6.7480 Cheat Sheet

L1-2: Entropy, Mutual Information

Entropy is given by

$$H(X) = -\sum p(X)\log p(X)$$

while the **conditional entropy** is defined by

$$H(X|Y) = -\sum p(X,Y)\log p(X|Y)$$

and has the following properties:

Properties of Entropy:

(a) (Positivity) $H(X) \geq 0$ with equality iff X is constant.
(b) (Uniform distribution maximizes entropy) For finite $\mathcal{X}$, $H(X) \leq \log|X|$, with equality iff X is uniform on X
(c) (Invariance under relabling) $H(X) = H(f(X))$ for any bijective f
(d) (Conditioning reduces entropy) $H(X|Y) \leq H(X)$ with equality iff X and Y are independent
(e) (Chain Rule) $H(X,Y) = H(X) + H(Y|X) \leq H(X) + H(Y)$
(f) (Entropy under deterministic transformation) $H(X) = H(X, f(X)) \geq H(f(X))$ with equality iff f is injective on the support of P_X
(g) (Full Chain Rule) $H(X_1, \cdots, X_n) = \sum_{i=1}^n H(X_i|X^{i-1}) \leq \sum H(X_i)$ with equality iff the X_i are independent.

Method of Types: Let $n_1, \cdots, n_k \geq 0$ be such that $n = \sum n_j$ and denote $P = (p_1, \cdots, p_k)$ with $p_i = n_i/n$. Then,

$$\frac{1}{(1+n)^{k-1}\exp\{nH(p)\}} \leq \binom{n}{n_1, \ldots, n_k} \leq nH(p)$$

KL Divergence is given by

$$D(P\|Q) = \begin{cases} \int p\log\frac{p}{q}\,d\mu, & P \ll Q, \\ +\infty, & \text{otherwise.} \end{cases}$$

Properties of KL Divergence:

(a) **(Information Inequality)** $D(P||Q) \geq 0$ with equality iff $P = Q$
(b) **(Min characterization)** $\min_Q \mathbb{E}[\log 1/Q(X)] = H(X)$
(c) **(Chain Rule)** $D(P_{X,Y}||Q_{X,Y}) = D(P_X||Q_X) + D(P_{Y|X}||Q_{Y|X}|P_X)$
(d) **(Monotonicity)** $D(P_{X,Y}||Q_{X,Y}) \geq D(P_Y||Q_Y)$
(e) **(Full Chain Rule)** $D(P_{X_1,\cdots,X_n}||Q_{X_1,\cdots,X_n}) = \sum_{i=1}^n D(P_{X_i|X^{i-1}}||Q_{X_i|X^{i-1}}|P_{X^{i-1}}) \geq \sum_{i=1}^n D(P_{X_i}||Q_{X_i})$
(f) **(Conditioning increases divergence)** Given $P_{Y|X}, Q_{Y|X}$ and P_X, let $P_Y = P_{Y|X} \circ P_X$ and $Q_Y = Q_{Y|X} \circ P_X$, Then

$$D(P_Y||QY) \leq D(P_{Y|X}||Q_{Y|X}|P_X)$$

, with equality iff $D(P_{Y|X}||Q_{Y|X}|P_X) = 0$.
(g) **(DPI)**: For $P_Y = P_{Y|X} \circ P_X$ and $Q_Y = P_{Y|X} \circ Q_X$,

$$D(P_Y||Q_Y) \leq D(P_X||Q_X)$$

with equality iff $D(P_{X|Y}||Q_{X|Y}|P_Y) = 0$.

By DPI, under a deterministic transform f, divergence is $\leq$ unless surjective (see Corollary 2.18).

Locally, **KL divergence is χ^2-like**:

$$\liminf_{\lambda\to 0} \frac{1}{\lambda^2} D(\lambda P + \bar{\lambda}Q||Q) = \frac{\log e}{2}\chi^2(P||Q)$$

where both sides are finite or infinite simultaneously.

Fano's Inequality: Let $|\mathcal{X}| = M < \infty$ and $X \to Y \to \hat{X}$. Let $P_e = \mathbb{P}[X \neq \hat{X}]$. Then,

$$H(X|Y) \leq P_e \log(M-1) + h(P_e)$$

where h is the binary entropy. Furthermore, if $P_{max} = \max_{x\in\mathcal{X}} P_X(x) > 0$ then regardless of $|\mathcal{X}|$,

$$I(X;Y) \geq (1-P_e)\log\frac{1}{P_{max}} - h(P_e)$$

L3: Mutual Information

The **Mutual Information** is given by

$$I(X;Y) = D(P_{XY}\|P_XP_Y) = \mathbb{E}\left[\log\frac{p(y|x)}{p(y)}\right] = H(X) - H(X|Y)$$

The **conditional MI** is

$$I(X;Y|Z) = \mathbb{E}_Z\, D(P_{XY|Z}\|P_{X|Z}P_{Y|Z}) = H(X|Z) - H(X,Y|Z)$$

Properties of Mutual Information:

(a) **(Mutual information as conditional divergence)** If $\mathcal{Y}$ is standard Borel,

$$I(X;Y) = D\big(P_{Y|X}||P_Y||P_X\big)$$

(b) **(Symmetry)** $I(X;Y) = I(Y;X)$.
(c) **(Positivity)** $I(X;Y) \geq 0$, with equality iff $X \perp Y$.
(d) **(Deterministic maps)** For any measurable f,

$$I\big(f(X);Y\big) \leq I(X;Y).$$

If f is one-to-one with a measurable inverse, then $I(f(X);Y) = I(X;Y)$.
(e) **(More data ⇒ more information)** ; $I(X_1,X_2;Z) \geq I(X_1;Z)$, (by symmetry also $I(X_1,X_2;Z) \geq I(X_2;Z)$).
(f) $I(X;Z|Y) \geq 0$ with equality iff $X \to Y \to Z$
(g) (**Chain Rule**) $I(X,Y;Z) = I(X;Z) + I(Y;Z|X) = I(Y;Z) + I(X;Z|Y)$
(h) (**DPI**): If $X \to Y \to Z$ then

$$I(X;Z) \leq I(X;Y)$$

with equality iff $X \to Z \to Y$
(i) If $X \to Y \to Z \to W$, then $I(X;W) \leq I(Y;Z)$
(j) (Injective Invariance) If f and g are injective, then $I(f(X),g(Y)) = I(X,Y)$

Example Channels: AWGN (Additive White Gaussian Noise), BSC (Binary Symmetric Channel)

Golden Formula for Mutual Information: For any Q_Y we have

$$D(P_{Y|X}||Q_Y|P_X) = I(X;Y) + D(P_Y||Q_Y)$$

Thus, if $D(P_Y||Q_Y) < \infty$,

$$I(X;Y) = D(P_{Y|X}||Q_Y|P_X) - D(P_Y||Q_Y)$$

Mutual Information as Center of Gravity: For any Q_Y,

$$I(X;Y) \leq D(P_{Y|X}||Q_Y|P_X)$$

and thus

$$I(X;Y) = \min_{Q_Y} D(P_{Y|X}||Q_Y|P_X)$$

If $I(X;Y) < \infty$, the unique minimizer is $Q_Y = P_Y$.

Mutual Information as Distance to Product Distributions:

$$I(X;Y) = \min_{Q_X, Q_Y} D(P_{X,Y}||Q_X Q_Y)$$

with unique minimizer $(Q_X, Q_Y) = (P_X, P_Y)$.

L4: Variational Characterization/Convexity

Donsker–Varadhan (DV): For P, Q probability measures on $\mathcal{X}$ and denote
$\mathcal{C}_Q = \{f : \mathcal{X} \to \mathbb{R} \cup \{-\infty\} : 0 < \mathbb{E}_Q[\exp\{f(X)\}] < \infty\}$. Then,

$$D(P\|Q) = \sup_{f \in \mathcal{C}_Q} \left\{\mathbb{E}_P[f] - \log \mathbb{E}_Q[e^f]\right\}.$$

This can be strengthened to bounded continuous functions.
A similar characterization is given by Gibbs:

Gibbs Variational Principle: for $f : \mathcal{X} \to \mathbb{R} \cup \{-\infty\}$ measurable and Q a probability measure on X,

$$\log \mathbb{E}_Q[\exp\{f(X)\}] = \sup_P \mathbb{E}_P[f(X)] - D(P||Q)$$

where the supremum is taken over all $D(P||Q) < \infty$. The unique maximizer if finite is $P = Q^f$.

Convexity of KL (Theorem 5.1): $(P, Q) \mapsto D(P||Q)$ is convex (but not strictly convex). It is *strongly convex* with respect to Total Variation.

Concavity of Entropy (Theorem 5.2): $P_X \mapsto H(P_X)$ is concave. If $P_{Y|X}$ is any channel, then $P_X \mapsto H(X|Y)$ is concave. If $\mathcal{X}$ is finite, $P_X \mapsto H(X|Y)$ is continuous.

Convexity/Concavity of MI (Theorem 5.3):

1. For fixed $P_{Y|X}$, $P_X \mapsto I(P_X, P_{Y|X})$ is concave.
2. For fixed P_X, $P_{Y|X} \mapsto I(P_X, P_{Y|X}$ is convex.

L5: Capacity Saddle Point

We call $C := \sup_{P_X \in \mathcal{P}} I(X;Y)$ the **capacity**.

Capacity Saddle Point (Theorem 5.4): Let $\mathcal{P}$ be a convex set of distributions on $\mathcal{X}$. Suppose there exists $P_X^* \in \mathcal{P}$, a capacity-achieving distribution such that

$$\sup_{P_X \in \mathcal{P}} I(P_X, P_{Y|X}) = I(P_X^*, P_{Y|X}) := C$$

Let $P_Y^* := P_{Y|X} \circ P_X^*$ be called the capacity-achieving output distribution. Then, for all $P_X \in \mathcal{P}$ and Q_Y, we have

$$D(P_{Y|X}||P_Y^*|P_X) \le D(P_{Y|X}||P_Y^*||P_X^*) \le D(P_{Y|X}||Q_Y||P_X^*)$$

Similarly, we get the famous minimax theorem:

Minimax Theorem (Theorem 5.6): Under the same conditions,

$$\begin{aligned} C :&= \max_{P_X \in \mathcal{P}} \min_{Q_Y} D(P_{Y|X}||Q_Y|P_X) \\ &= \min_{Q_Y} \sup_{P_X \in \mathcal{P}} D(P_{Y|X}||Q_Y|P_X) \end{aligned}$$

