Part I: Information Geometry

Color code

Definition

Theorem

Key Problem

1. The Center Piece of Information Theory

Definition: K-L Divergence (Kullback-Liebler)

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

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

- Entropy and mutual information are both special cases

$$\begin{aligned} H(\mathsf{x}) &= H(P_{\mathsf{x}}) = H(U) - D(P_{\mathsf{x}}||U) \\ I(\mathsf{x};\mathsf{y}) &= D(P_{\mathsf{xy}}||P_{\mathsf{x}}P_{\mathsf{y}}) \end{aligned}$$

- **Information inequality**

  $D(P||Q) \geq 0, \qquad \text{equality iff } P = Q$

- **Convexity**

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

- **Continuity**

Why K-L divergence matters?

Notation:

- Empirical distribution

  $\check{P}_{\mathsf{x}}(a; x_1^n) = \frac{1}{n} \sum_{i=1}^{n} 1_{(x_i == a)}$

- Type class:

$T_Q \triangleq \{x_1^n \in \mathcal{X}^n : \check{P}_{\mathsf{x}}(\cdot; x_1^n) = Q(\cdot)\}$

- Empirical Average

  $\frac{1}{n}\sum_{i=1}^n g(x_i) = \mathbb{E}_{\mathsf{x} \sim \check{P}_{\mathsf{x}}(\cdot; x_1^n)}[g(\mathsf{x})]$

Sanov's Theorem

1. Probability of type class

   $\mathbb{P}_{\mathsf{x}_1^n \sim \text{i.i.d.} P}\ (\mathsf{x}_1^n \in T_Q\ ) \doteq e^{-nD(Q||P)}$

2. One type dominates: E is a subset of distributions on $\mathcal{X}$,

   $\mathbb{P}_{\mathsf{x}_1^n \sim \text{i.i.d.} P}\ (\mathsf{x}_1^n \in \cup_{Q \in E} T_Q\ ) \doteq e^{-n \cdot D(Q^*||P)}$

   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 : $(\tilde{x}_1^n[j], y_1^n) \sim P_{\mathsf{x}} P_{\mathsf{y}}$

- It's unlikely for an incorrect codeword to appear typical with the received

  $\mathbb{P}_{(\tilde{x}_1^n, y_1^n) \sim P_{\mathsf{x}} P_{\mathsf{y}}}((\tilde{x}_1^n, y_1^n) \in T_{P_{\mathsf{xy}}}\ ) \doteq e^{-n \cdot D(P_{\mathsf{xy}}||P_{\mathsf{x}} P_{\mathsf{y}})} = e^{-n \cdot I(\mathsf{x};\mathsf{y})}$

- With $M = e^{nR}$ incorrect codewords, $R < I(\mathsf{x}; \mathsf{y})$, 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 $f(1) = 0$

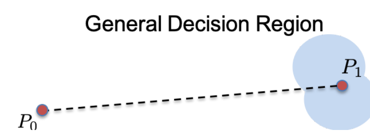
  $$D_f(P||Q) = \sum_x Q(x) \cdot f\left(\frac{P(x)}{Q(x)}\right)$$

i-Projection, the binary hypothesis testing story

Consider $\mathsf{x}_1, \ldots, \mathsf{x}_n$ i.i.d. distributed from either P_0 or P_1.

- Log-likelihood ratio test

  $$\frac{1}{n}\sum_{i=1}^{n} \log \frac{P_1(x_i)}{P_0(x_i)} \underset{\hat{H}_0}{\overset{\hat{H}_1}{\gtrless}} \gamma$$



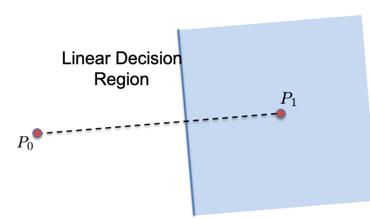
- The statistic is an empirical average

  $$\frac{1}{n}\sum_{i=1}^{n} \log \frac{P_1(x_i)}{P_0(x_i)} = \mathbb{E}_{\mathsf{x}\sim\check{P}(\cdot;x_1^n)}\left[\log \frac{P_1}{P_0}(\mathsf{x})\right]$$

- The decision region is a subset of *type classes*

  $$E \triangleq \{Q : \mathbb{E}_{\mathsf{x}\sim Q}\left[\log \frac{P_1}{P_0}(\mathsf{x})\right] \geq \gamma\}$$

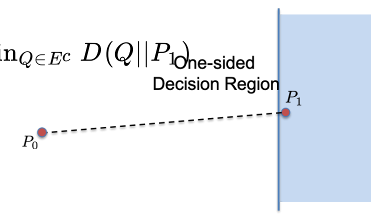
  claim $\hat{H}_1$ iff $\mathsf{x}_1^n = x_1^n \in \bigcup_{Q\in E} T_Q$



- Probability of Error

$$\mathbb{P}(H_0 \to \hat{H}_1) = \mathbb{P}\left(\mathsf{x}_1^n \in \bigcup_{Q \in E} T_Q \middle| H_0 \right) \doteq e^{-n \cdot \min_{Q \in E} D(Q||P_0)}$$

$$\mathbb{P}(H_1 \to \hat{H}_0) = \mathbb{P}\left(\mathsf{x}_1^n \in \bigcup_{Q \in E^c} T_Q \middle| H_1 \right) \doteq e^{-n \cdot \min_{Q \in E^c} D(Q||P_1)}$$



- The optimization problem

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

  - Dominating error event $Q^*_{0 \to 1, \gamma}$
  - testing statistic $f(x) = \log P_1(x)/P_0(x)$

Definition: Exponential Family (1-D)

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

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

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

  also called the log-moment generation function.
- viewed as "exponential tilting" on P_0 according to $f(\cdot)$.

