Part I: Information Geometry

1. The Center Piece of Information Theory

**Definition: K-L Divergence (Kullback-Liebler)**

For $P, Q$ both probability distributions on the same finite alphabet $\mathcal{X}$

\[
D(P\|Q) \triangleq \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)}
\]

- Entropy and mutual information are both special cases

\[
\begin{align*}
H(x) &= H(P_x) = H(U) - D(P_x\|U) \\
I(x; y) &= D(P_{xy}\|P_xP_y)
\end{align*}
\]

- **Information inequality**

\[
D(P\|Q) \geq 0, \quad \text{equality iff } P = Q
\]

- **Convexity**

$D(P\|Q)$ is convex in $(P, Q)$

- **Continuity**

**Why K-L divergence matters?**

**Notation:**

- Empirical distribution

\[
\hat{P}_x(a; x^n_1) = \frac{1}{n} \sum_{i=1}^{n} 1(x_i = a)
\]

- Type class:
Empirical Average

Sanov’s Theorem

1. Probability of type class
\[\Pr_{x_1^n \sim \text{i.i.d.} P} (x_1^n \in T_Q) \doteq e^{-nD(Q||P)}\]

2. One type dominates: \(E\) is a subset of distributions on \(\mathcal{X}\),
\[\Pr_{x_1^n \sim \text{i.i.d.} P} (x_1^n \in \bigcup_{Q \in E} T_Q) \doteq e^{-n \cdot D(Q^*||P)}\]
where \(Q^* = \arg \min_{Q \in E} D(Q||P)\)

Quick review of the Channel Coding Theorem

- Transmit a codeword \(x_1^n\), receive \(y_1^n\),

\((x_1^n, y_1^n)\) jointly typical w.r.t. \(P_{xy}\)

- Have some other “incorrect” codewords: \(\tilde{x}_1^n[j]\), \(j = 1, \ldots, M\)

  each \(j:\) \((\tilde{x}_1^n[j], y_1^n) \sim P_x P_y\)

- It’s unlikely for an incorrect codeword to appear typical with the received
\[\Pr_{(\tilde{x}_1^n, y_1^n) \sim P_x P_y} ((\tilde{x}_1^n, y_1^n) \in T_{P_{xy}}) \doteq e^{-n \cdot D(P_{xy}||P_x P_y)} = e^{-n \cdot I(x; y)}\]

- With \(M = e^{nR}\) incorrect codewords, \(R < I(x; y)\), the union bound of the above is still small.
Similar stories in rate distortion, error exponents,

2. Distance and Projection

- K-L divergence is a measure of distance between distributions
- There are many other ways to define divergence
  
  eg. \( f(\cdot) \) convex, continuous, and \( f(1) = 0 \)
  
  \[
  D_f(P || Q) = \sum_x Q(x) \cdot f\left(\frac{P(x)}{Q(x)}\right)
  \]

i-Projection, the binary hypothesis testing story

Consider \( x_1, \ldots, x_n \) i.i.d. distributed from either \( P_0 \) or \( P_1 \).

- Log-likelihood ratio test
  
  \[
  \frac{1}{n} \sum_{i=1}^{n} \log \frac{P_1(x_i)}{P_0(x_i)} \overset{\hat{H}_1}{\geq} \gamma
  \]

- The statistic is an empirical average
  
  \[
  \frac{1}{n} \sum_{i=1}^{n} \log \frac{P_1(x_i)}{P_0(x_i)} = \mathbb{E}_{x \sim P} x^n \left[ \log \frac{P_1}{P_0} (x) \right]
  \]

- The decision region is a subset of type classes
  
  \[
  E \triangleq \{ Q : \mathbb{E}_{x \sim Q} \left[ \log \frac{P_1}{P_0} (x) \right] \geq \gamma \}
  \]

  claim \( \hat{H}_1 \) iff \( x_1^n = x_1^n \in \bigcup_{Q \in E} T_Q \)
• Probability of Error

\[ P(H_0 \to \hat{H}_1) = P \left( x^n_1 \in \bigcup_{Q \in E} T_Q \mid H_0 \right) = e^{-n \cdot \min_{Q \in E} D(Q \parallel P_0)} \]

\[ P(H_1 \to \hat{H}_0) = P \left( x^n_1 \in \bigcup_{Q \in E^c} T_Q \mid H_1 \right) = e^{-n \cdot \min_{Q \in E^c} D(Q \parallel P_0)} \]

• The optimization problem

\[ Q^* = \operatorname{arg\,min}_{Q : \mathbb{E}_Q[f(x)] \geq \gamma} D(Q \parallel P_0) \]

