\psi_{\mathbb{Z}}(t) \leq \frac{t^2}{2} \quad \text{for } t > 0
\]

iff \( E_Q[Z] - E[Z] \leq \sqrt{2}D(Q \parallel P) \)

for all \( Q \ll P \).

**Proof:** Variational Characterization (Gibbs) says:

\[
\psi_{\mathbb{Z}}(t) = \sup_{Q \ll P} \left[ E(Q \mathbb{1}_Z - E_Z) - D(Q \parallel P) \right]
\]

So \( \psi_{\mathbb{Z}}(t) \leq \frac{t^2}{2} \) iff \( E(Q \mathbb{1}_Z - E_Z) - D(Q \parallel P) - \frac{t^2}{2} \leq 0 \quad \forall Q \ll P \).

Can restrict to looking at \( Q \) where \( E_Q \mathbb{1}_Z - E_Z \geq 0 \), else trivial.

The left side is quadratic in \( t \) with nonnegative coefficients, so maximum is at \( t = \frac{a_1}{2a_3} \). \( (a_1 t - a_2 - a_3 t^2) \) Max value is \( \frac{a_1^2}{2a_3} - a_2 - a_3 \cdot \frac{a_1^2}{2a_3} = \frac{a_1^2}{2a_3} - a_q \).

So for \( \frac{a_1^2}{2a_3} - a_q \leq 0 \), we need \( a_1 \leq \sqrt{a_q a_3} \). Thus \( \psi_{\mathbb{Z}}(t) \leq \frac{t^2}{2} \) \( \forall t > 0 \) iff

\[
E_Q \mathbb{1}_Z - E_Z \leq \sqrt{2}D(Q \parallel P) \quad \forall Q \ll P
\]

for which left side is positive.

Same as \( \forall Q \ll P \). \( \square \)
General Setup of Transportation Method:

\[ X_1, \ldots, X_n \text{ independent}, \quad f: \mathbb{R}^n \to \mathbb{R}, \quad d: \mathbb{R} \times \mathbb{R} \to [0, \infty) \text{ is a pseudo-metric}. \]

Assume regularity property:

\[ f(y) - f(x) \leq \sum_{i=1}^{n} c_i d(y_i, x_i) \quad \text{for } x, y \in \mathbb{R}^n, \quad c_i \geq 0. \]

We will focus on the example \[ d(a, b) = \begin{cases} 1, & a \neq b \\ 0, & a = b \end{cases}. \]

The regularity property is a bounded differences condition w/ \( c_1, \ldots, c_n \).

Let \( Z = f(X_1, \ldots, X_n) \). To use transportation lemma to show \( \Psi^2(z) \leq \frac{\mathbb{E}^2(z)}{2} \)

we need to bound \( E_Q Z^2 - E_Z Z \). We consider a coupling \( \mathcal{B} P, Q \); \( P \), the distribution of \( (X, Y) \) where \( X \) has distr. \( P \) and \( Y \) has distr. \( Q \).

Let \( \mathcal{P}(P, Q) \) be the set of all such joint distributions \( P \).

In our case \( X = (X_1, \ldots, X_n) \), \( P \) is joint distribution. If \( Q \preceq P \),

then \( Y = (Y_1, \ldots, Y_n) \) with distr. \( Q \) and \( \inf_{P \in \mathcal{P}(P, Q)} \) we use

regularity property:

\[ E_Q Z^2 - E_Z Z = E_P (Z(Y)) - E_P (Z(X)) = E_P (Z(Y) - Z(X)) \leq \sum_{i=1}^{n} c_i E_p (d(X_i, Y_i)) \]

\[ \leq \left( \sum_{i=1}^{n} c_i^2 \right)^{1/2} \left( \mathbb{E} \left[ E_p [d(X_i, Y_i)]^2 \right] \right)^{1/2}. \]

So proving Gaussian Concentration

Candy Shao's

\[ \inf_{P \in \mathcal{P}(P, Q)} \sum_{i=1}^{n} E_p (d(X_i, Y_i))^2 \leq \frac{2\gamma}{\gamma} D(Q \parallel P) \text{ for } Q \preceq P \text{ and } C = \sum_{i=1}^{n} c_i. \]

\[ \text{min. "cost" of transporting } X \text{ w/ distr. } P \text{ to } Y \text{ w/ distr. } Q \text{ on same space, } \text{Transportation Cost Inequality}. \]
In our example \( d(a, b) = \sum_i \begin{cases} 0 & a = b \\ 1 & a \neq b \end{cases} \), the Tr mpg. Cert Ieg.

\[
\inf_{P \in \mathcal{P}(P, \Omega)} \sum_{i=1}^{n} (P(x_i \neq y_i))^2 \leq \frac{2V}{\mathcal{C}} D(Q \parallel P)
\]

**Defn:** If \( P, Q \) are prob. measures on the same space \((\mathcal{F}, \mathcal{F})\), the total variation distance between them is defined as:

\[
V(P, Q) = \sup_{A \in \mathcal{F}} |P(A) - Q(A)|.
\]

The sup is attained at some \( A^* \in \mathcal{F} \).