As information radius, we get the following:

Capacity as Information Radius (Corollary 5.8):

$$\begin{aligned} C = \sup_{P_X} I(X;Y) &\le \inf_Q \sup_{x \in \mathcal{X}} D(P_{Y|X=x}||Q) \\ &\le \sup_{x,x'} D(P_{Y|X=x}||P_{Y|X=x'}) \end{aligned}$$

L6-8: f-Divergences

f-divergences: For convex f with $f(1) = 0$,

$$D_f(P\|Q) = \mathbb{E}_Q\left[f\left(\frac{p}{q}\right)\right] = \sum_x Q(x) f\left(\frac{P(X)}{Q(X)}\right)$$

with $f(0) = f(0+)$ and $0\log(0/0) = 0$ and $0f(a/0) = af'(\infty)$.

Common f-divergences

(a) (**KL-Divergence**) $f(x) = x\log x$ and

$$D(P||Q) = \sum_x P(x) \log\left(\frac{P(X)}{Q(X)}\right)$$

(b) (χ^2-**Divergence**) $f(x) = (x-1)^2$ and

$$\begin{aligned} \chi^2(P||Q) &= \sum_x Q(x)\left(\frac{P(X)}{Q(X)} - 1\right)^2 \\ &= \sum_{x \in \mathcal{X}} \frac{P(X)^2}{Q(X)} - 1 \end{aligned}$$

(c) (**Squared Hellinger-Distance**): $f(x) = (\sqrt{x} - 1)^2$ and

$$\begin{aligned} H^2(P,Q) &= \sum_{x \in \mathcal{X}} \left(\sqrt{P(X)} - \sqrt{Q(X)}\right)^2 \\ &= 2 - 2\sum_{x \in \mathcal{X}} \sqrt{P(X)Q(X)} \end{aligned}$$

(d) (**Total-Variation**): $f(x) = \frac{1}{2}|x-1|$ and

$$TV(P,Q) = \frac{1}{2}\sum_{x \in \mathcal{X}} |P(X) - Q(X)|$$

(e) (**Le Cam Divergence**) $f(x) = \frac{1-x}{2x+2}$ and

$$LC(P,Q) = \frac{1}{2}\sum_{x \in \mathcal{X}} \frac{(P(X) - Q(X))^2}{P(X) + Q(X)}$$

(f) (**Jenson-Shannon**) $f(x) = x\log\frac{2x}{x+1} + \log\frac{2}{x+1}$:

$$JS(P,Q) = D\left(P||\frac{P+Q}{2}\right) + D\left(Q||\frac{P+Q}{2}\right)$$

Properties of f-Divergence:

(a) (**Additivity**): $D_{f_1+f_2}(P||Q) = D_{f_1}(P||Q) + D_{f_2}(P||Q)$

(b) (**Non-Negativity**): $D_f(P||Q) \ge 0$. If f is strictly convex at 1, then $D_f(P||Q) = 0$ iff $P = Q$

(c) (**Joint-Convexity**): $(P, Q) \mapsto D_f(P||Q)$ is jointly convex. $P \mapsto D_f(P||Q)$ and $Q \mapsto D_f(P||Q)$ are also convex.

(d) (**Zero Condition**) $D_f(P||Q) = 0$ for all $P \ne Q$ iff $f(x) = c(x-1)$ and for any other f we have $D_f(P||Q) = f(0) + f'(\infty)$

(e) If $P_X = Q_X$ and only the Markov Kernel differs,

$$D_f(P_{X,Y}||Q_{X,Y}) = D_f(P_{Y|X}||Q_{Y|X}|P_X)$$

(f) If the Markov Kernel is the same $P_{X|Y} = Q_{X|Y}$ then

$$D_f(P_{X,Y}||Q_{X,Y}) = D_f(P_X||Q_X)$$

and

$$D_f(P_X P_Y||Q_X P_X) = D_f(P_X||Q_X)$$

(g) (**Center of Gravity**) We have

$$D_f(P_Y||Q_Y) \le D_f(P_{Y|X}||Q_{Y|X}|P_X)$$

(h) (**DPI**): for a channel $P_Y = P_{Y|X} \circ P_X$ and $Q_Y = P_{Y|X} \circ Q_X$ we have

$$D_f(P_Y||Q_Y) \le D_f(P_X||Q_X)$$

Note, unlike KL divergence, there is no exact Chain Rule, Donsker Varadhan, and (later) tensorization for general f-divergences.

Tensorization

For the following f-divergences there are is a nice tensorization:

(a) **KL-divergence**: We have

$$D(P^{\otimes n}||Q^{\otimes n}) = \sum_{i=1}^{n} D(P_i||Q_i)$$

(b) **Squared-Hellinger**: We have

$$H^2(P^{\otimes n}, Q^{\otimes n}) = 2 - 2\prod\left(1 - \frac{H^2(P_i, Q_i)}{2}\right)$$

(c) χ^2**-divergence**: We have

$$1 + \chi^2(P^{\otimes n}||Q^{\otimes n}) = \prod_{i=1}^{n}(1 + \chi^2(P_i||Q_i))$$

Note, Total variation has no nice tensorization!