• Dominating error event \( Q_0^* \to 1, \gamma \)
• testing statistic \( f(x) = \log P_1(x) / P_0(x) \)

**Definition: Exponential Family (1-D)**

\[ \mathcal{E}(P_0, f) \triangleq \{ P_t, t \in [0, 1] : P_t(x) = P_0(x) \cdot e^{t \cdot f(x) - \alpha(t)}, \forall x \} \]

• \( P_0 \): a starting point
• \( f(\cdot) \): natural statistic (meaning later)
• \( \alpha(t) \): normalization factor

\[ e^{\alpha(t)} = \sum_x P_0(x) \cdot e^{t \cdot f(x)} = \mathbb{E}_{x \sim P_0} [e^{t \cdot f(x)}] \]

also called the log-moment generation function.

• viewed as “exponential tilting” on \( P_0 \) according to \( f(\cdot) \).

• Empirical average

\[ \eta(t) \triangleq \mathbb{E}_{x \sim P_t} [f(x)] \]

• A number of nice properties
\[
\frac{\partial^2}{\partial t^2} D(P_t || P_0) = \frac{\partial}{\partial t} \eta(t) = \text{var}_{x \sim P_t} [f(x)] = \mathcal{I}_t
\]

- \( \eta(t) \) monotonically increase with \( t \)
- Connection between Fisher information \( \mathcal{I}_t \) and K-L divergence

**Definition: Linear Family**

\[
\mathcal{L}(f, \gamma) \triangleq \{ Q : \mathbb{E}_{x \sim Q} [f(x)] = \gamma \}
\]

**Theorem: Pythagorean**

\[
\forall Q \in \mathcal{L}(f, \gamma) : \\
D(Q || P_0) = D(Q || Q^*) + D(Q^* || P_0)
\]

where \( Q^* \in \mathcal{L}(f, \gamma) \cap \mathcal{E}(P_0, f) \)

- Unique intersection since \( \eta(t) \triangleq \mathbb{E}_{x \sim P_t} [f(x)] \) monotonic increase with \( t \).

\[ Q^* = P_{t^*} \in \mathcal{E}, \text{ with } \mathbb{E}_{x \sim P_{t^*}} [f(x)] = \gamma \]
Corollary: Typical Error Event Occurs on Exponential Family

\[ Q^* = \arg \min_{Q: \mathbb{E}_Q[f(x)] > \gamma} D(Q || P_0) \]

has \( Q^* \in \mathcal{E}(P_0, f) \), with \( f(x) = \log \left( \frac{P_1(x)}{P_0(x)} \right) \) and \( \mathbb{E}_{x \sim Q^*} [f(x)] = \gamma \).

Definition: \( Q^* \) is called the i-projection of \( P_0 \) to the linear family \( \mathcal{L}(f, \gamma) \).

Takeaway message:

- Hypothesis testing is about operations on the empirical distribution, in functional space;
- Each problem has a pair \( P_0, P_1 \), and the exponential family associated;
- Projection of the observed empirical distribution to the exponential family.

m-Projection: the Learning Story

Suppose we observe some data samples \( x_1^n \) with empirical distribution \( \tilde{P}(\cdot; x_1^n) \).

We know that the true model belongs to a parameterized family

\[ \mathcal{P} \triangleq \{ P(\cdot; \theta); \theta \in \mathbb{R} \} \]

often chosen as an exponential family.

Maximum Likelihood estimate of the unknown parameter \( \theta \).
\[ \hat{\theta}_{\text{ML}}(x^n_1) = \arg \max_{\theta} \frac{1}{n} \sum_{i=1}^{n} \log P(x_i; \hat{\theta}) \]

- Usually assume the family to be smooth, \( \frac{\partial}{\partial \theta} P(x, \theta) \) exist, finite
- Can have higher dimensional parameters
- Why do maximum likelihood estimate?
- Distribution matching

\[
\arg \max_{\theta} \frac{1}{n} \sum_{i=1}^{n} \log P(x_i; \hat{\theta}) = \arg \max_{\theta} \mathbb{E}_{x \sim \tilde{P}_x(\cdot; x^n_1)} \left[ \log P(x; \hat{\theta}) \right] \\
= \arg \min_{\theta} \mathbb{E}_{x \sim \tilde{P}(\cdot; x^n_1)} \left[ \log \frac{\tilde{P}(x; x^n_1)}{P(x; \hat{\theta})} \right] \\
= \arg \min_{\theta} D \left( \tilde{P}(\cdot; x^n_1) || P(\cdot; \hat{\theta}) \right)
\]

**Definition:** \( P(\cdot; \hat{\theta}_{\text{ML}}) \) is called the m-projection of \( \tilde{P} \) to the model family \( \mathcal{P} \).