**Pf:** Defint \( J = P + Q \). \( P, Q \ll J \). Let \( f = \frac{dP}{dJ} \), \( g = \frac{dQ}{dJ} \).

Let \( E = \{w: f(w) \geq g(w)\} \). For \( A \in \mathcal{F} \),

\[
|P(A) - Q(A)| = |\sum_{A} (f-g) dJ| = |\sum_{A \in \mathcal{E}} (f-g) dJ - \sum_{A \in \mathcal{E}^c} (g-f) dJ|
\]

so

\[
|P(A) - Q(A)| \leq \max \sum_{A \in \mathcal{E}} (f-g) dJ, \sum_{A \in \mathcal{E}^c} (g-f) dJ
\]

\[
\leq \max \sum_{E} (f-g) dJ, \sum_{E^c} (g-f) dJ = \max \sum (P(E) - Q(E), Q(E^c) - P(E^c))
\]

but these are equal. Upper bound is \( P(E) - Q(E) \). Take \( A = E \), and the bound is obtained. \( \square \)
The Pinsker's Inequality: Let $P, Q$ be prob. measures on $(\Omega, F)$, s.t. $Q \ll P$. Then $\psi(P, Q)^2 \leq \frac{1}{2} D(Q \| P)$.

**Proof:** Let $A^* \in F$ s.t. $\psi(P, Q) = Q(A^*) - P(A^*)$, $\mathbb{P} = \mathbb{1}_{A^*}$ to hold, so by Hoeffding's Inequality

$$\psi^2(t) \leq \frac{t^2}{8} \quad \forall t,$$

On the other hand, let $\gamma = \frac{dQ}{dP}$, then the Gibbs Var Prize gives

$$\psi^2(t) = t(E_Q \mathbb{E} - E_Z) - D(Q \| P) = t(Q(A^*) - P(A^*)) - D(Q \| P)$$

$$= t \psi(Q, P) - D(Q \| P).$$

Choose $t = 4 \psi(Q, P)$

$$\frac{1}{8} \frac{16 \psi(Q, P)^2}{8} \geq \psi^2(4 \psi(Q, P)) = \psi(Q, P)^2 - D(Q, P).$$

$$\psi(P, Q)^2 \leq \frac{1}{2} D(Q \| P).$$

So now proving $\psi^2(t) \leq \frac{t^2}{2}$ is reduced to proving

$$\inf_{P \in \mathcal{P}(P, Q)} \sum_{i=1}^n (P(X \neq Y_i))^2 \leq \frac{4}{\psi} \psi(Q, P)^2.$$

We'll do this by induction on $n$. For base case we need $\psi^2(0) \leq \frac{1}{2} D(Q \| P)$.
Theorem: (Base Case) If $P, Q$ are probability distributions on $\mathbb{R}^k$, then

$$\min_{P \in \mathcal{P}(P, Q)} P(X \neq Y) = V(P, Q)$$

where $(X, Y)$ has distribution $P$.

Proof: (i) Let $P \in \mathcal{P}(P, Q)$, $A \in \mathcal{F}$. $|P(A) - Q(A)| = |P(X \in A) - P(Y \in A)|$

$$\leq P(\{X \in A\} \Delta \{Y \in A\}) \leq P(X \neq Y).$$

(Holds for all $A \in \mathcal{F}$, so in particular for $A^*$ achieving $V(P, Q)$.

(ii) Set $\alpha = V(P, Q)$. We may assume $\alpha > 0$, else $P = Q$ and let $X = Y$. Furthermore, if $\alpha = 1$, $P \perp Q \perp (is mutually singular to), so \forall P$ coupling, $P(X \neq Y) = 1$. If $\alpha < 1$, want to define $P$ on $\mathbb{R} \times \mathbb{R}$ to be $\alpha P_1 + (1-\alpha) P_2$ where $P_1$ is concentrated on $\{(x, y) : x \neq y\}$, $P_2$ on $\{(x, y) : x = y\}$. If we do this so marginal of $X$ is $P$, marginal of $Y$ is $Q$, then

$$P(X \neq Y) = \alpha P_1(X \neq Y) + (1-\alpha) P_2(X \neq Y) = \alpha$$

Set $\alpha = P + Q$, $f = \frac{dP}{d\lambda}$, $g = \frac{dQ}{d\lambda}$.