- Empirical average

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

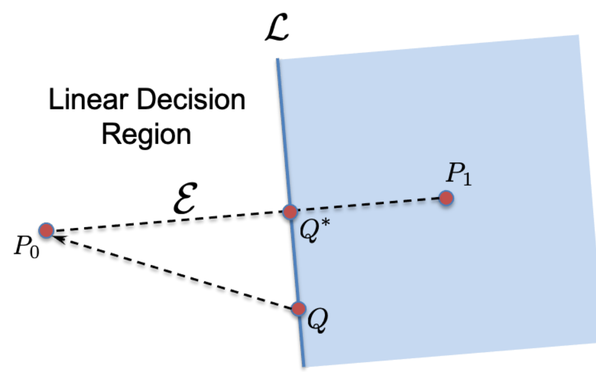
- A number of nice properties

$$\frac{\partial^2}{\partial t^2} D(P_t || P_0) = \frac{\partial}{\partial t}\eta(t) = \mathrm{var}_{\mathsf{x}\sim P_t}[f(\mathsf{x})] = \mathcal{I}_t$$

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

Definition: Linear Family

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



Theorem: Pythagorean

$\forall Q \in \mathcal{L}(f,\gamma)$:

$$\boxed{D(Q||P_0) = D(Q||Q^*) + D(Q^*||P_0)}$$

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

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

  $Q^* = P_{t^*} \in \mathcal{E}$, with $\mathbb{E}_{\mathsf{x}\sim P_{t^*}}[f(\mathsf{x})] = \gamma$:

$$
\begin{aligned}
D(Q||Q^*) &= \mathbb{E}_{\mathsf{x}\sim Q}\left[\log\frac{Q(\mathsf{x})}{Q^*(\mathsf{x})}\right] = \mathbb{E}_{\mathsf{x}\sim Q}\left[\log\frac{Q(\mathsf{x})}{P_{t^*}(\mathsf{x})}\right] \\
&= \mathbb{E}_{\mathsf{x}\sim Q}\left[\log\frac{Q(\mathsf{x})}{P_0(\mathsf{x})\cdot e^{t^*\cdot f(\mathsf{x})-\alpha(t^*)}}\right] \\
&= \mathbb{E}_{\mathsf{x}\sim Q}\left[\log\frac{Q(\mathsf{x})}{P_0(\mathsf{x})}\right] - E_{\mathsf{x}\sim Q}[t^* f(\mathsf{x}) - \alpha^*(t)] \\
&= \mathbb{E}_{\mathsf{x}\sim Q}\left[\log\frac{Q(\mathsf{x})}{P_0(\mathsf{x})}\right] - E_{\mathsf{x}\sim Q^*}[t^* f(\mathsf{x}) - \alpha^*(t)] \\
&= \mathbb{E}_{\mathsf{x}\sim Q}\left[\log\frac{Q(\mathsf{x})}{P_0(\mathsf{x})}\right] - E_{\mathsf{x}\sim Q^*}\left[\log\frac{P_0(\mathsf{x})\cdot e^{t^* f(\mathsf{x})-\alpha^*(t)}}{P_0(\mathsf{x})}\right] \\
&= D(Q||P_0) - D(Q^*||P_0)
\end{aligned}
$$

C

Corollary: Typical Error Event Occurs on Exponential Family

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

has $Q^* \in \mathcal{E}(P_0, f)$, with $f(x) = \log\left(\frac{P_1(x)}{p_0(x)}\right)$ and $\mathbb{E}_{\mathsf{x}\sim Q^*}[f(\mathsf{x})] = \gamma$.

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 $\check{P}(\cdot; x_1^n)$. 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 θ.

$\hat{\theta}_{\mathsf{ML}}(x_1^n) = \arg\max_{\hat{\theta}} \frac{1}{n}\sum_{i=1}^n \log P(x_i;\hat{\theta})$

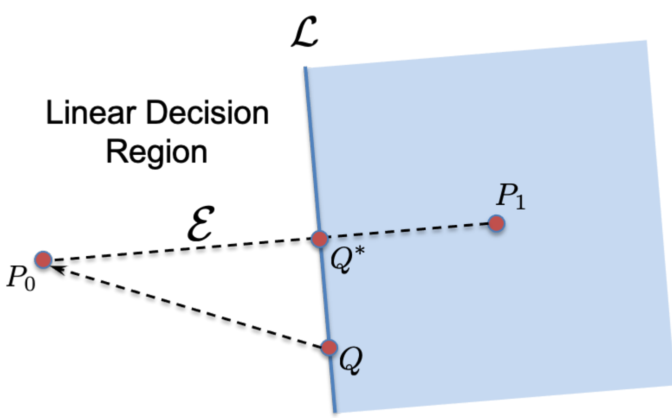
- Usually assume the family to be smooth, $\frac{\partial}{\partial\theta}P(x,\theta)$ exist, finite
- Can have higher dimensional parameters
- Why do maximum likelihood estimate?
- Distribution matching