**Corollary:** if \( \mathcal{P} \) is an exponential family, \( \mathcal{P} = \mathcal{E}(P_0, f) \), and suppose 
\( \mathbb{E}_{x \sim \tilde{P}}[f(x)] = \gamma \), then
\[
P(\cdot; \hat{\theta}_{\text{ML}}) \in \mathcal{P} \cap \mathcal{L}(f, \gamma)
\]
Takeaway message:

- A model family is a manifold/plane in the space of distributions;
- Learning is also a projection, from the observed empirical distribution to the model family.

3. The Geometry of Information Theory and Learning

- A number of information theory results presented as geometric stories:
  - Error Exponent
  - Csiszar's book

What is difficult about this?

- The geometry is complex.

Fisher information

\[
\frac{\partial^2}{\partial t^2} D(P_t || P_0) = \frac{\partial}{\partial t} \eta(t) = \text{var}_{x \sim P_t} [f(x)] = \mathcal{I}_t
\]
but $I_t \geq 0$ can be an arbitrary function of $t$.

So $D(P_t || P_0)$ is a convex function of $t$, but not clear how convex.

- If you have learned Cramer-Rao bound ...

- What we need is a lot more.
  - Broadcast channels: $P_{y|x}, P_{z|x}$. Even if $I(x; y) > I(x; z)$, doesn’t mean the channel $x \rightarrow z$ is degraded.

**Dependence is not a single dimensional concept.**

- Mismatched detection, universal detection: what happens if we didn’t use the right $f(x) = \log \frac{P_1(x)}{P_0(x)}$ to make decision, but used a different $f'(\cdot)$?

**How bad are imperfect statistic models?**

- Increasing the dimensionality of $\mathcal{E}$, what collection/sequence of statistics

$$P_x(x; \theta) = P_0(x) \cdot \exp \left[ \sum_{i=1}^{k} \theta_i \cdot f_i(x) - \alpha(\theta) \right]$$

**What statistic is more valuable in learning?**

- What happens with each iteration and each mini-batch of samples?

**Evolution and convergence of learned models in functional space.**

- From input/output neural networks to Transfer Learning, Multi-Modal Learning

**Network information theory and more complex learning tasks.**

- There are often too many distributions to worry about
  - The ground truth
  - The parameterized family
  - The empiricals
  - The current model and the updates
  - Restrictions, side information, loss
  - Tuning of design parameters
- **Basically**: we cannot write it very clean for 1-D problems with 2 distributions, but we need some analysis for multi-dimensional problems with many distributions.

**What is Geometry and Why Geometry?**

- Distance $\longrightarrow$ inner product, projection, basis, coordinates (Hilbert Space for distributions)

- Space of functions and Space of distributions.
Part II: The Local Geometry

Notation
- True model $P$, Observed empirical distribution $\tilde{P}$, Estimated model $\hat{P}$.

Color code
- Definition
- Theorem
- Key Problem

4. Fisher Information Metric

**Definition: Fisher Information**

For a parameterized family of distributions $\mathcal{P} = \{P_x(\cdot; \theta), \theta \in \mathbb{R}^k\}$, the Fisher information matrix $\mathcal{I}(\theta) \in \mathbb{R}^{k \times k}$ is

$$[\mathcal{I}(\theta)]_{ij} \triangleq \mathbb{E}_{x \sim P(x; \theta)} \left[ \left( \frac{\partial}{\partial \theta_i} \log P_x(x; \theta) \right) \left( \frac{\partial}{\partial \theta_j} \log P_x(x; \theta) \right) \right]$$

- Can be shown to be Positive Semi-Definite
- Can be shown to be a valid metric
- Has a lot of good applications

**Understand the Definition**

- Every distribution involved is close to $P_x(\cdot; \theta)$
  - Reference distribution:
    $$R_x \triangleq P_x(\cdot; \theta = 0)$$
  - Think of all entries in $\theta$ are restricted to be within $(\epsilon, +\epsilon)$
  - Each $\theta_i$ corresponds to a curve
    $$\theta = [0, \ldots, 0, \theta_i, 0, \ldots, 0] \quad \rightarrow \quad P_x(x; \theta) \triangleq R_x(x) \cdot (1 + \theta_i \cdot f_i(x)), \ x \in \mathcal{X}$$
Log likelihood ratio

\[ \log P_x(x; \theta) - \log P_x(x; 0) = \log(1 + \theta \cdot f_i(x)) = \theta_i \cdot f_i(x) + O(\epsilon^2), \forall x \in \mathcal{X} \]

Perturbation accumulates

\[ \theta = [\theta_1, \ldots, \theta_k] \longrightarrow P_x(x; \theta) \triangleq R_x(x) \cdot \left(1 + \sum_i \theta_i \cdot f_i(x) + O(\epsilon^2)\right), x \in \mathcal{X} \]

Locally viewed as exponential family with natural statistic \( f_i(\cdot) \).

