# Information Geometry and Its Applications to Statistics

We shall begin by considering the problem of learning a probability distribution (popularly known as "distribution learning").

- Domain = $S$ [e.g., set of all english words in a corpus]
- Measured features for $x \in S$:
  - $T_1(x) = \mathbb{1}(\text{Length of } x \text{ is } > 5)$
    ↳ = 1 if length of $x > 5$
    = 0 otherwise
  - $T_2(x) = \mathbb{1}(x \text{ ends in 'e'})$

and so on...

**Main problem:** find a distribution on $S$ that satisfies constraints of the form

$$E[T_1(X)] = 0.3, \quad E[T_2(X)] = 0.45 \quad \text{and so on...}$$

but makes no other assumptions.

is as random as possible

the notion of randomness we typically use is that of Shannon entropy. However, there are other generalisations of Shannon entropy such as Rényi entropy, Tsallis entropy that can also be used as measures of randomness.

**Shannon entropy:** For $X \in S$ with distribution $p = \{p_x : x \in S\}$,

$$H(X) = H(p) = \sum_{x \in S} p_x \log_2 \frac{1}{p_x}$$

**Examples:**

- For a fair coin, $S = \{H, T\}$, $p = \{1/2, 1/2\}$, $H(p) = 1$

- For a coin with bias $3/4$,

$$S = \{H, T\}, \quad p = \{3/4, 1/4\},$$

$$H(p) = \frac{3}{4} \log_2 \frac{4}{3} + \frac{1}{4} \log_2 4 = 0.81$$

- For a coin with bias $0.99$,

$$S = \{H, T\}, \quad p = \{0.99, 0.01\},$$

$$H(p) = 0.99 \log_2 \frac{1}{0.99} + 0.01 \log_2 \frac{1}{0.01} = 0.08$$

Back to our problem

Main problem: find a distribution on $S$ that satisfies constraints of the form

$$E[T_1(X)] = 0.3, \quad E[T_2(X)] = 0.45 \text{ and so on} \dots$$

but makes no other assumptions.

is as random as possible → maximum Shannon entropy

Mathematically, we would like to solve the following problem:

$$\max_p \sum_{x \in S} p_x \log \frac{1}{p_x}$$

subject to $\sum_{x \in S} p_x T_1(x) = 0.3$

$$\sum_{x \in S} p_x T_2(x) = 0.45 \text{ and so on} \dots$$

$$p_x \geq 0 \ \forall x \in S, \quad \sum_{x \in S} p_x = 1$$

More generally, our problem of interest is of the following kind:

$$\max_{p} \sum_{x \in S} p_x \log \frac{1}{p_x}$$

$$\text{Subject to} \quad \sum_{x \in S} p_x T_i(x) = b_i \;, \quad i = 1, \ldots, k.$$

$$p_x \geq 0 \;\; \forall x \in S \;, \quad \sum_{x \in S} p_x = 1$$

We can use the method of Lagrange multipliers to solve this problem. Let

$$F(p, \lambda, \nu) = \sum_{x \in S} p_x \log \frac{1}{p_x} - \sum_{i=1}^{k} \lambda_i \left( \sum_{x \in S} p_x T_i(x) - b_i \right) - \nu \left( \sum_{x \in S} p_x - 1 \right).$$

Setting $\frac{\partial F}{\partial p_x} = 0 \;\; \forall x \in S$, we get

$$-1 + \log \frac{1}{p_x} - \sum_{i=1}^{k} \lambda_i T_i(x) - \nu = 0$$

$$\Rightarrow p_x = e^{-\sum_i \lambda_i T_i(x)} \cdot e^{-(\nu+1)}.$$

Using the condition that $\sum_{x \in S} p_x = 1$, we get

$$p_x = \frac{1}{Z} e^{-\sum_{i=1}^{k} \lambda_i T_i(x)}, \text{ where } Z = \sum_{x \in S} e^{-\sum_{i=1}^{k} \lambda_i T_i(x)} \text{ is the normalisation term.}$$

## An alternative formulation

Let us write Shannon entropy in a different way. Given a

probability distribution $p = \{p_x : x \in S\}$, we have

$$H(p) = \sum_{x \in S} p_x \log \frac{1}{p_x}$$

$$= - \sum_{x \in S} p_x \log p_x$$

$$= - \sum_{x \in S} p_x \log \frac{p_x}{1/|S|} + \log |S| \quad \longrightarrow \text{Here, } |S| = \text{no of elements in } S$$

$$= - D_{KL}(p_x \| u) + \log |S|,$$

where $u = \{\frac{1}{|S|}, \ldots, \frac{1}{|S|}\}$ denotes the uniform distribution on $S$, and $D_{KL}$ denotes KL divergence.

