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

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

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

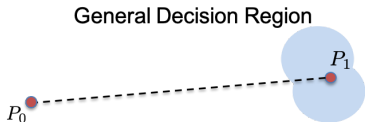
Consider $\mathrm{x}_1, \ldots, \mathrm{x}_n$ i.i.d. distributed from either P_0 or P_1.

- 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

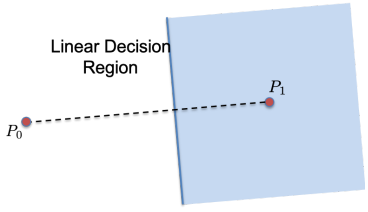
  $$\frac{1}{n}\sum_{i=1}^n \log\frac{P_1(x_i)}{P_0(x_i)} = \mathbb{E}_{\mathrm{x}\sim\check{P}(\cdot;x_1^n)}\left[\log\frac{P_1}{P_0}(\mathrm{x})\right]$$



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

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

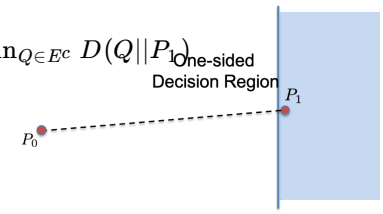
  claim $\hat{H}_1$ iff $\mathrm{x}_1^n = x_1^n \in \bigcup_{Q\in E} T_Q$



- 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

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

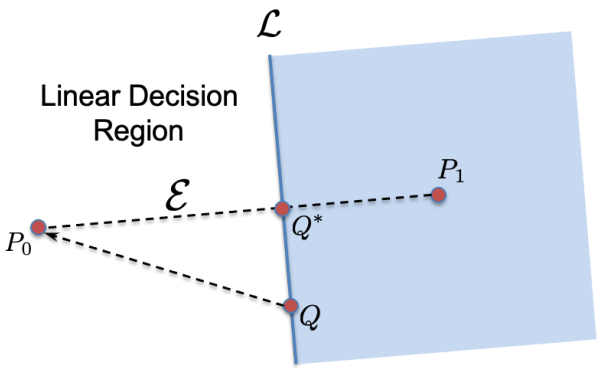
- A number of nice properties

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

- $\eta(t)$ 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.

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

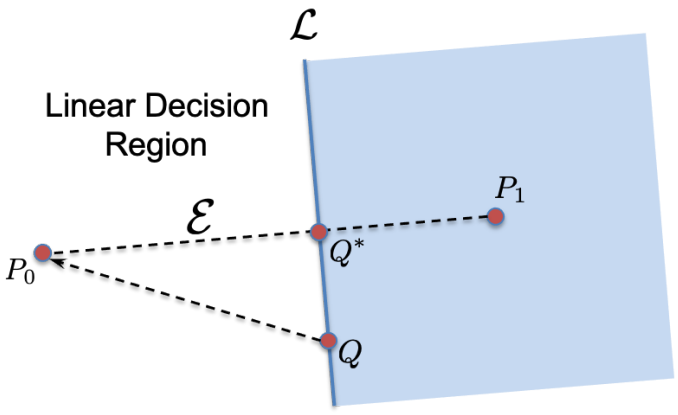
- Usually assume the family to be smooth, $\frac{\partial}{\partial\theta}P(x,\theta)$ exist, finite
- Can have higher dimensional parameters
- Why do maximum likelihood estimate?
- Distribution matching

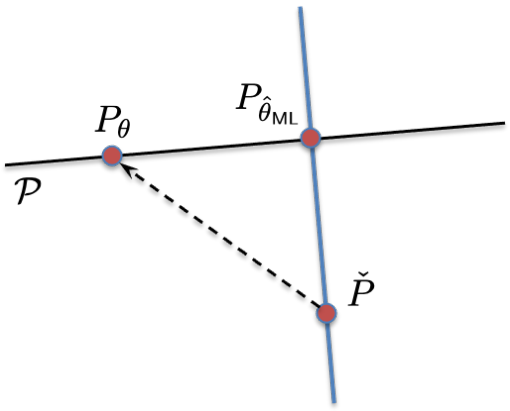
$$\begin{aligned}\arg\max_{\hat{\theta}} \frac{1}{n}\sum_{i=1}^n \log P(x_i;\hat{\theta}) &= \arg\max_{\hat{\theta}} \mathbb{E}_{\mathsf{x}\sim\check{P}_{\mathsf{x}}(\cdot;x_1^n)}\left[\log P(\mathsf{x};\hat{\theta})\right] \\ &= \arg\min_{\hat{\theta}} \mathbb{E}_{\mathsf{x}\sim\check{P}(\cdot;x_1^n)}\left[\log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

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

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



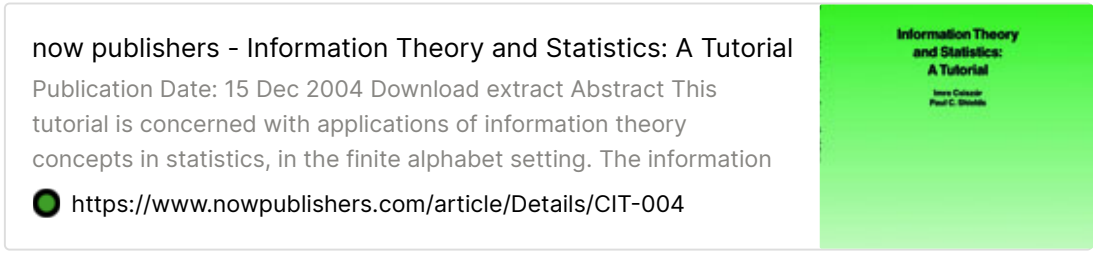


Takeaway message:

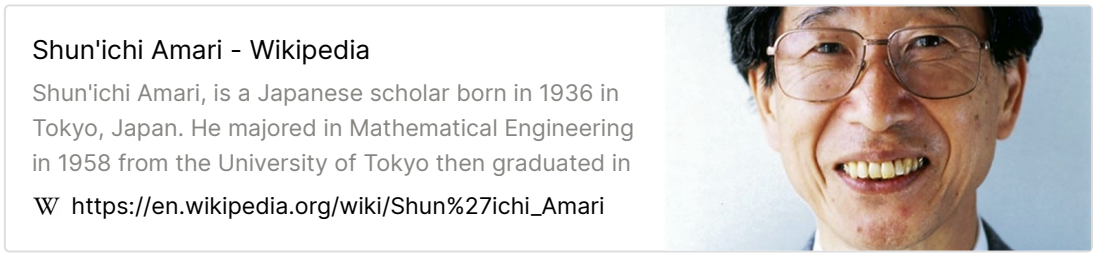
- A model family is a manifold/plane in the space of distributions;
- Learning is also a projection, from the observed empirical distribution to the model family.

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



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



  - Fisher information

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

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 $\mathsf{x} \to \mathsf{z}$ 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.