Fisher information:

\[ [\mathcal{I}(\theta = 0)]_{ij} = \mathbb{E}_{x \sim R_x} [f_i(x)f_j(x)], \quad \forall i, j \]

- obviously positive semi-definite
- obviously a valid inner product:

\[ \langle f_i, f_j \rangle \triangleq \mathbb{E}_{x \sim R_x} [f_i(x)f_j(x)] \]

- has to stay on the simplex:

\[ \mathbb{E}_{x \sim R_x} [f_i(x)] = 0, \forall i \]

Wait! now we are talking about both distributions and functions.

For given two distributions \( P_1, P_2 \in \mathcal{N}_s(R_x) \), define \( f_1, f_2 \) by

\[ P_i(x) = R_x(x) \cdot (1 + f_i(x)), \quad i = 1, 2 \]
5. Information Vector

**Definition:**
- Fix a finite alphabet: \( \mathcal{X} \),
- Fix a reference distribution: \( R \) on \( \mathcal{X} \),

1. For any function \( f : \mathcal{X} \mapsto \mathbb{R} \), with \( \mathbb{E}_{x \sim R}[f(x)] = 0 \)
   The information vector for \( f \) is written as \( \phi(f) \in \mathbb{R}^\mathcal{X} \), with
   \[
   \phi(f) \triangleq \left[ \sqrt{R(x)} \cdot f(x), \; x \in \mathcal{X} \right]^T
   \]
2. For any distribution \( P \in \mathcal{N}_\epsilon(R) \),
   The information vector for \( P \) is written as \( \phi(P) \in \mathbb{R}^\mathcal{X} \), with
   \[
   \phi(P)(x) \triangleq \sqrt{R(x)} \cdot \left( \frac{P(x)}{R(x)} - 1 \right) = \frac{1}{\sqrt{R(x)}} \cdot (P(x) - R(x)), \; x \in \mathcal{X}
   \]

**First Properties:**
1. Inner product and Covariance:
   \[
   \langle \phi(f_1), \phi(f_2) \rangle = \mathbb{E}_{x \sim R}[f_1(x)f_2(x)]
   \]
2. Norm and variance:
   \[
   \| \phi(f) \|^2 = \text{var}_{x \sim R}[f(x)]
   \]
3. Orthogonal functions iff uncorrelated (w.r.t. \( R \))

**K-L Divergence**
For \( P, Q \in \mathcal{N}_\epsilon(R) \),
\[
D(P \parallel Q) = \frac{1}{2} \| \phi(P) - \phi(Q) \|^2 + o(\epsilon^2)
\]

\[
D(Q \parallel P) = \frac{1}{2} \| \phi(P) - \phi(Q) \|^2 + o(\epsilon^2)
\]

**Proof:**

This is our first local geometric result, let's start with notations.

Write

\[
f \leftrightarrow P \leftrightarrow \phi(P) : P(x) = R(x) \cdot (1 + f(x)) = R(x) \cdot \left(1 + \frac{\phi(P)(x)}{\sqrt{R(x)}}\right) = R(x) + \sqrt{R(x)} \cdot \phi(P)(x)
\]

\[
g \leftrightarrow Q \leftrightarrow \phi(Q) : Q(x) = R(x) \cdot (1 + g(x)) = R(x) \cdot \left(1 + \frac{\phi(Q)(x)}{\sqrt{R(x)}}\right) = R(x) + \sqrt{R(x)} \cdot \phi(Q)(x)
\]

Now we have

\[
D(P \parallel Q) = \sum_x P(x) \cdot \log \frac{P(x)}{Q(x)} = \sum_x P(x) \cdot \left(\log \frac{P(x)}{R(x)} - \log \frac{Q(x)}{R(x)}\right)
\]

\[
= \sum_x P(x) \cdot \left[\log (1 + f(x)) - \log (1 + g(x))\right]
\]

\[
= \sum_x \left[R(x) + R(x) \cdot f(x)\right] \cdot \left[f(x) - \frac{1}{2} f^2(x) - g(x) + \frac{1}{2} g^2(x) + O(\epsilon^3)\right]
\]

\[
= \sum_x R(x) (f(x) - g(x))
\]

\[
+ \sum_x R(x) \left(- \frac{1}{2} f^2(x) + \frac{1}{2} g^2(x) + f(x) \cdot (f(x) - g(x))\right) + o(\epsilon^2)
\]

\[
= \frac{1}{2} \mathbb{E}_{x \sim R} [(f(x) - g(x))^2] + o(\epsilon^2)
\]