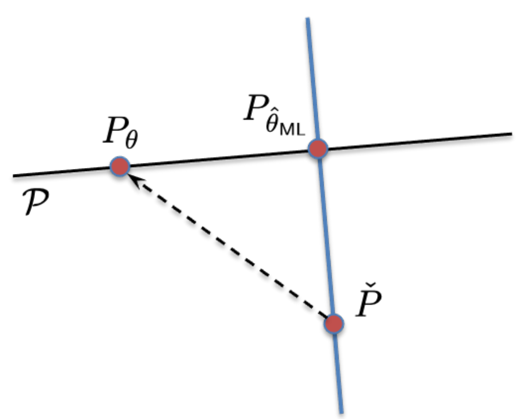
$$
\begin{aligned}
\arg\max_{\hat{\theta}} \frac{1}{n}\sum_{i=1}^n \log P(x_i;\hat{\theta}) &= \arg\max_{\hat{\theta}} \mathbb{E}_{\mathsf{x}\sim\check{P}_{\mathsf{x}}(\cdot;x_1^n)}\left[\log P(\mathsf{x};\hat{\theta})\right] \\
&= \arg\min_{\hat{\theta}} \mathbb{E}_{\mathsf{x}\sim\check{P}(\cdot;x_1^n)}\left[\log \frac{\check{P}(\mathsf{x};x_1^n)}{P(\mathsf{x};\hat{\theta})}\right] \\
&= \arg\min_{\hat{\theta}} D\left(\check{P}(\cdot;x_1^n)||P(\cdot;\hat{\theta})\right)
\end{aligned}
$$

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

Corollary: if $\mathcal{P}$ is an exponential family, $\mathcal{P}=\mathcal{E}(P_0,f)$, and suppose $\mathbb{E}_{\mathsf{x}\sim\check{P}}[f(\mathsf{x})]=\gamma$, then

$$P(\cdot;\hat{\theta}_{\mathsf{ML}}) \in \mathcal{P} \cap \mathcal{L}(f,\gamma)$$





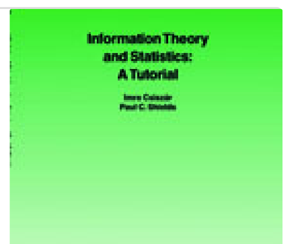
Takeaway message:

- 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:
  - Rate Distortion, "Rate-distortion theory: A mathematical basis for data compression, Englewood Cliffs, NJ: Prentice-Hall, 1971."
  - Error Exponent
  - Csiszar's book

now publishers - Information Theory and Statistics: A Tutorial
Publication Date: 15 Dec 2004 Download extract Abstract This tutorial is concerned with applications of information theory concepts in statistics, in the finite alphabet setting. The information
https://www.nowpublishers.com/article/Details/CIT-004



- What is difficult about this?
  - The geometry is complex.

Shun'ichi Amari - Wikipedia
Shun'ichi Amari, is a Japanese scholar born in 1936 in Tokyo, Japan. He majored in Mathematical Engineering in 1958 from the University of Tokyo then graduated in
W https://en.wikipedia.org/wiki/Shun%27ichi_Amari

  - Fisher information

$$\frac{\partial^2}{\partial t^2} D(P_t||P_0) = \frac{\partial}{\partial t}\eta(t) = \mathrm{var}_{\mathsf{x}\sim P_t}[f(\mathsf{x})] = \mathcal{I}_t$$

but $\mathcal{I}_t \geq 0$ 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: $P_{\mathsf{y}|\mathsf{x}}, P_{\mathsf{z}|\mathsf{x}}$. Even if $I(\mathsf{x};\mathsf{y}) > I(\mathsf{x};\mathsf{z})$, doesn't mean the channel $\mathsf{x} \rightarrow \mathsf{z}$ 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

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

    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 $\check{P}$, Estimated model $\hat{P}$.

Color code

Definition

Theorem

Key Problem

4. Fisher Information Metric

Definition: Fisher Information

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

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

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

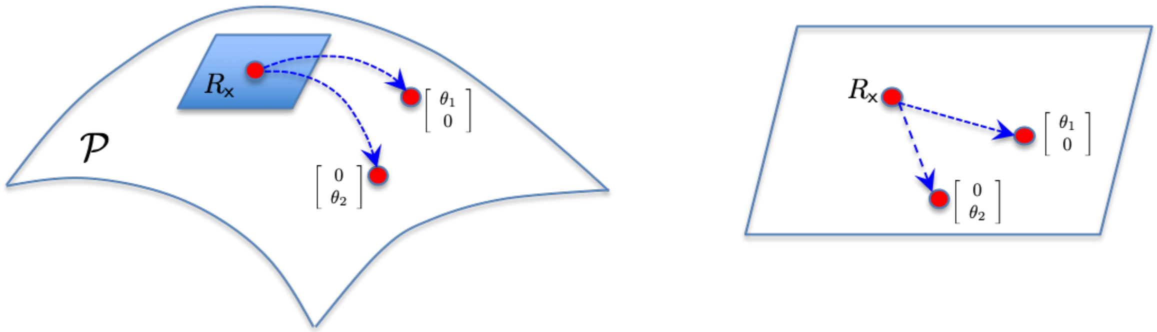
Understand the Definition

- Every distribution involved is close to $P_{\mathsf{x}}(\cdot;\underline{\theta})$
  - Reference distribution:

$$R_{\mathsf{x}} \triangleq P_{\mathsf{x}}(\cdot;\underline{\theta}=\underline{0})$$

- Think of all entries in $\underline{\theta}$ are restricted to be within $(-\epsilon,+\epsilon)$
- Each θ_i corresponds to a curve

$$\underline{\theta} = [0,\dots,0,\theta_i,0,\dots,0] \quad\longrightarrow\quad P_{\mathsf{x}}(x;\underline{\theta}) \triangleq R_{\mathsf{x}}(x)\cdot(1+\theta_i\cdot f_i(x)),\ x\in\mathcal{X}$$



- Log likelihood ratio

$$\log P_{\mathsf{x}}(x;\underline{\theta}) - \underbrace{\log P_{\mathsf{x}}(x;\underline{0})}_{R_{\mathsf{x}}(x)} = \log(1+\theta_i\cdot f_i(x)) = \theta_i\cdot f_i(x) + O(\epsilon^2),\ \forall x\in\mathcal{X}$$