## KL divergence

Given a domain $S$ and two probability distributions $p = \{p_x : x \in S\}$ and $q = \{q_x : x \in S\}$ on $S$, the KL divergence between $p$ and $q$ is defined as

$$D_{KL}(p \| q) := \sum_{x \in S} p_x \log \frac{p_x}{q_x}.$$

## Some properties of KL divergence

1. $D_{KL}(p \| q) \geq 0$. $D_{KL}(p \| q) = 0$ if and only if $p = q$.
2. $D_{KL}(p \| q) \neq D_{KL}(q \| p)$ in general.

Because of the above properties, information theorists regard $D_{KL}$ as a measure of "distance" between probability distributions.

Because $H(p) = \log|S| - D_{KL}(p \| u)$, Shannon entropy is a measure of how far a distribution is from the uniform distribution. Maximising Shannon entropy is equivalent to minimising $D_{KL}$ with respect to the uniform distribution.

Therefore, our central problem of learning an unknown distribution may be written as below:

$$\min_{p} \; D_{KL}(p \| u)$$

$$\text{subject to} \quad \sum_{x \in S} p_x \, T_i(x) = b_i, \quad i = 1, \dots, k$$

$$p_x \geq 0 \;\; \forall \, x \in S, \quad \sum_x p_x = 1.$$

We know that the solution to the above problem is

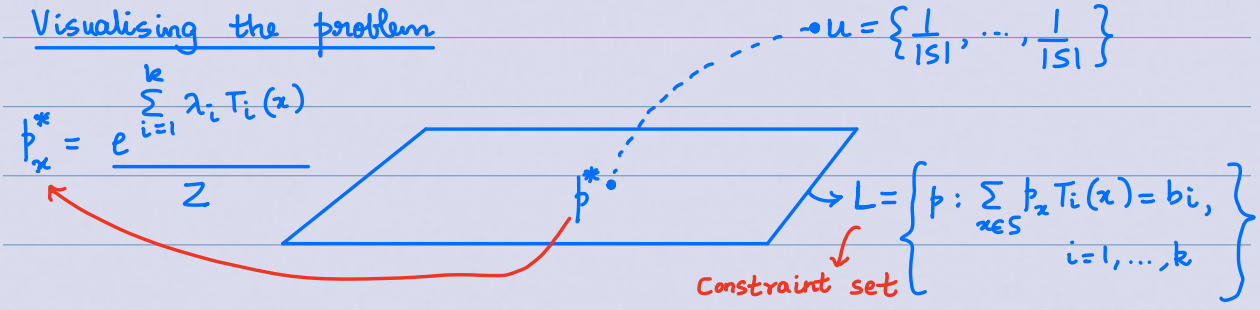
$$p_x = \frac{e^{\sum_{i=1}^{k} \lambda_i T_i(x)}}{Z}, \quad \text{where} \quad Z = \sum_{x \in S} e^{\sum_{i=1}^{k} \lambda_i T_i(x)}$$

(Z is sometimes called the partition function)

Here, $\lambda_i$'s are a function of $b_i$'s.

(We will see later how to determine $\lambda_i$'s)

Visualising the problem



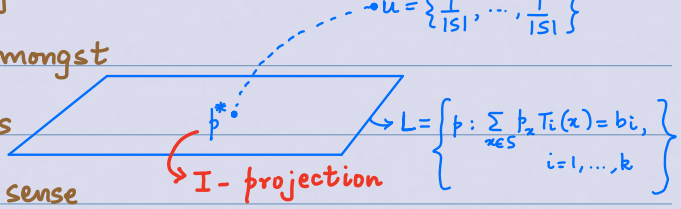
$S$ = domain (set of all words in a corpus)

Features = $T_i$, $i = 1, \ldots, k$ (does the word end in 'e', is the length of the word > 5 etc.)