Comparison Inequalities

The main ones:

(a) **(f-divergences are locally** χ^2**)**:

$$D_f(P||Q) = \frac{f''(1)}{2}\chi^2(P||Q) + o(\chi^2(P||Q))$$

as $P \to Q$

(b) (**Pinsker's Inequality**)

$$TV^2(P, Q) \le \frac{1}{2\log e} D(P\|Q)$$

(c) (χ^2 **vs. TV**) We have

$$4TV^2(P, Q) \le f(TV(P, Q)) \le \chi^2(P||Q)$$

for any convex increasing bijection $f : [0, 1) \to (0, \infty)$

(d) (**TV vs. Hellinger**)

$$\frac{1}{2}H^2(P, Q) \le TV(P, Q) \le H(P, Q)\sqrt{\left(1 - \frac{H^2(P, Q)}{4}\right)} \le 1$$

(e) (**KL vs.** χ^2**)**:

$$D(P||Q) \le \log(1 + \chi^2(P||Q)) \le (\log e)\chi^2(P||Q)$$

Variational Characterization

Generalized Donsker-Varadhan: Let P and Q be probability measures on $\mathcal{X}$. Fix an extension (typically, set f_{ext} to be ∞ on $x < 0$) f_{ext} of f and let f^*_{ext} be its convex conjugate, i.e.,

$$f^*_{\text{ext}}(y) = \sup_{x \in \mathbb{R}}\{xy - f_{\text{ext}}(x)\}.$$

Denote $\text{dom}(f^*_{\text{ext}}) \triangleq \{y : f^*_{\text{ext}}(y) < \infty\}$. Then

$$D_f(P\|Q) = \sup_{g:\, \mathcal{X} \to \text{dom}(f^*_{\text{ext}})} \left\{\mathbb{E}_P[g(X)] - \mathbb{E}_Q\left[f^*_{\text{ext}}(g(X))\right]\right\}$$

where the supremum can be taken over either (a) all simple g, or (b) all g satisfying $\mathbb{E}_Q\left[f^*_{\text{ext}}(g(X))\right] < \infty$.

All of these inequalities are special cases of the Joint-Range Theorem:

Joint Range Theorem (Harremoës-Vajda): for two f-divergences $D_f(P||Q)$ and $D_g(P||Q)$ their joint range is a subset of $[0, \infty)^2$

$$\mathcal{R} := \{D_f(P||Q), D_g(P||Q))\}$$

Let $\mathcal{R}_k$ be thje same but on measures on $[k]$. Then,

$$\mathcal{R} = \text{co}(\mathcal{R}_2) = \mathcal{R}_4$$

where co is the convex hull.

Golden/Inf–Sup for MI:

$$I(X; Y) = \inf_{Q_Y} \sup_{f} \left\{\mathbb{E}_{P_{XY}}[f] - \log \mathbb{E}_{P_X Q_Y}[e^f]\right\}.$$

L9: Statistics

Fisher Information:

$$J(\theta) := \mathbb{E}_\theta\left[\left(\nabla_\theta \log p_\theta(X)\right)^2\right] = \mathbb{E}_\theta[-\nabla^2_\theta \log p_\theta(X)]$$

under certain regularity conditions. As a Hessian, we have

$$D(P_\theta||P_{\theta+\delta}) = \frac{1}{2}\delta^T J_F(\theta)\delta + o(\|\delta\|^2)$$

so

$$J_F(\theta) = \nabla^2_h D(P_\theta||P_{\theta+h})|_{h=0}$$

Suppose we have a latent parameter θ and have n i.i.d. observations $X_1, \cdots, X_n \sim P_\theta$. Based on this, we can estimate an empirical $\hat{\theta}$. We introduce three bounds on the quadratic loss: the first two are frequentist, and assume θ is given. The last one (BCR/Van-Trees) is Bayesian and assumes θ comes from some prior.

Hammersley-Chapman-Robbins (HCR) Bound: We have

$$R_\theta(\hat{\theta}) := \mathbb{E}_\theta[(\hat{\theta} - \theta)^2] \ge \text{Var}_\theta(\hat{\theta}) \ge \sup_{\theta' \ne \theta} \frac{(\mathbb{E}_\theta[\hat{\theta}] - \mathbb{E}_{\theta'}[\hat{\theta}])^2]}{\chi^2(P_{\theta'}||P_\theta)}$$

For an *unbiased estimator* $\mathbb{E}_\theta[\hat{\theta}] = \theta$ for all $\theta \in \Theta$, we have under some regularity conditions:

Cramer-Rao:

$$\text{Var}_\theta(\hat{\theta}) \ge \frac{1}{J_F(\theta)}$$

Van-Trees (Bayesian Cramer-Rao): Let π be a differentiable prior density on the interval $[\theta_0, \theta_1]$ such that $\pi(\theta_0) = \pi(\theta_1) = 0$. Define the Fischer information of the prior $J(\pi) = \mathbb{E}[(\partial_\theta \log \pi(\theta))^2]$. Then, the quadratic risk

$$R^*_\pi = \inf_{\hat{\theta}} \mathbb{E}[(\theta - \hat{\theta})^2] \ge \frac{1}{\mathbb{E}_{\theta\sim\pi}[J_F(\theta)] + J(\pi)}$$

L10-12: Compression

Goal: compress $\mathcal{X}^n \to \{0,1\}^* \to \mathcal{X}^n$.