For $A, B \in \mathcal{F}$, define

$$P_1(A \times B) = \frac{1}{\alpha} \int_{A \times B} (f(x) - g(x)) + (g(y) - f(y))_+ \, dA(x) \, dA(y)$$

$$P_2(\{(x, y) : x \neq y\}) = \frac{1}{1-\alpha} \int_A \min \{f(x), g(x)\} \, dA(x)$$

$P_1$ is supported on $\{(x, y) : x \neq y\}$ (integrands is 0 when $x = y$)

$P_2$ is supported on $\{(x, y) : x = y\}$. Furthermore $P_1(N \times N) = P_2(N \times N) = 1$. Marginals

$$P_1(N \times \mathbb{R}) = \frac{1}{\alpha} \int f(x) - g(x) \, dA(x) \int \min \{f(y), g(y)\} \, dA(y) = 1$$

$[see \ page \ 399]$

$$P_2(N \times \mathbb{R}) = \frac{1}{1-\alpha} \int g(y) \, dA(x) \int [f(x) - g(x)]_+ \, dA(y)$$

$$= \frac{1}{2(1-\alpha)} \int [f(x) - g(x)]_+ \, dA(x) = \frac{1}{2(1-\alpha)} \int [f(x) - g(x)]_+ \, dA(x)$$

$[see \ page \ 399]$
Lemma 1.3.1: For positive integers $n_1, n_2, \ldots, n_d$ with $n_i > 1$ for $i = 1, 2, \ldots, d$, let $X = (X_1, X_2, \ldots, X_d)$ be a random vector with joint distribution $P$. Then, for any $i = 1, 2, \ldots, d$, we have

$$D_i(D, P) = \sum_{i=1}^{d} D_i(D, P).$$

Theorem 1.3.1: For a random vector $X = (X_1, X_2, \ldots, X_d)$ with joint distribution $P$, let $P_{i\mid X_i}$ be the conditional distribution of $X_i$ given $X_i$. Then, for any $i = 1, 2, \ldots, d$, we have

$$D_i(D, P) = \sum_{i=1}^{d} D_i(D, P).$$

Proof: The proof follows from the definition of $D_i(D, P)$ and the properties of conditional expectation.

See pg 25 for more details.
Then by Jensen's,
\[
D(Q, P) \geq 2 \sum_{i=1}^{n-1} (\int f(x_i, y_i) dQ_n(y) )^2 + 2 P_n(x_n = y_n)^2
\]
\[
= 2 \sum_{i=1}^{n-1} (P(x_i \neq y_i)^2 + 2P(x_n \neq y_n)) \geq 2 \sum_{i=1}^{n} (P(x_i \neq y_i)^2).
\]
(by previous line)

Summary: Bounded Differences by transportation:

Let \( X = (X_1, \ldots, X_n) \) be independent and \( f: \mathbb{R}^n \to \mathbb{R} \) satisfy bounded differences condition w/ \( c_1, \ldots, c_n \geq 0 \)

\[
|f(x_1, \ldots, x_n) - f(x_1, \ldots, x_i', \ldots, x_n)| \leq c_i \quad \forall x_1, x_1', \ldots, x_n, x_n' \in \mathbb{R}
\]

Then for \( Z = f(X) \), \( \mathbb{E}_Z - \mathbb{E}_Z^2(t) \leq \frac{t^2}{2} \) for \( t > 0 \) with \( v = \frac{c}{4} \).

\[
P(f(X) - Ef(X) > t) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^{n} c_i^2}\right) \quad \text{for } t > 0.
\]

\( Pf: \) Transp. Lemma gives \( \Psi_Z(t) \leq \frac{t^2}{2} \) for \( t > 0 \) iff

\[
EQZ - EQ2 \leq \sqrt{2vD(Q||P)} \quad \text{for } Q \ll P.
\]

By earlier results,

\[
EQZ - EQ2 \leq \inf_{P \in P(Q, P)} \left( \sum_{i=1}^{n} (P(X_i \neq Y_i))^2 \right)^{1/2}
\]

By Marton's Inequality

\[
\leq \sqrt{\frac{c}{2}} D(Q||P). \quad \text{So } \Psi_Z(t) \leq \frac{t^2}{2} \text{ for } t > 0
\]

with \( v = \frac{c}{4} \).