Probability distribution = ?

In the absence of any prior information about the distribution of the words in the corpus, it is reasonable to assume that the words are uniformly distributed. However, after making some measurements (features), we may want to update the distribution to satisfy the measured features. This updated distribution is $p^*$.

In the language of information geometry, $p^*$ is the I-projection of $u$ onto $L$.




The speciality of $p^*$ is that amongst all the probability distributions in $L$, it is the closest (in the sense of $D_{KL}$) to the uniform distribution (i.e., $p^*$ has the max. Shannon entropy).

## Availability of prior information

Suppose, through prior measurements, we know that the distribution of words in the corpus is $\pi = \{\pi_x : x \in S\}$. Now, say measurements are made and $\pi$ needs to be updated to a distribution that satisfies the measurements. We then solve the following problem:

$$\min_{p} \quad D_{KL}(p \| \pi)$$

↳ notice the presence of $\pi$ instead of $u$

$$\text{subject to} \quad \sum_{x \in S} p_x \, T_i(x) = b_i, \quad i = 1, \dots, k$$

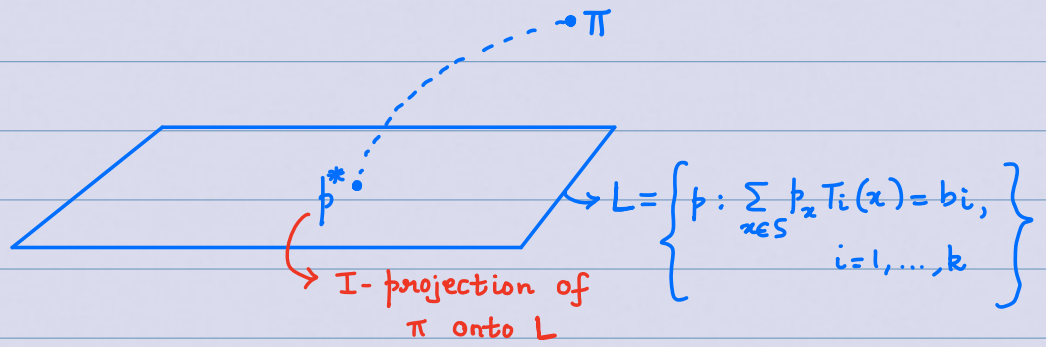
$$p_x \geq 0 \;\; \forall \, x \in S, \quad \sum_{x \in S} p_x = 1.$$

It can be shown that in this case (exercise)

$$p_x^* = \frac{1}{Z} \, \pi_x \cdot e^{\sum_{i=1}^{k} \lambda_i T_i(x)}$$

is the solution to the above problem.

Here, $Z = \sum_{x \in S} \pi_x \, e^{\sum_{i=1}^{k} \lambda_i T_i(x)}$ is the partition function (normalising term)




In the language of information geometry, we say $p^*$ is a member of the exponential family generated by $\pi$.

## Exponential families - A short primer

Most of the commonly used probability distributions such as Bernoulli, Gaussian, Poisson, multinomial are exponential families.

To define an exponential family, we need the following:

- Domain $S \subseteq \mathbb{R}^d$
- Base measure $h : \mathbb{R}^d \to \mathbb{R}_+$
- Features $T(x) = \left[ T_1(x), \cdots, T_k(x) \right]$, $x \in S$.

The exponential family generated by $h$ and $T_1, \ldots, T_k$ is a parametric family of distributions given by

$$p_\lambda(x) = \frac{1}{Z} h(x) \cdot e^{\sum_{i=1}^{k} \lambda_i T_i(x)}$$

, where $Z$ is the normalising term that makes $p_\lambda$ a probability distribution

Let

$$G(\lambda) := \log Z = \log \sum_{x \in S} h(x)\, e^{\sum_{i=1}^{k} \lambda_i T_i(x)}.$$

Then, we can write

$$p_\lambda(x) = h(x) \cdot e^{\sum_{i=1}^{k} \lambda_i T_i(x) - G(\lambda)}.$$

Here, $\lambda = [\lambda_1, \cdots, \lambda_k]$ is the vector of parameters.