Shannon Optimal Encoder (Theorem 10.2/3): The optimal encoder for some $X \in \mathcal{X}$ has codeword length $\ell(f^*(i)) = \lfloor \log_2 i \rfloor$ and minimum expected length. Namely,

$$H(X) \text{ bits} - \log_2[e(H(X)+1)] \le \mathbb{E}[\ell(f^*(X))] \le H(X) \text{ bits}$$

Code-Length Distribution Theorem: $\forall \tau > 0, k \ge 0$,

$$\begin{aligned}\mathbb{P}\left[\log_2 \frac{1}{P_X(X)} \le k\right] &\le \mathbb{P}[l(f^*(X) \le k] \\ &\le \mathbb{P}\left[\log_2 \frac{1}{P_X(X)} \le k+\tau\right] + 2^{-\tau+1}\end{aligned}$$

Entropy Rate

The **entropy rate** is defined $\bar{H} = \lim_{n\to\infty} \frac{1}{n} H(X^n)$ if it exists.

Shannon-McMillan-Breiman (Theorem 12.5): Let $\mathbb{S} = \{S_1, S_2, \cdots\}$ be stationary and ergodic with entropy rate $\bar{H}$. Then

$$\frac{1}{n} \log \frac{1}{\mathbb{P}_{S^n}(S^n)} \to \mathcal{H}$$

almost surely in L^1.

Asymptotic Equipartition Property (AEP) (Theorem 12.7): Let $\{S_1, S_2, \cdots\}$ be stationary and ergodic. For any δ, define

$$T_n^\delta = \left\{ s^n : \left| \frac{1}{n} \log \frac{1}{P_{S^n}(S^n)} - \mathcal{H} \right| \le \delta \right\}$$

Then

(a) $\mathbb{P}[S^n \in T_n^\delta] \to 1$ as $n \to \infty$
(b) $\exp\{n(\mathcal{H}-\delta)\}(1+o(1)) \le |T_n^\delta| \le \exp\{(H+\delta)n\}(1+o(1))$

Universal Compression

How do we actually get the compression? Storing a table and using Shannon's encoder for $\mathcal{X}^n$ is unreasonable.

Arithmetic Encoder: extend an ordering of $\mathcal{X}$ onto a lexicographical ordering on $\mathcal{X}^n$. Define

$$F_n(X^n) = \sum_{y^n < x^n} Q_{x^n}(y^n)$$

and associate to each x_n the interval $I_{x^n} = [F_n(x^n), F_n(x^n) + Q_{X^n}(x^n)]$. Take the largest dyadic interval D_{x^n} contained in I_{x^n}, say $[m2^{-k}, (m+1)2^{-k}]$ and use zero padded k-digit binary. The code has length:

$$\log_2 \frac{1}{Q_{X^n}(x^n)} \le \ell(f(x^n)) \le \left\lceil \log_2 \frac{1}{Q_{x^n}(x^n)} \right\rceil + 1$$

Sequential Computation of AE: since

$$F_n(x^n) = F_{n-1}(x^{n-1}) + Q_{X^{n-1}}(x^{n-1}) \sum_{y<x^n} Q_{X_n|X^{n-1}}(y|x^{n-1})$$

if you can easily compute the sequential marginals $Q_{X^{n-1}}$ and $Q_{X_n|X^{n-1}}$, the A.E¿ can be made to be fast.

Also, note *any* Q defines a valid A.E. although it is not clear that the average length will be small (this is a joint learning problem).

Fitingof ("Add-1" construction): simply define Q uniform on all types:

$$Q_{X^n}(X^n) = 1/\binom{n+k-1}{k-1}\binom{n}{n_1, \cdots, n_k}$$

and then we get the **add-1 rule**:

$$Q_{X^n}(X^n) = \left(\prod_{t=1}^n (N(X_t, X_1^{t-1}) + 1)\right) \frac{1}{k(k+1)\cdots(n+k-1)}$$

where $N(a,s) = \#$ of as in a string, and

$$Q_{X_t|X^{t-1}}(\cdot|X^{t-1}) = \frac{N(\cdot; X_1^{t-1}) + 1}{t-1+k}$$

For this Q, we get

$$\mathbb{E}\,\ell = \mathbb{E}\,H(\hat{p}) + O(k \log n)$$

Redundancy-Capacity

Capacity-Redundancy Theorem: We have that the minimax redunancy

$$R_n^* = R_n^*(\Theta) := \min_{Q_{X^n}} \sup_{\theta_0 \in \Theta} D(P_{X^n|\theta=\theta_0} || Q_{X^n})$$

satisfies

$$R_n^* = \sup_\pi \min_{Q_{X^n}} D(P_{X^n|\theta} || Q_{X^n} | \pi) = \sup_\pi I(\theta; X^n)$$

where the redundancy is over all priors $\pi \in \mathcal{P}(\Theta)$. That is, the redundancy is precisely the capacity.