- Perturbation accumulates

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

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

- Fisher information:

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

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

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

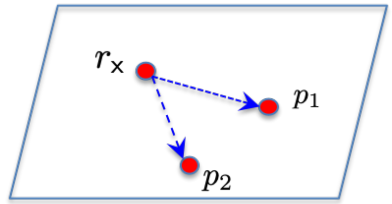
  - has to stay on the simplex:

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

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

    For given two distributions $P_1, P_2 \in \mathcal{N}_\epsilon(R_{\mathsf{x}})$, define f_1, f_2 by

$$P_i(x) = R_{\mathsf{x}}(x)\cdot(1+f_i(x)), \quad i=1,2$$



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 $\mathbb{E}_{\mathsf{x}\sim R}[f(x)] = 0$

       The information vector for f is written as $\phi^{(f)} \in \mathbb{R}^{\mathcal{X}}$, with

       $$\boxed{\phi^{(f)} \triangleq [\sqrt{R(x)} \cdot f(x),\ x \in \mathcal{X}]^T}$$

    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

       $$\boxed{\phi^{(P)}(x) \triangleq \sqrt{R(x)} \cdot \left(\frac{P(x)}{R(x)} - 1\right) = \frac{1}{\sqrt{R(x)}} \cdot (P(x) - R(x)),\ x \in \mathcal{X}}$$

First Properties:

1. Inner product and Covariance:

   $$\langle \phi^{(f_1)}, \phi^{(f_2)} \rangle = \mathbb{E}_{\mathsf{x}\sim R}[f_1(\mathsf{x}) f_2(\mathsf{x})]$$

2. Norm and variance:

   $$\|\phi^{(f)}\|^2 = \mathrm{var}_{\mathsf{x}\sim R}[f(\mathsf{x})]$$

3. Orthogonal functions iff uncorrelated (w.r.t. R)

K-L Divergence

For $P, Q \in \mathcal{N}_\epsilon(R)$,

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

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

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

$$\begin{aligned} D(P||Q) &= \sum_x P(x) \cdot \log \frac{P(x)}{Q(x)} = \sum_x P(x) \cdot \left(\log \frac{P(x)}{R(x)} - \log \frac{Q(x)}{R(x)}\right) \\ &= \sum_x P(x) \cdot [\log(1 + f(x)) - \log(1 + g(x))] \\ &= \sum_x [R(x) + \underbrace{R(x) \cdot f(x)}_{O(\epsilon)}] \cdot \left[\underbrace{f(x)}_{O(\epsilon)} - \underbrace{\frac{1}{2}f^2(x)}_{O(\epsilon^2)} - \underbrace{g(x)}_{O(\epsilon)} + \underbrace{\frac{1}{2}g^2(x)}_{O(\epsilon^2)} + O(\epsilon^3)\right] \\ &= \underbrace{\sum_x R(x)(f(x) - g(x))}_{0} \\ &\quad + \sum_x R(x)\left(-\frac{1}{2}f^2(x) + \frac{1}{2}g^2(x) + f(x) \cdot (f(x) - g(x))\right) + o(\epsilon^2) \\ &= \frac{1}{2}\mathbb{E}_{\mathsf{x} \sim R}[(f(\mathsf{x}) - g(\mathsf{x}))^2] + o(\epsilon^2) \end{aligned}$$

CLT and Asymptotic Normality

Recall Large Deviations / Sanov Theorem

$$\begin{aligned} \mathbb{P}_{\mathsf{x}_1^n \sim \text{i.i.d.} P}\left(\mathsf{x}_1^n \in T_Q\right) &\doteq e^{-nD(Q||P)} \\ &= e^{-n \cdot \frac{1}{2}\|\phi^{(Q)} - \phi^{(P)}\|^2 + o(\epsilon^2)} \end{aligned}$$

- The empirical distribution $\check{P}_{\mathsf{x}}(\cdot; \mathsf{x}_1^n) = Q$ is random, corresponding $\phi^{(Q)}$ 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 $\hat{\theta}_{\mathrm{ML}}$

- CLT: if $f(\mathsf{x})$ has zero-mean and unit variance w.r.t. R, $\frac{1}{\sqrt{n}} \sum_i f(\mathsf{x}_i) \to N(0,1)$
- Asymptotic efficiency of ML estimate:

$$\sqrt{n} \cdot (\hat{\theta}_{\mathrm{ML}} - \theta) \quad \to \quad N(0, \tfrac{1}{\mathcal{I}(\theta)})$$

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

6. Example: Akaike Information Criterion

Akaike information criterion - Wikipedia

The Akaike information criterion (AIC) is an estimator of prediction error and thereby relative quality of statistical models for a given set of data. Given a collection of models

W https://en.wikipedia.org/wiki/Akaike_information_criterion

- Consider a sequence of nested parameterized families $\mathcal{P}_1 \subset \mathcal{P}_2 \subset \ldots \subset \mathcal{P}_m$
- with increasing dimensionality of parameters

  $\mathcal{P}_k = \{P_{\mathsf{x}}(\cdot; \theta^k), \ \theta^k \in \mathbb{R}^k\}, \quad k = 1, \ldots, m$
- Observe $\mathsf{x}_1^n = x_1^n$, solve for each family

  $$\hat{P}_{\mathrm{ML}}^k = \arg\min_{\hat{P} \in \mathcal{P}_k} \ D(\check{P}(\cdot; \mathsf{x}_1^n) || \hat{P})$$
- Larger k, better matching,
- Which k to choose? How to penalize bigger families? Avoid over-fitting.