Examples:

① Bernoulli:

- $S = \{0, 1\} \subseteq \mathbb{R}$
- $h(x) = 1 \quad \forall x \in S$ (h need not be a prob. dist.)
- $T(x) = x$

Then,

$$G(\lambda) = \log \left( e^{\lambda \cdot 0} + e^{\lambda \cdot 1} \right) = \log \left( 1 + e^{\lambda} \right).$$

Thus, $p_\lambda(x) = \dfrac{e^{\lambda x}}{1+e^{\lambda}}$, $x \in \{0,1\}$.

We are accustomed to using

$p_\lambda(1) = \theta$, $p_\lambda(0) = 1-\theta$. Here, $\theta$ and $\lambda$ are related as

$$\theta = \frac{e^{\lambda}}{1+e^{\lambda}}.$$

② Poisson:

$$S = \{0, 1, 2, \dots\}$$

$$h(x) = \frac{1}{x!}, \quad T(x) = x$$

Then, we have

$$G(\lambda) = \log\left(\sum_{x=0}^{\infty} \frac{1}{x!} e^{\lambda x}\right) = e^{\lambda}.$$

Thus, $p_\lambda(x) = \dfrac{1}{x!} e^{\lambda x - e^{\lambda}}$.

We are accustomed to the parametrisation $\mathrm{Poi}(\theta)$, where $\log\theta = \lambda$

③ Gaussian:

$$S = (-\infty, \infty)$$

$$h(x) = \frac{1}{\sqrt{2\pi}}, \quad T(x) = (x, x^2)$$

If we are interested in $N(\mu, \sigma^2)$, then $\lambda = [\lambda_1, \lambda_2]$ with

$$\lambda_1 = \mu/\sigma^2, \quad \lambda_2 = \frac{-1}{2\sigma^2}.$$

(notice the difference in parametrisation)

## Properties of exponential families

Recall $G(\lambda) = \log\left[\sum_{x \in S} h(x) \exp\left(\sum_{i=1}^{k} \lambda_i T_i(x)\right)\right]$.

- $G(\cdot)$ is strictly convex (recall, for Poisson, $G(\lambda) = e^{\lambda}$)
  - ↳ This means $\frac{dG}{d\lambda}$ (or $\left[\frac{\partial G}{\partial \lambda_1}, \cdots, \frac{\partial G}{\partial \lambda_k}\right]$ is 1-1 mapping).
- $\frac{\partial G}{\partial \lambda_i} = E[T_i]$, $i = 1, \cdots, k$ (exercise)

### How to determine $\lambda_1 \cdots \lambda_k$ ?

Example 1:

Fit a Poisson distribution to integer data with mean 7.5.

Answer: Feature is $T(x) = x$.

Also, for Poisson, we know that $G(\lambda) = e^{\lambda}$.

$$\frac{dG}{d\lambda} = E[T_1] = 7.5 \Rightarrow e^{\lambda} = 7.5 \Rightarrow \lambda = \ln(7.5).$$

Example 2:

We are told that a distribution over $\mathbb{R}$ has mean 0 and variance 5, but is otherwise as random as possible. Write down the form of the distribution.

Answer:

The measured features are $E[X] = 0$, $E[X^2] = 10$.

$\Rightarrow T_1(x) = x$, $T_2(x) = x^2$

$\Rightarrow$ these define a Gaussian dist.

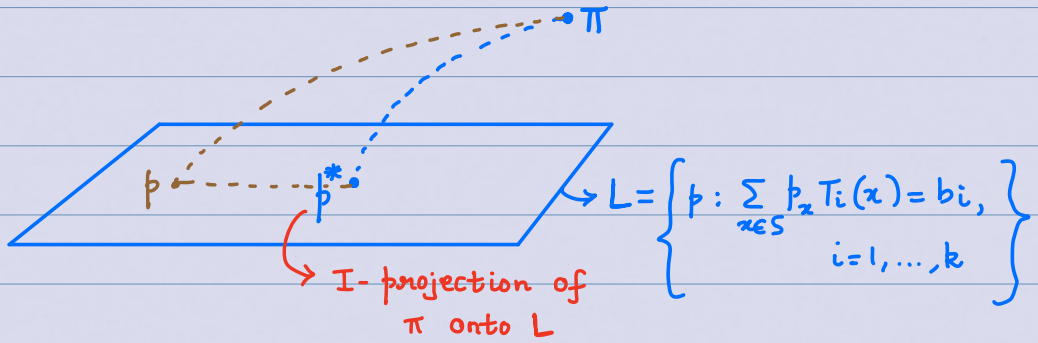
Given $\mu = 0$, $\sigma^2 = 10$, we have $\lambda_1 = \frac{\mu}{2\sigma^2}$, $\lambda_2 = \frac{-1}{2\sigma^2}$.