**CLT and Asymptotic Normality**

Recall Large Deviations / Sanov Theorem

\[
\mathbb{P}_{x_1 \sim \text{i.i.d. } P} \left( x_1^n \in T_Q \right) = e^{-nD(Q \parallel P)}
\]

\[
= e^{-n \frac{1}{2} \| \phi(Q) - \phi(P) \|^2 + o(\epsilon^2)}
\]

Part II: The Local Geometry
• The empirical distribution $\hat{P}_x(\cdot; x^n_i) = Q$ is random, corresponding $\phi^{(g)}$ is also random
• Gaussian distributed around the ensemble distribution $P \leftrightarrow \phi^{(P)}$
• With approximate a Gaussian distribution, white, with variance $1/n$ per dimension.

Local parameter estimate: empirical average $\propto$ estimate $\hat{\theta}_{ML}$
• CLT: if $f(x)$ has zero-mean and unit variance w.r.t. $R_k\frac{1}{\sqrt{n}} \sum_i f(x_i) \rightarrow N(0, 1)$
• Asymptotic efficiency of ML estimate:

$$\sqrt{n} \cdot (\hat{\theta}_{ML} - \theta) \rightarrow N(0, \frac{1}{I(\theta)})$$

• The business between a finite alphabet and a continuous alphabet.

6. Example: Akaike Information Criterion

Consider a sequence of nested parameterized families $P_1 \subset P_2 \subset \ldots \subset P_m$
• with increasing dimensionality of parameters
  $$P_k = \{P_x(\cdot; \theta^k), \theta^k \in \mathbb{R}^k\}, \quad k = 1, \ldots, m$$
• Observe $x^n_1 = x^n_1$, solve for each family
  $$\hat{P}^k_{ML} = \arg \min_{\hat{P} \in P_k} D(\hat{P}(\cdot; x^n_i) || \hat{P})$$
• Larger $k$, better matching,
• Which $k$ to choose? How to penalize bigger families? Avoid over-fitting.

Akaike's observation:
We really want to minimize
but observe $D(\hat{P}\|\bar{P})$.

Locally:
1. All ML estimates are just projections

$$\hat{P}_{\text{ML}}^k \leftrightarrow \pi_k(\tilde{\phi})$$

2. What to compare?

$$\|\pi_2(\tilde{\phi}) - \tilde{\phi}\| < \|\pi_1(\tilde{\phi}) - \tilde{\phi}\|$$

but how about

$$\|\pi_2(\tilde{\phi}) - \phi\| \geq \|\pi_1(\tilde{\phi}) - \phi\|$$

3. $\tilde{\phi} - \phi$ is asymptotically normal with variance $1/n$ per dimension.

Given $\|\tilde{\phi} - \pi_k(\tilde{\phi})\|^2$, need to

- subtract the average power of $(\tilde{\phi} - \phi) \perp \mathcal{P}$, $\sim \frac{1}{n}(|\mathcal{X}| - k)$
- add the average power of $(\tilde{\phi} - \phi) \parallel \mathcal{P}$, $\sim \frac{k}{n}$

$$\min_k \|\tilde{\phi} - \pi_k(\tilde{\phi})\|^2 + \frac{k}{n} - \frac{|\mathcal{X}| - k}{n} \rightarrow \min_k D(\hat{P}\|\hat{P}_{\text{ML}}^k) + \frac{k}{n}$$

7. Projections and Inner Products

What does the direction of information vectors represent?