Akaike's observation:

We really want to minimize

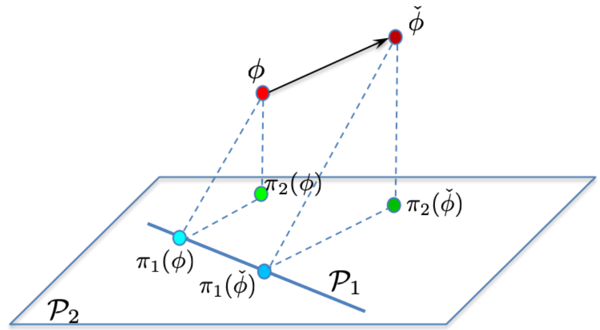
$D(P||\hat{P})$

but observe $D(\check{P}||\hat{P})$.

Locally:

1. All ML estimates are just projections

   $\hat{P}^k_{\mathrm{ML}} \leftrightarrow \pi_k(\check{\phi})$



2. What to compare?

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

   but how about

   $\|\pi_2(\check{\phi}) - \phi\| \;\gtrless\; \|\pi_1(\check{\phi}) - \phi\|$

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

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

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

$$\min_k \; \|\check{\phi} - \pi_k(\check{\phi})\|^2 + \frac{k}{n} - \frac{|\mathcal{X}| - k}{n} \quad \longrightarrow \quad \min_k \; D(\check{P}||\hat{P}^k_{\mathrm{ML}}) + \frac{k}{n}$$

7. Projections and Inner Products

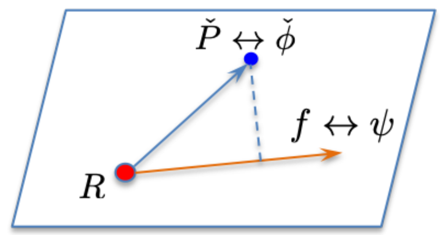
What does the direction of information vectors represent?

- Consider $\mathsf{x}_1, \ldots, \mathsf{x}_n \sim$ i.i.d. R_{x} ,
- but we observe empirical distribution $\check{P}$

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

The empirical average

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



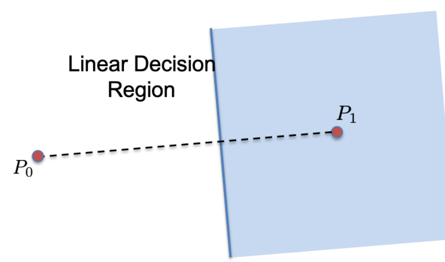
Back to Binary Hypothesis Testing

- Recall linear decision region

  $\frac{1}{n}\sum_{i=1}^{n} \log f(x_i) = \mathbb{E}_{\mathsf{x}\sim\check{P}}[f(\mathsf{x})] \underset{\hat{H}_0}{\overset{\hat{H}_1}{\gtrless}} \gamma$

  and the optimal

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

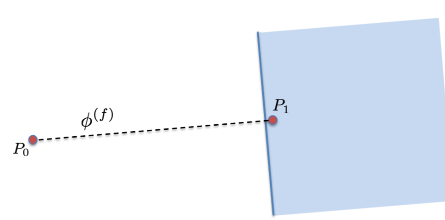


- A binary query corresponds to a vector

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

  for the log-likelihood function

  - LLR = project observed empirical distribution $\check{\phi}$ on ψ
  - length = $D(P_1||P_0)$, maximum one-sided error exponent

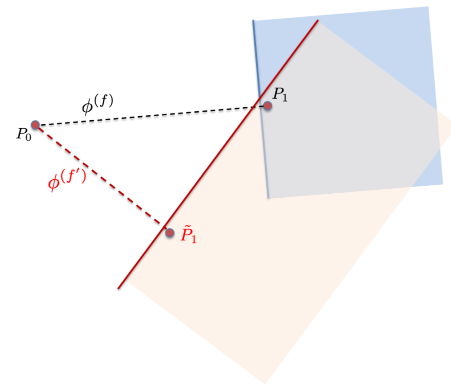


- What if we use a different statistic $f' \neq \log\frac{P_1}{P_0}$?

$$\tilde{P}_1 \triangleq \arg \min_{Q: \mathbb{E}_Q[f'] = \mathbb{E}_{P_1}[f']} D(Q \| P_0)$$

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 $R_{\mathsf{xy}} = P_{\mathsf{x}} \cdot P_{\mathsf{y}}$, independent with the same marginals

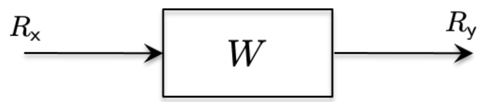
Definition: Canonical Dependence Matrix (CDM) $B \in \mathbb{R}^{\mathcal{X} \times \mathcal{Y}}$

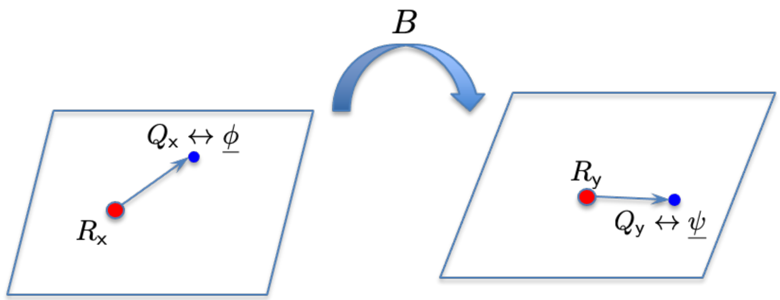
$$B(x, y) \triangleq \frac{P_{\mathsf{xy}}(x, y) - P_{\mathsf{x}}(x) P_{\mathsf{y}}(y)}{\sqrt{P_{\mathsf{x}}(x) P_{\mathsf{y}}(y)}}, \quad (x, y) \in \mathcal{X} \times \mathcal{Y}$$