Thus, $\lambda_1 = 0$, $\lambda_2 = -0.05$. We then have

$$p_\lambda(x) \propto \exp\left(-0.05\, x^2\right).$$

Note: Going from $(\mu, \sigma^2)$ to $(\lambda_1, \lambda_2)$ in the Gaussian example or from $\theta$ to $\lambda$ in the Poisson example was relatively easy since $\left\{\frac{\partial G}{\partial \lambda_i}\right\}_i$ or $\frac{dG}{d\lambda}$ could be easily inverted. This may not always be the case. So, how do we find $\lambda = (\lambda_1, \dots, \lambda_k)$ in more complicated examples?

## Pythagorean Property




We know that $p^*$ is the closest to $\pi$ among all the members of $L$.

Now, for any $p \in L$, we have

$$D_{KL}(p\|\pi) - D_{KL}(p^*\|\pi) = \sum_x p_x \log \frac{p_x}{\pi_x} - \sum_x p_x^* \log \frac{p_x^*}{\pi_x}$$

$$\overset{(a)}{=} \sum_x p_x \log \frac{p_x}{\pi_x} - \sum_x p_x^* \left(\sum_{i=1}^k \lambda_i T_i(x) - G(\lambda)\right) \quad \text{(using the form for } p^*\text{)}$$

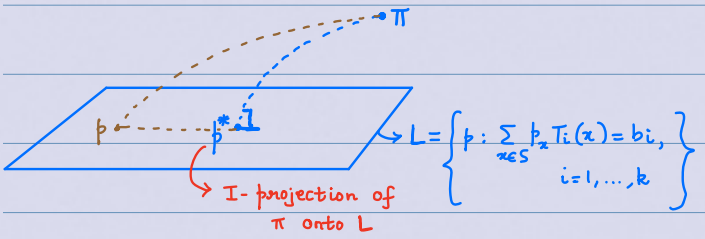
$$\overset{(b)}{=} \sum_x p_x \log \frac{p_x}{\pi_x} - \sum_x p_x \left( \sum_{i=1}^{k} \lambda_i T_i(x) - G(\lambda) \right)$$ (because $p$ & $p^*$ both lie in $L$)

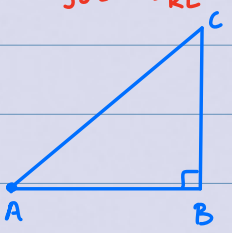
$$= \sum_x p_x \log \frac{p_x}{\pi_x} - \sum_x p_x \log \frac{p^*_x}{\pi_x}$$

$$= \sum_x p_x \log \frac{p_x}{p^*_x} = D_{KL}(p \| p^*).$$

Therefore, we have

$$\boxed{D_{KL}(p\|\pi) = D_{KL}(p\|p^*) + D_{KL}(p^*\|\pi)}$$ ← Pythagorean theorem for $D_{KL}$




$AC^2 = AB^2 + BC^2$

↳ Pythagorean theorem for squared distance

An important application of the Pythagorean property is the following I-projection algorithm by Csiszar:

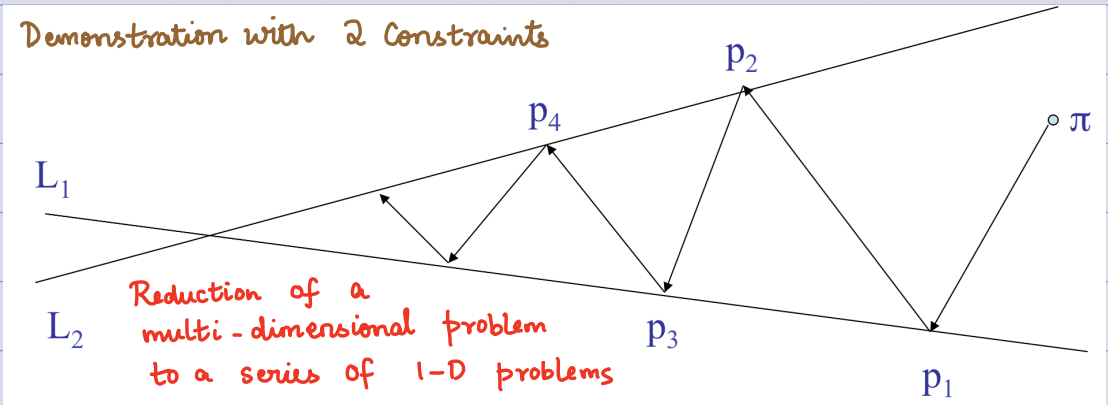
## Csiszar's Successive Projection Algorithm

Let

$$L_i = \{p : \sum_{x \in S} p_x T_i(x) = b_i\}$$ ← only $i^{th}$ constraint

Csiszar's algo is as below:

- Set $p_0 = \pi$ (given prior information)
- Project $p_t$ onto $L_{t \bmod k}$ to get $p_{t+1}$
- Loop until convergence




At any given time $t$, projecting $p_t$ onto $L_i$ simply involves finding $\lambda_i$ such that the distribution

$$p_t(x) \cdot e^{\lambda_i T_i(x) - G(\lambda_i)}$$

satisfies $E[T_i] = b_i$. However, we know that

$$\frac{dG}{d\lambda_i} = E[T_i] = b_i.$$

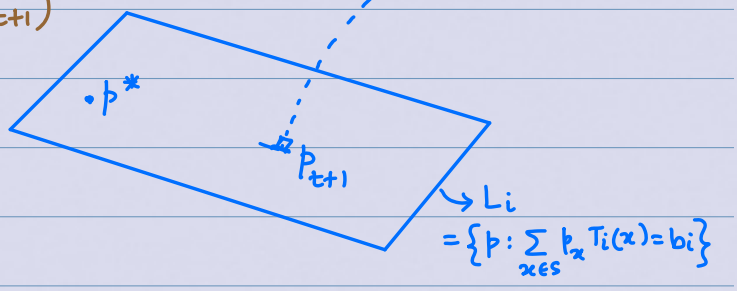
Therefore, we have to find $\lambda_i$ such that $G'(\lambda_i) = b_i$.

This can be accomplished using line search!

## Proof of Convergence

We know $p^* \in L_i$. By the Pythagorean property,

$$D_{KL}(p^* \| p_t) = D_{KL}(p^* \| p_{t+1}) + D_{KL}(p_{t+1} \| p_t)$$
$$\geq D_{KL}(p^* \| p_{t+1})$$




In this way, step by step, we can find the $\lambda_i$'s.

## Forward and Reverse I-projections

Given a prior $\pi$ and a constraint set $L$, the problem of **forward I-projection** can be described as

$$\min_{p \in L} D_{KL}(p \| \pi).$$

The element $p^* \in L$ which attains the minimum is called the forward I-projection of $\pi$ onto $L$.

On the contrary, the problem of **reverse I-projection** involves keeping the first argument of $D_{KL}$ fixed and minimising the second argument across $L$:

$$\min_{p \in L} D_{KL}(\pi \| p).$$

One instance of reverse I-projections may be found in the problem of maximum likelihood estimation.

## Maximum Likelihood Estimation

Let $L = \{p_\theta : \theta \in \Theta\}$ be a parametric family of probability distributions. Given $X_1, X_2, \ldots, X_n$ sampled independently according to $p_{\theta_0}$ for some unknown $\theta_0 \in \Theta$, our goal is to determine the most likely $p_\theta$ according to which $X_1, \ldots, X_n$ could have been sampled.

That is, we would like to obtain

$$\theta^* \in \arg\max_{\theta \in \Theta} p_\theta(X_1, \ldots, X_n).$$

Such a $\theta^*$, if it exists, is called the maximum likelihood estimate

Suppose $X_i$'s take value in a finite set $S = \{1, 2, \cdots, d\}$.