- Consider $x_1, \ldots, x_n \sim \text{i.i.d.} \mathcal{R}_x$,
- but we observe empirical distribution $\hat{P}$

$$\tilde{\phi}(x) = \frac{P(x) - R(x)}{\sqrt{R(x)}}, \forall x$$

- We would like to evaluate the empirical average of a function $f : \mathcal{X} \mapsto \mathbb{R}$
- w.l.o.g. assume $\mathbb{E}_{x \sim \mathcal{R}}[f(x)] = 0$
- Information vector $\psi \leftrightarrow f$, with $\psi(x) = \sqrt{R(x)} \cdot f(x), \forall x$
The empirical average

\[
\frac{1}{n} \sum_{i=1}^{n} f(x_i) = \mathbb{E}_{X \sim P}[f(x)] = \sum_{x} \hat{P}(x) \cdot f(x) \\
= \sum_{x} \left( R(x) + \sqrt{R(x)} \cdot \hat{\phi}(x) \right) \cdot \frac{\psi(x)}{\sqrt{R(x)}} \\
= \langle \hat{\phi}, \psi \rangle
\]

**Back to Binary Hypothesis Testing**

- Recall linear decision region

  \[
  \frac{1}{n} \sum_{i=1}^{n} \log f(x_i) = \mathbb{E}_{X \sim \hat{P}[f(x)]} \underset{H_0}{\overset{H_1}{\gtrless}} \gamma
  \]

  and the optimal

  \[
  f(x) = \log \frac{P_1(x)}{P_0(x)}, \quad \forall x
  \]

- A binary query corresponds to a vector

  \[
  \psi \triangleq \phi(P_1) - \phi(P_2) \\
  = \frac{P_1(x) - R(x)) - (P_0(x) - R(x))}{\sqrt{R(x)}} \\
  = \sqrt{R(x)} \cdot \left[ \left( \frac{P_1(x)}{R(x)} - 1 \right) - \left( \frac{P_0(x)}{R(x)} - 1 \right) \right] \\
  \approx \sqrt{R(x)} \cdot \left[ \log \left( \frac{P_1(x)}{R(x)} \right) - \log \left( \frac{P_0(x)}{R(x)} \right) \right] \\
  = \sqrt{R(x)} \cdot f(x) = \phi(f)
  \]

  for the log-likelihood function

  - LLR = project observed empirical distribution \( \hat{\phi} \) on \( \psi \)
  - length = \( D(P_1 \parallel P_0) \), maximum one-sided error exponent

  - What if we use a different statistic \( f' \neq \log \frac{P_1}{P_0} \)?
Maximum one-sided error exponent reduced by factor of
\[ \left| \cos \left( \angle(\phi(f), \phi(f')) \right) \right|^2 \]

- Measures how much is \( f'(x) \) useful in answering a question about \( f \)!!!

8. Information Vector for Joint Distributions, CDM

- \( P_{xy} \) with reference \( R_{xy} \)
- Choose \( R_{xy} = P_x \cdot P_y \), independent with the same marginals

**Definition:** Canonical Dependence Matrix (CDM) \( B \in \mathbb{R}^{X \times Y} \)

\[ B(x, y) \triangleq \frac{P_{xy}(x, y) - P_x(x)P_y(y)}{\sqrt{P_x(x)P_y(y)}}, \quad (x, y) \in X \times Y \]

- Inherited property
  \[ \frac{1}{2} \| B \|^2 \approx D(P_{xy}||P_xP_y) = I(x; y) \]

- \( x, y \) have symmetric positions
- Describes how the two random variables are dependent
- Can be viewed as a channel
- $W = P_{y|x}$ defines a channel
- By definition, if input is $R_x = P_x$, the output is $R_y = P_y$
- If we change input to be $Q_x \leftrightarrow \phi \in \mathbb{R}^X$
  
  $$Q_x(x) = R_x(x) + \sqrt{R_x(x)} \cdot \phi(x), \quad x \in X$$

  The output would be $Q_y \leftrightarrow \psi \in \mathbb{R}^Y,$
  
  $$Q_y(y) = \sum_x P_{y|x} (y|x) \cdot Q_x(x)$$
  
  $$= \sum_x P_{y|x} (y|x) \cdot R_x(x) \cdot \left(1 + \frac{\phi(x)}{\sqrt{R_x(x)}}\right)$$
  
  $$= R_y(y) + \sum_x P_{xy} (x, y) \cdot \frac{\phi(x)}{\sqrt{R_x(x)}}$$
  
  $$= R_y(y) + \sum_x (P_{xy} (x, y) - P_x(x)P_y(y)) \cdot \frac{\phi(x)}{\sqrt{R_x(x)}}$$
  
  $$= R_y(y) + \sqrt{R_y(y)} \cdot \left(\sum_x \frac{P_{xy} (x, y) - P_x(x)P_y(y)}{\sqrt{R_x(x)} \sqrt{R_y(y)}} \cdot \phi(x)\right)$$

**Theorem:** $B$-matrix as a map

$$\psi = B \cdot \phi$$

**Map of functions**
Suppose
\[ \phi \leftrightarrow f : \sqrt{R_x(x)} \cdot f(x) = \phi(x), \quad x \in \mathcal{X} \]
\[ \psi \leftrightarrow g : \sqrt{R_y(y)} \cdot g(y) = \psi(y), \quad y \in \mathcal{Y} \]