- Inherited property

$$\frac{1}{2} \|B\|^2 \approx D(P_{\mathsf{xy}} \| P_{\mathsf{x}} P_{\mathsf{y}}) = I(\mathsf{x}; \mathsf{y})$$

- x, y have symmetric positions
- Describes how the two random variables are dependent
- Can be viewed as a channel





- $W = P_{\mathsf{y}|\mathsf{x}}$ defines a channel
- By definition, if input is $R_{\mathsf{x}} = P_{\mathsf{x}}$, the output is $R_{\mathsf{y}} = P_{\mathsf{y}}$
- If we change input to be $Q_{\mathsf{x}} \leftrightarrow \underline{\phi} \in \mathbb{R}^{\mathcal{X}}$

$$Q_{\mathsf{x}}(x) = R_{\mathsf{x}}(x) + \sqrt{R_{\mathsf{x}}(x)} \cdot \phi(x), \quad x \in \mathcal{X}$$

The output would be $Q_{\mathsf{y}} \leftrightarrow \underline{\psi} \in \mathbb{R}^{\mathcal{Y}}$,

$$\begin{aligned}
Q_{\mathsf{y}}(y) &= \sum_x P_{\mathsf{y}|\mathsf{x}}(y|x) \cdot Q_{\mathsf{x}}(x) \\
&= \sum_x P_{\mathsf{y}|\mathsf{x}}(y|x) \cdot R_{\mathsf{x}}(x) \cdot \left(1 + \frac{\phi(x)}{\sqrt{R_{\mathsf{x}}(x)}}\right) \\
&= R_{\mathsf{y}}(y) + \sum_x P_{\mathsf{xy}}(x,y) \cdot \frac{\phi(x)}{\sqrt{R_{\mathsf{x}}(x)}} \\
&= R_{\mathsf{y}}(y) + \sum_x \left(P_{\mathsf{xy}}(x,y) - P_{\mathsf{x}}(x) P_{\mathsf{y}}(y)\right) \cdot \frac{\phi(x)}{\sqrt{R_{\mathsf{x}}(x)}} \\
&= R_{\mathsf{y}}(y) + \sqrt{R_{\mathsf{y}}(y)} \cdot \left(\sum_x \underbrace{\frac{P_{\mathsf{xy}}(x,y) - P_{\mathsf{x}}(x) P_{\mathsf{y}}(y)}{\sqrt{R_{\mathsf{x}}(x)}\sqrt{R_{\mathsf{y}}(y)}}}_{B(x,y)} \cdot \phi(x)\right)
\end{aligned}$$

Theorem: B- matrix as a map

$$\boxed{\underline{\psi} = B \cdot \underline{\phi}}$$

Map of functions

Suppose

$$\underline{\phi} \leftrightarrow f: \quad \sqrt{R_{\mathsf{x}}(x)} \cdot f(x) = \phi(x), \quad x \in \mathcal{X}$$
$$\underline{\psi} \leftrightarrow g: \quad \sqrt{R_{\mathsf{y}}(y)} \cdot g(y) = \psi(y), \quad y \in \mathcal{Y}$$

Define $\underline{\psi} = B \cdot \underline{\phi}$, what function operation is this?

$$\begin{aligned}
g(y) &= \frac{1}{\sqrt{R_{\mathsf{y}}(y)}} \cdot \psi(y) = \frac{1}{\sqrt{R_{\mathsf{y}}(y)}} \cdot \left(\sum_x B(x, y) \cdot \phi(x) \right) \\
&= \frac{1}{\sqrt{R_{\mathsf{y}}(y)}} \cdot \left(\sum_x \frac{P_{\mathsf{xy}}(x, y) - P_{\mathsf{x}}(x) P_{\mathsf{y}}(y)}{\sqrt{P_{\mathsf{x}}(x) P_{\mathsf{y}}(y)}} \cdot \phi(x) \right) \\
&= \sum_x \frac{P_{\mathsf{xy}}(x, y)}{P_{\mathsf{y}}(y)} \cdot \frac{\phi(x)}{\sqrt{R_{\mathsf{x}}}(x)} \\
&= \mathbb{E}[f(\mathsf{x}) | \mathsf{y} = y], \qquad \forall y
\end{aligned}$$

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

B matrix is the conditional expectation operator.

Part III: Machine Learning

9. Example: Conjugator Prior Family

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

$\mathcal{P} = \{P_{\mathsf{x}}(\cdot;\theta), \theta \in \mathbb{R}\}$ is called the conjugate prior family if for any value of y, $P_{\mathsf{x}|\mathsf{y}}(\cdot|y) \in \mathcal{P}$.