Then,

$$\log p_\theta(x_1, \cdots, x_n) = \sum_{i=1}^{n} \log p_\theta(x_i)$$

$$= \sum_{x \in S} N(x) \log p_\theta(x)$$

($N(x)$: #times '$x$' has appeared amongst $x_1, \cdots, x_n$)

$$= n \sum_{x \in S} \frac{N(x)}{n} \log p_\theta(x)$$

$$= n \left( \sum_{x \in S} \frac{N(x)}{n} \log \frac{p_\theta(x)}{\frac{N(x)}{n}} + \sum_{x \in S} \frac{N(x)}{n} \log \frac{N(x)}{n} \right)$$

Defining $\hat{p}(x) := \frac{N(x)}{n}$, $x \in S$, we get

$$\log p_\theta(x_1, \cdots, x_n) = n \left[ \sum_{x \in S} \hat{p}(x) \log \frac{p_\theta(x)}{\hat{p}(x)} + \sum_{x \in S} \hat{p}(x) \log \hat{p}(x) \right]$$

$$= -n \left( D_{KL}(\hat{p} \| p_\theta) + H(\hat{p}) \right).$$

Therefore, we have

$$\theta^* \in \arg\max_{\theta \in \Theta} p_\theta(x_1, \cdots, x_n)$$

$$= \arg\max_{\theta \in \Theta} \log p_\theta(x_1, \cdots, x_n) = \arg\min_{\theta \in \Theta} D_{KL}(\hat{p} \| p_\theta).$$

reverse I-projection

## Maximum Likelihood Estimation in Exponential Families

Let $\pi$ be a prior. Given $T_1, \dots, T_k$, the exponential family generated by $\pi$ and $T_1, \dots, T_k$ is a parametric family defined by $\lambda = [\lambda_1, \dots, \lambda_k]$ as

$$p_\lambda(x) = \pi(x) \cdot e^{\sum_{i=1}^{k} \lambda_i T_i(x) - G(\lambda)}$$

where $G(\lambda) = \log \left[ \sum_{x \in S} \pi(x) \cdot e^{\sum_{i=1}^{k} \lambda_i T_i(x)} \right]$.

Given $X_1, \dots, X_n \overset{iid}{\sim} p_{\lambda_0}$ for some unknown $\lambda_0$, what is the maximum likelihood estimate of $\lambda$?

Notice that

$$\log p_\lambda(X_1, \dots, X_n) = \sum_{i=1}^{k} \sum_{j=1}^{n} \lambda_i T_i(X_j) - nG(\lambda) + \sum_{j=1}^{n} \log \pi_{X_j}$$

Therefore, the max. likelihood estimate is given by $\lambda$ such that

$$\frac{\partial G}{\partial \lambda_i} = \frac{1}{n} \sum_{j=1}^{n} T_i(X_j) \quad , \quad i = 1, \dots, k.$$

Therefore, ML estimate = $\lambda$ for which the above condition holds!

## Summary

- Shannon entropy is a common measure of randomness.
- Maximum Shannon entropy principle is equivalent to minimizing KL divergence with respect to uniform dist.

- Forward I-projection:

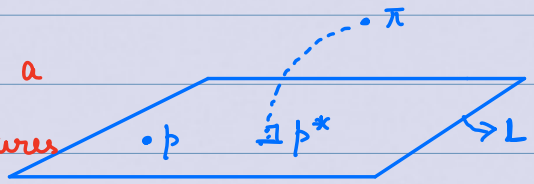
$$\min_{p \in L} D_{KL}(p \| \pi)$$

The minimiser is a member of the exponential family generated by $\pi$:

$$p_\lambda(x) \propto \pi(x) \exp\left(\sum_{i=1}^{k} \lambda_i T_i(x)\right).$$

- The I-projection satisfies Pythagorean property

$$D_{KL}(p \| \pi) = D_{KL}(p \| p^*) + D_{KL}(p^* \| \pi)$$




- Bregman's divergences is a class of divergence measures which satisfy Pythagorean property. KL divergence and Squared Euclidean distance are examples of Bregman's divergences.

- Reverse I-projection:

$$\min_{p \in L} D_{KL}(\pi \| p)$$

Maximum likelihood estimation is an instance of reverse I-projection.