It turns out for general smooth parametric families, **Jeffery's Prior** maximizes the value of general parametric families,

$$\pi(\theta) \propto \sqrt{\det J_F(\theta)}$$

which gives the Redundancy in the form (for smooth priors):

$$\mathcal{R}_n^*(\Theta) = \frac{d}{2} \log \frac{n}{2\pi e} + \log \int_\Theta \sqrt{\det J_F(\theta)} d\theta + o(1)$$

Lempel-Ziv Type Compressors

Kac-Lemma: Suppose $(\cdots, X_{-1}, X_0, X_1, \cdots)$ is stationary and ergodic. Then

$$\mathbb{E}[\tau | X_0 = u] = \frac{1}{\mathbb{P}[X_0 = u]}$$

for $\tau := \inf\{t > 0 : X_{-t} = X_0\}$.

Lempel–Ziv Universal, no model is needed and makes use of the Kac-Lemma. Instead of compressing everthing, compress "strings" of length at most k and look back.

L15-16: Hypothesis Testing

Suppose we have two distributions P and Q on a space $\mathcal{X}$. There are two possibilities summarized by a pair of hypothesis: $H_0 : X \sim P$ called the **null hypothesis** and $H_1 : X \sim Q$ called the **alternative hypothesis**.

A **test** is a random variable $P_{Z|X} : \mathcal{X} \to \{0, 1\}$ giving the probability of rejecting the null.

We denote the two parameters for any test $P_{Z|X}$:

$\alpha := \pi_{0|0} = P[Z = 0]$ (probability of success when H_0 is true)

$\beta := \pi_{0|1} = Q[Z = 0]$ (probability of error when H_1 is true)

In the **Neyman-Pearson formulation** of hypothesis testing, we minimize the type-II error subject to the success probability under the null being at least α. Specifically, given (P, Q), the Neyman-Pearson region consists of all

$$\mathcal{R}(P, Q) = \{(\alpha, \beta) : P_{Z|X} : \mathcal{X} \to \{0, 1\}\} \subseteq [0, 1]^2 \tag{1}$$

The lower boundary is denoted $B_\alpha(P, Q) = \inf_{P[Z=0] \geq \alpha} Q[Z = 0]$. Note $\mathcal{R}(P, Q)$ is closed, convex, contains the diagonal, and is symmetric about the x-axis.

Randomized Tests are Convex Closures of Deterministic Tests: $\mathcal{R}(P, Q) = \text{cl}(\text{co}(\mathcal{R}_{det}(P, Q)))$.

Log-Likelihood Ratio: given the hypothesis testing setting with distributions P and Q, define $T(x) := \log \frac{dP}{dQ}(x)$ to be the LLR. T is a sufficient statistic for binary hypothesis testing.

Converse Statements on $\mathcal{R}(P, Q)$

In the weak converse, we need only the expected value of the LLR (the KL divergence) while the strong-converse, we need the actual value.

Weak Converse (Theorem 14.8): $\forall(\alpha, \beta) \in \mathcal{R}(P, Q)$, we have $d(\alpha||\beta) \leq D(P||Q)$ and $d(\beta||\alpha) \leq D(Q||P)$.

Strong Converse (Theorem 14.10):
$\forall(\alpha, \beta) \in \mathcal{R}(P, Q), \forall\gamma > 0$,

$$\alpha - \gamma\beta \leq P\left[\log \frac{dP}{dQ} > \log \gamma\right], \quad \beta - \frac{\alpha}{\gamma} \leq Q\left[\log \frac{dP}{dQ} < \log \gamma\right]$$

Achievability Bounds on $\mathcal{R}(P, Q)$

L17-19: Channel Coding

Suppose we have a sequence $[M]$ and a channel $P_{Y|X}$. An **M-code** for $P_{Y|X}$ is an encoder/decoder pair (f, g) of randomized functions with the $c_i = f(i)$ being called the **codewords** (the set $\mathcal{C} = \{c_i\}$ the **codebook**) and $D_i = g^{-1}(\{i\})$ is the **decoding region** for i. In general for a message W, we encode it into some encoding X, pass it to the channel $P_{Y|X}$ to get Y, and decode with the decoder to get $\hat{W}$.

Define the following two notions of error for an M-code (f, g):

- The **maximum error probability** $P_e(f, g) := \max_{m \in [M]} \mathbb{P}[\hat{W} \neq m | W = m]$
- The **average error probability** $P_e(f, g) := \mathbb{P}[\hat{W} \neq W]$

We say a code (f, g) is an (M, ϵ)-**code** for $P_{Y|X}$ if $P_e(f, g) \leq \epsilon$ and similar definition for an $(M, \epsilon)_{max}$ code. Ideally we want a large M (transmit many bits) while keeping ϵ small (small error).

Random Coding Bounds

Given two random variables X and Y with joint density $P_{X,Y}$ and densities P_X and P_Y define the **Information Density** (assumign absolute continuity of each of the measures) by:

$$i(x; y) = \log \frac{dP_{X,Y}}{dP_X P_Y}(x, y)$$

The key feature is that $I(X; Y) = \mathbb{E}[i(X; Y)]$. We also get a bound on the **information tails**:

$$\mathbb{P}[i(x, Y) > t] \leq \exp(-t)$$
$$\mathbb{P}[i(\bar{X}, Y) > t] \leq \exp(-t)$$

Now, for a fixed input distribution, maximizing the MAP is equivalent to maximizing the information density. We get the following achievability bounds on (M, ϵ)-codes:

Shannon Achievability Bound: for a channel $P_{Y|X}$ and input distribution P_X, $\tau > 0$, $M \geq 1$ there exists an (M, ϵ) code with

$$\epsilon \leq \mathbb{P}[i(X; Y) \leq \log M + \tau] + \exp(-\tau)$$

The main idea is to use the probabilistic method: for any distribution P_X, draw the codes (encode them) iid using P_X and then analyze the expected error. If you can bound the expected error, you must have at least one such encoding that has error less than or equal to the expectation. A slightly stronger bound is the following:

DT Bound: for a channel $P_{Y|X}$ and input distribution P_X, $M \geq 1$, there exists an (M, ϵ)-code with

$$\epsilon \leq \mathbb{E}\left[\exp\left\{-\left(i(X; Y) - \log \frac{M-1}{2}\right)^+\right\}\right]$$

which also allows you to not optimize over τ, although even in that case the DT bound is stronger.

Maximal Coding

Another set of approaches due to Feinstein uses a greedy approach rather than the probabilistic method. The main theorem is the following:

Feinstein's Lemma: For a channel $P_{Y|X}$ and arbitrary input distribution P_X, for every $\gamma > 0$ and $\epsilon \in (0, 1)$ there is an $(M, \epsilon)_{max}$-code with

$$M \geq \gamma(\epsilon - \mathbb{P}[i(X; Y) < \log \gamma])$$

RCU and Gallagher Bounds

Even stronger bounds on the errors are given by the random-coding-union (RCU) bound and Gallagher:

RCU Bound: for $M \geq 1$, there exists an (M, ϵ)-code such that

$$\epsilon \leq \mathbb{E}[\min\{1, (M-1)\mathbb{P}[i(\bar{X}, Y) \geq i(X, Y) \mid X, Y]\}]$$

where $\bar{X} \perp (X, Y)$.

Gallagher's Bound: Fix a channel $P_{Y|X}$, an arbitrary input distribution P_X and $\rho \in [0, 1]$. There exists an (M, ϵ)-code such that

$$\epsilon \leq M^\rho \, \mathbb{E}\left(\mathbb{E}\left[\exp \frac{i(\bar{X}; Y)}{1+\rho} \mid Y\right]\right)^{1+\rho}$$

Linear Codes

Suppose the input and output space of your channel $\mathcal{X} = \mathcal{Y} = \mathbb{F}_q^n$. A codebook $\mathcal{C}$ of size $M = q^k$ is a **linear code** if $\mathcal{C}$ is a k-dimensional linear subspace of F_q^n.

Linear codes can in fact be provably quite good as stated as follows:

DT Bound for Linear Codes: let $P_{Y|X}$ be an additive noise channel over $\mathbb{F}_q^n$. For all integers $k \geq 1$ there exists a linear code $f : \mathbb{F}_q^k \to \mathbb{F}_q^n$ with error probability

$$P_e = P_{e,max} \leq \mathbb{E}\left[q^{-(n-k\log_q(1/P_Z(Z))^+}\right]$$

Channel Capacity

For the P_e case, for any fixed M, the optimal decoder is trivially the MAP: $g^*(y) = \arg\max_{m\in[M]} \mathbb{P}[Y = y|M = m]$ which can WLOG be chosen to be deterministic. What is the relationship between M and ϵ though? First, we have the following negative result:

Weak Converse: Any (M, ϵ)-code for $P_{Y|X}$ satisfies

$$\log M \leq \frac{\sup_{P_X} I(X;Y) + h(\epsilon)}{1-\epsilon}$$

We also make more rigorous the definition of a **channel**: a sequence of Markov kernels $P_{Y^n|X^n} : \mathcal{A}^n \to \mathcal{B}^n$ is called a **channel** with the length of the input n known as the **blocklength**.

The (n, M, ϵ)-code is thus defined in the same way. The **ϵ-capacity** C_ϵ and the **Shannon capacity** C are defined as

$$C_\epsilon := \liminf_{n\to\infty} \frac{1}{n}\log M^*(n,\epsilon), \qquad C = \lim_{\epsilon\to 0^+} C_\epsilon$$

We have

$$\tau M^*(n, \epsilon(1-\tau)) \leq M^*_{max}(n,\epsilon) \leq M^*(n,\epsilon)$$

and $C_\epsilon^{max} = C_\epsilon$ for $\epsilon > 0$ and $C^{max} = C$.

The **information capacity** of a channel is

$$C^{(I)} = \liminf_{n\to\infty} \frac{1}{n}\sup_{P_{X^n}} I(X^n;Y^n)$$

Bounds on the ϵ-capacity: we have for a stationary, memoryless channel, $\epsilon \in (0,1)$

$$\sup_{P_X} I(X;Y) \leq C_\epsilon \leq \frac{C^{(I)}}{1-\epsilon}$$

We then have the following key theorem:

Shannon's Channel Coding Theorem: For a stationary memoryless channel

$$C = C^{(I)} = \sup_{P_X} I(X;Y)$$

L21-25: Rate-Distortion Theory

We now move onto the topic of compression but for either continuous variables, where we now care about *lossy data compression*. We start with the classical view where the signal lives on a low dimensional space and then move onto the Shannon view.

Scalar and Vector Quantization

Given a random variable $X \in [-A/2, A/2] \subseteq \mathbb{R}$ what type of quantizer $q(X)$ minimizes the **distortion**, ie. $D(N) = \mathbb{E}\,|X - q(X)|^2$?