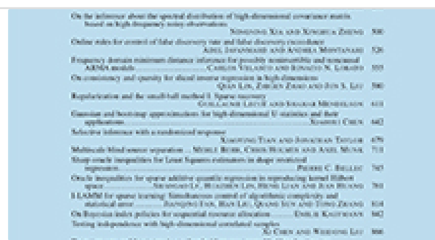
- Update knowledge turned into update parameters
- Bernoulli/Beta; Categorical/ Dirichlet, Poisson/Gamma, Normal (fix σ^2)/ Normal

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 Θ through the property of

https://projecteuclid.org/journals/annals-of-statistics/volume-7/issue-2/Conjugate-Priors-for-Exponential-Families/10.1214/aos/1176344611.full



If the observation model is an exponential family

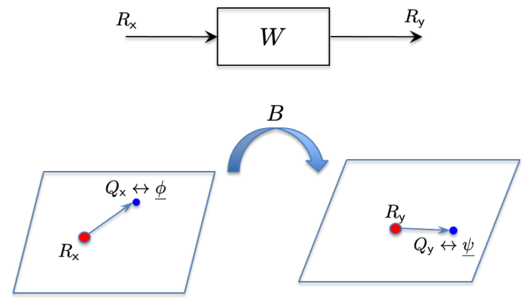
$P_{\mathsf{y}|\mathsf{x}}(y|x) = \exp(x \cdot t(y) - \alpha(x))$

then the conjugate prior $P_{\mathsf{x}}(\cdot;\theta)$ must satisfy that for all θ, evaluated w.r.t. $P_{\mathsf{x}}(\cdot;\theta) \cdot P_{\mathsf{y}|\mathsf{x}}$,

$\mathbb{E}[\mathbb{E}[t(\mathsf{y})|\mathsf{x}]|\mathsf{y}] = a \cdot t(\mathsf{y}) + b$, for some constants a, b.

The geometric view:

1. Observe a sequence $\check{y}_1, \ldots, \check{y}_n$ with empirical distribution $\check{P}_{\mathsf{y}} = Q_{\mathsf{y}} \leftrightarrow \underline{\psi}$
2. Symmetric story, the posterior $P_{\mathsf{x}|\mathsf{y}_1^n}(\cdot|\check{y}_1^n) = Q_{\mathsf{x}} \leftrightarrow \underline{\phi} = B^T \cdot \underline{\psi}$
3. Conjugate prior: regardless of $\underline{\psi}$, the posterior is always in a 1-D family : $\underline{\phi}$ remains in the same direction.



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

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

10. The multi-dimensional nature of dependence

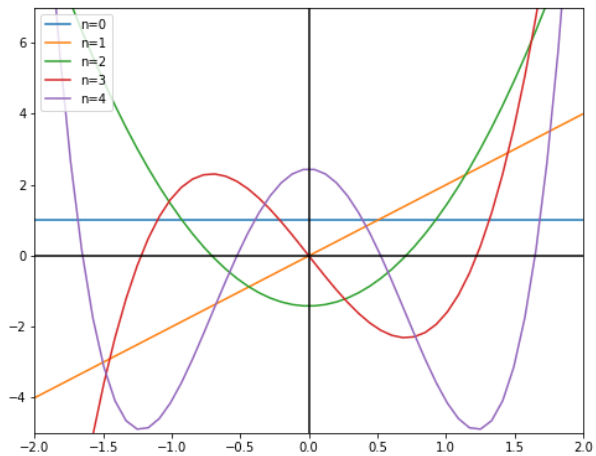
$$B = \sum_i \sigma_i \cdot \underline{u}_i \cdot \underline{v}_i^T$$

- Dependence over multiple modes
  $I(\mathsf{x};\mathsf{y}) \approx \frac{1}{2} \cdot \sum_i \sigma_i^2$

- **Example**: broadcast channel,
  - $I(\mathsf{x};\mathsf{y}) > I(\mathsf{x};\mathsf{z})$ does not mean we cannot transmit a private message $\mathsf{x} \to \mathsf{z}$ that is not decodable by y.
  - More capable (El-Gammal 79') : B_{xy} dominates B_{xz} in every mode.
    $||B_{\mathsf{xy}} \cdot \underline{\phi}_{\mathsf{x}}||^2 \geq ||B_{\mathsf{xz}} \cdot \underline{\phi}_{\mathsf{x}}||^2, \quad \forall \underline{\phi}_{\mathsf{x}}$

- **Example:** Strong DPI
  - All singular values of B are less than or equal to 1.
    $||B_{\mathsf{xy}} \cdot \underline{\phi}_{\mathsf{x}}||^2 \leq ||\underline{\phi}_{\mathsf{x}}||^2, \quad \forall \underline{\phi}_{\mathsf{x}} \quad \Leftrightarrow \quad D(P_{\mathsf{y}}||Q_{\mathsf{y}}) \leq D(P_{\mathsf{x}}||Q_{\mathsf{x}}),$
  - 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



11. Renyi Correlation, CCA

Definition: Hirschfeld-Gebelein-Renyi Maximal correlation:

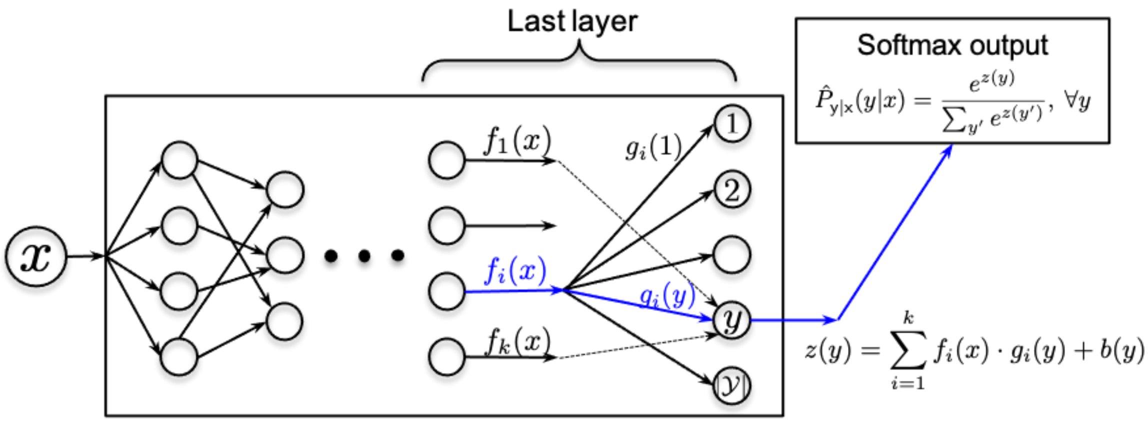
Given P_{xy}:

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

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

- 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:

$$\hat{P}^{(f,g)}_{\mathsf{y|x}}(y|x) = \frac{\exp\left[\sum_{i=1}^{k} f_i(x) \cdot g_i(y) + b(y)\right]}{\sum_{y'} \exp\left[\sum_{i=1}^{k} f_i(x) \cdot g_i(y') + b(y')\right]}, \quad y \in \mathcal{Y}$$

- Cross-Entropy Loss, ML for discriminative model.

$$\arg\min_{f,g} \; D(\check{P}_{\mathsf{x}} \cdot \check{P}_{\mathsf{y|x}} || \check{P}_{\mathsf{x}} \cdot \hat{P}^{(f,g)}_{\mathsf{y|x}})$$

```
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

$$R_{\mathsf{x}}(x) = \check{P}_{\mathsf{x}}(x), \quad \forall x$$

$$R_{\mathsf{y}}(y) \propto e^{b(y)}, \quad \forall y$$

$$\hat{P}^{(f,g)}_{\mathsf{y}|\mathsf{x}}(y|x) = R_{\mathsf{y}}(y) \cdot \frac{\exp\left[\sum_{i=1}^{k} f_i(x) \cdot g_i(y)\right]}{\sum_{y'} \exp\left[\sum_{i=1}^{k} f_i(x) \cdot g_i(y')\right]}, \quad y \in \mathcal{Y}$$

2. Learned model

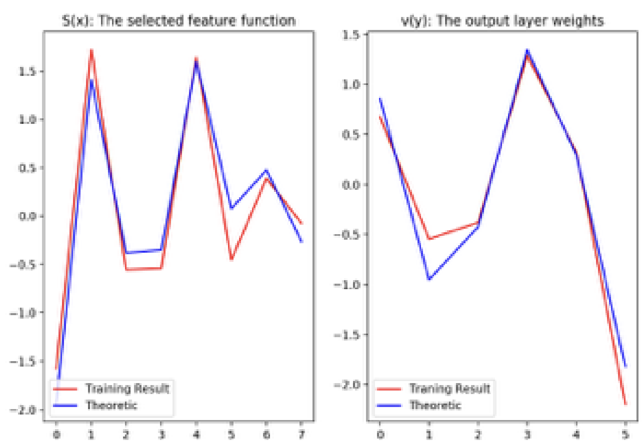
$$\hat{P}^{(f,g)}_{\mathsf{y}|\mathsf{x}} \longrightarrow R_{\mathsf{y}}, \hat{B}^{(f,g)} :$$

$$\hat{B}^{(f,g)}(x,y) = \sqrt{R_{\mathsf{x}}(x) R_{\mathsf{y}}(y)} \cdot \left(\sum_{i=1}^{k} f_i(x) \cdot g_i(y) \right) \quad \forall x, y$$

3. Optimization

$$\arg\min_{f,g} \|\check{B} - \hat{B}^{(f,g)}\|^2$$

4. Solution: SVD



5. How was this numerically solved?

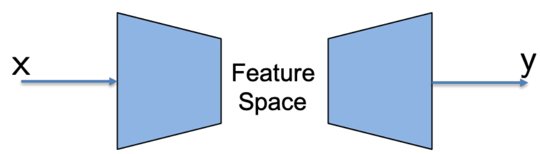
BackProp:

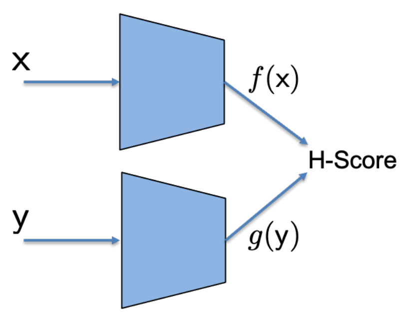
- Fix f: $g(y) \leftarrow \mathbb{E}[f(\mathsf{x})|\mathsf{y} = y], \ \forall y$, equivalent to $\underline{\psi}^{(g)} \leftarrow B \cdot \underline{\phi}^{(f)}$
- Fix g: $f(x) \leftarrow \mathbb{E}[g(\mathsf{y})|\mathsf{x} = x], \ \forall x$, equivalent to $\underline{\phi}^{(f)} \leftarrow B^T \cdot \underline{\psi}^{(g)}$

13. What is this good for?

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

$$
\begin{aligned}
H(\underline{f},\underline{g}) = \|\check{B} - \hat{B}^{(f,g)}\|^2 &\triangleq \mathbb{E}_{\mathsf{x},\mathsf{y}\sim\check{P}_{\mathsf{xy}}}\left[\underline{f}^T(\mathsf{x})\cdot\underline{g}(\mathsf{y})\right] - \frac{1}{2}\mathrm{trace}\left(\mathrm{cov}(\underline{f})\cdot\mathrm{cov}(\underline{g})\right)\\
H(\underline{f}) = H(\underline{f},\underline{g}^*) &\triangleq \mathbb{E}_{\mathsf{y}\sim p_{\mathsf{y}}}\left[\mathbb{E}[\underline{f}(\mathsf{x})|\mathsf{y}=y]^T\cdot\mathrm{cov}(\underline{f})^{-1}\cdot\mathbb{E}[\underline{f}(\mathsf{x})|\mathsf{y}=y]\right]
\end{aligned}
$$





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

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