Define \( \psi = B \cdot \phi \), what function operation is this?

\[
g(y) = \frac{1}{\sqrt{R_y(y)}} \cdot \psi(y) = \frac{1}{\sqrt{R_y(y)}} \cdot \left( \sum_x B(x, y) \cdot \phi(x) \right) \\
= \frac{1}{\sqrt{R_y(y)}} \cdot \left( \sum_x \frac{P_{xy}(x, y) - P_x(x)P_y(y)}{\sqrt{P_x(x)P_y(y)}} \cdot \phi(x) \right) \\
= \sum_x \frac{P_{xy}(x, y)}{P_y(y)} \cdot \frac{\phi(x)}{\sqrt{R_x(x)}} \\
= \mathbb{E}[f(x)|y = y], \quad \forall y
\]

Similarly \( f(x) = \mathbb{E}[g(y)|x = x], \quad \forall x \)

\( B \) matrix is the conditional expectation operator.
Part III: Machine Learning

9. Example: Conjugator Prior Family

**Definition:** Given an observation model $P_{y|x}$, a parameterized family of prior distribution

$$P = \{ P_x(\cdot; \theta), \theta \in \mathbb{R} \}$$

is called the conjugate prior family if for any value of $y$, $P_{x|y}(\cdot|y) \in P$.

- Update knowledge turned into update parameters
- Bernoulli/Beta; Categorical/Dirichlet, Poisson/Gamma, Normal (fix $\sigma^2$)/ Normal

Diaconis, Ylvisker (1979)

Conjugate Priors for Exponential Families
Let $X$ be a random vector distributed according to an exponential family with natural parameter $\theta \in \Theta$. We characterize conjugate prior measures on $\Theta$ through the property of


If the observation model is an exponential family

$$P_{y|x}(y|x) = \exp(x \cdot t(y) - \alpha(x))$$

then the conjugate prior $P_x(\cdot; \theta)$ must satisfy that for all $\theta$, evaluated w.r.t. $P_x(\cdot; \theta) \cdot P_y$,

$$\mathbb{E}\mathbb{E}[t(y)|x] = a \cdot t(y) + b$$

for some constants $a, b$.

The geometric view:
1. Observe a sequence $\tilde{y}_1, \ldots, \tilde{y}_n$ with empirical distribution $\tilde{P}_y = \tilde{Q}_y \leftrightarrow \psi$

2. Symmetric story, the posterior

$$P_{x|y^n} (\cdot | \tilde{y}_1^n) = Q_x \leftrightarrow \phi = B^T \cdot \psi$$

3. Conjugate prior: regardless of $\psi$, the posterior is always in a 1-D family: $\phi$ remains in the same direction.

$B$ is a rank-1 matrix: $B = \sigma \cdot \psi \cdot \phi^T$,

$$B \cdot B^T \cdot \psi = \sigma^2 \cdot \psi \quad \Leftrightarrow \quad \mathbb{E}[\mathbb{E}[t(y)|x]|y] = a \cdot t(y)$$

10. The multi-dimensional nature of dependence

$$B = \sum_i \sigma_i \cdot u_i \cdot v_i^T$$

- Dependence over multiple modes

$$I(x; y) \approx \frac{1}{2} \cdot \sum \sigma_i^2$$

- **Example**: broadcast channel,

  - $I(x; y) > I(x; z)$ does not mean we cannot transmit a private message $x \rightarrow z$ that is not decodable by $y$.

  - More capable (El-Gammal 79'): $B_{xy}$ dominates $B_{xz}$ in every mode.

    $$\|B_{xy} \cdot \phi_x\|^2 \geq \|B_{xz} \cdot \phi_x\|^2, \quad \forall \phi_x$$

- **Example**: Strong DPI

  - All singular values of $B$ are less than or equal to 1.\n
    $$\|B_{xy} \cdot \phi_x\|^2 \leq \|\phi_x\|^2, \quad \forall \phi_x \quad \Leftrightarrow \quad D(P_y||Q_y) \leq D(P_x||Q_x),$$

  - But the contraction is really not a 1-D scaling issue.