Uniform Quantization gives $D_U(R) = \frac{A^2 2^{-2R}}{12}$, ie. each additional bit gives you 6dB improvement in SNR.

Non-Uniform Quantization: the main idea is to first take a monotone transform $f(X)$ of the original source and then perform uniform quantization in the transformed space. Then, $q(X) = f^{-1}(q_U(f(X))$ where f is called the **compander**. One popular source is the μ-compounding function:

$$f(X) = \text{sign}(X)\frac{\ln(1+\mu|X|)}{\ln(1+\mu)}$$

High Dimensional Scalar R-bit Quantization: the optimal quantizer density is given by $\lambda^*(x) = \frac{p^{1/3}(x)}{\int p^{1/3}(x)dx}$ so the distortion is (when given R bits):

$$D_{scalar}(R) \approx \frac{1}{12}2^{-2R}\left(\int p^{1/3}(x)dx\right)^3 \tag{2}$$

and can be realized with compander architecture with $f(x) = \int_{-\infty}^x p^{1/3}(t)dt / \int_{-\infty}^\infty p^{1/3}(t)dt$.

High Dimensional Entropy bounded Quantization: if you instead bound $H(q(X))$ we get simply uniform quantization is optimal.

Optimal Quantization and Lloyd's Method: given R bits for reconstruction you can update iteratively the quantization points to the centroids of the Voronoi regions of the points. This converges (potentially) to the optimal quantizer.

Information-Theoretic Formulation

A lossy compressor is an encoder-decoder pair (f, g) along with a **distortion metric** $d : \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R} \cup \{\infty\}$. We define an (n, M, D) code as an encoder-decoder pair that satisfies $\mathbb{E}[d(S^n; g(f(S^n)))] \leq D$. Define the **optimal blocklength** $M^*(n, D) = \min\{M : \exists(n, M, D) \text{ code}\}$ and the **rate distortion** by

$$R(D) := \limsup_{n\to\infty} \frac{1}{n}\log M^*(n, D)$$

General Converse

General Converse (Theorem 24.3): Define

$$\phi_X(D) = \inf_{P_{Y|X}:\mathbb{E}[d(X,Y)\leq D} I(X;Y)$$

Now, if $X \to W \to \hat{X}$ and $\mathbb{E}[d(X,\hat{X})] \leq D$ then

$$\log M \geq \phi_X(D)$$

Define $D_{max} = \inf_{\hat{x}\in\mathcal{X}} \mathbb{E}[d(X,\hat{x})]$. For all $D > D_{max}$ we have $\phi_X(D) = 0$. If $D_0 > D_{max}$ then also $\phi_X(D_{max}) = 0$.

The **information rate-distortion function** for a source $\{S_i\}$ is

$$R^{(I)}(D) = \limsup_{n\to\infty} \phi_{S^n}(D)$$

$$\phi_{S^n}(D) = \inf_{P_{\hat{S}^n|S^n}:\mathbb{E}[d(S^n;\hat{S}^n)]\le D} I(S^n;\hat{S}^n)$$

Then $\forall D, R(D) \ge R^{(I)}(D)$. We also have the following tensorization property:

Single-Letter Tensorization: For stationary memoryless source S_i iid P_S and separable distortion d

Achievability Bounds

Shannon's Rate Distortion Theorem: In the case of a stationary, memoryless channel, $S^n \sim P_S$ iid, with distortion metric d and target distortion D, we have if

1. $d(s^n;\hat{s}^n)$ is non-negative and separable
2. $D > D_0$ where $D_0 = \inf\{D : \phi_S(D) < \infty\}$
3. D_{max} is finite, ie.

$$D_{max} := \inf_{\hat{s}} \mathbb{E}[d(S,\hat{s})] < \infty$$

then

$$R(D) = R^{(I)}(D) = \phi_S(D) = \inf_{P_{\hat{S}|S}:\mathbb{E}[d(S,\hat{S})]\le D} I(S;\hat{S})$$

The main idea is that you can use a random codebook and optimize over that which leads to a constrained mutual information optimization.

Evaluating Rate-Distortion

We have the rate-distortion can be evaluated for several key sources:

Bernoulli: for $S \sim \text{Ber}(p)$ with Hamming distortion $d(S,\hat{S}) = \mathbb{1}[S \ne \hat{S}]$, Then,

$$R(D) = (h(p) - h(D))_+$$

Gaussian: let $S \sim N(0, \sigma^2 I_d)$ and $d(s,\hat{s}) = \|s - \hat{s}\|_2^2$ for $s, \hat{s} \in \mathbb{R}^d$. Then

$$R(D) = \frac{d}{2}\log^+ \frac{d\sigma^2}{D}$$

Lossy Joint Source-Channel Coding

The lossy channel coding can be extended to the case of joint codes $S^k = (S_1, \cdots, S_k)$ taking values on $\mathcal{S}$. A pair (f,g) is called a (k,n,D) **joint source channel code (JSCC)** in an analogous way to a source channel coding when we transmit k symbols now. We get that for stationary memoryless channel, that

$$R_{JSCC}(D) \le \frac{C}{R(D)} \tag{3}$$

Under some mild assumptions (see Assumption 26.1) the inequality becomes equality.