  - Literature of slightly different formulations of SDPI.
• **Example:** Hermite Polynomial for Additive Gaussian Noise Channel

![Graph showing Hermite Polynomials](image)

### 11. Renyi Correlation, CCA

**Definition:** Hirschfeld-Gebelein-Renyi Maximal correlation:

Given $P_{xy}$:

$$
\rho \triangleq \max_{f,g} \mathbb{E}_{x,y \sim P_{xy}} [f(x) \cdot g(y)]
$$

where $f, g$ satisfies $\mathbb{E}[f(x)] = \mathbb{E}[g(y)] = 0$, $\mathbb{E}[f^2(x)] = \mathbb{E}[g^2(y)] = 1$

- Defined as a measure of level of dependence 1959.
- Generalizes to multiple pairs of functions $f_1, \ldots, f_k; g_1, \ldots, g_k$.
- Canonical Dependence Analysis, Correspondence Analysis.

### 12. Operations in Neural Networks
Classification $y \in \{1, \ldots, |\mathcal{Y}|\}$.

- Last layer input: $f_1(x), \ldots, f_k(x)$,
- Last layer weights: $g_i(y), i = 1, \ldots, k; y \in \mathcal{Y}$,
- Softmax activation:
  \[
  \hat{P}^{(f,g)}_{y|x}(y|x) = \frac{\exp \left[ \sum_{i=1}^{k} f_i(x) \cdot g_i(y) + b(y) \right]}{\sum_{y'} \exp \left[ \sum_{i=1}^{k} f_i(x) \cdot g_i(y') + b(y') \right]}, \ y \in \mathcal{Y}
  \]

- Cross-Entropy Loss, ML for discriminative model.

$$
\arg \min_{f,g} D\left(\hat{P}_x \cdot \hat{P}_{y|x} \parallel \hat{P}_x \cdot \hat{P}^{(f,g)}_{y|x}\right)
$$

```python
model = Sequential()
model.add("")
model.add(Dense(yCard, activation='softmax', input_dim=k))
sgd = SGD(4, decay=1e-2, momentum=0.9, nesterov=True)
model.compile(loss='categorical_crossentropy', optimizer=sgd)
```

- In the local setup
  1. Reference
\[ R_x(x) = \hat{P}_x(x), \quad \forall x \]
\[ R_y(y) \propto e^{b(y)}, \quad \forall y \]
\[ \hat{P}_{y|x}^{(f,g)}(y|x) = R_y(y) \cdot \frac{\exp \left[ \sum_{i=1}^{k} f_i(x) \cdot g_i(y) \right]}{\sum_{y'} \exp \left[ \sum_{i=1}^{k} f_i(x) \cdot g_i(y') \right]}, \quad y \in Y \]

2. Learned model
\[ \hat{P}_{y|x}^{(f,g)} \rightarrow R_y, \hat{B}^{(f,g)} : \]
\[ \hat{B}^{(f,g)}(x, y) = \sqrt{R_x(x)R_y(y)} \cdot \left( \sum_{i=1}^{k} f_i(x) \cdot g_i(y) \right) \quad \forall x, y \]

3. Optimization
\[ \arg \min_{f,g} \| \hat{B} - \hat{B}^{(f,g)} \|^2 \]

4. Solution: SVD

5. How was this numerically solved?

BackProp:
- Fix \( f \): \( g(y) \leftarrow \mathbb{E}[f(x)|y = y], \quad \forall y \), equivalent to \( \hat{\psi}^{(g)} \leftarrow B \cdot \hat{\phi}^{(f)} \)
- Fix \( g \): \( f(x) \leftarrow \mathbb{E}[g(y)|x = x], \quad \forall x \), equivalent to \( \hat{\phi}^{(f)} \leftarrow B^T \cdot \hat{\psi}^{(g)} \)
13. What is this good for?

- It is good to know that NNs are SVD solvers;
- H-score implementation

\[
H(f, g) = \| \hat{B} - \hat{B}^{(f,g)} \|^2 \triangleq \mathbb{E}_{x,y \sim \hat{p}_{xy}} \left[ f^T(x) \cdot g(y) \right] - \frac{1}{2} \text{trace} \left( \text{cov}(f) \cdot \text{cov}(g) \right)
\]

\[
H(f) = H(f, g^*) \triangleq \mathbb{E}_{y \sim p_y} \left[ \mathbb{E}[f(x)|y = y]^T \cdot \text{cov}(f)^{-1} \cdot \mathbb{E}[f(x)|y = y] \right]
\]

- Allows aggressive dimension reduction
- Direct operation on the feature functions

- Choice of reference distribution \( R_{xy} \) and iterative algorithms, convergence analysis.
- Knowledge subspace: \( \text{span}(f_1, \ldots, f_k) \). Interpretation and evaluation of learning quality.
- Multi-variate, multi-modal, multi-task problems.