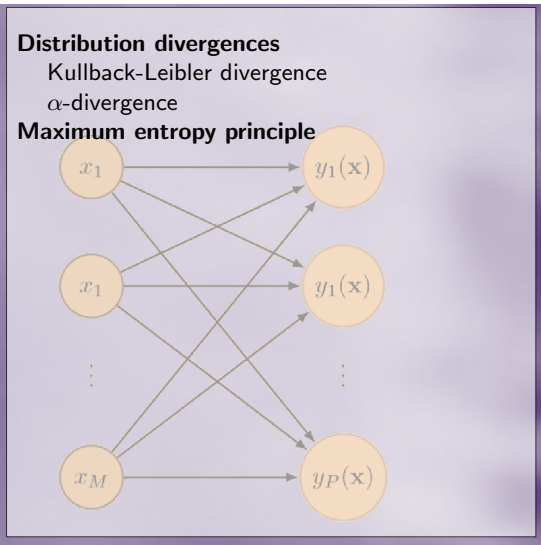

1 — Information theory for learning

1.1 Introduction

[TBW]

1.2 Distribution divergences

1.2.1 Kullback-Leibler divergence

One of the famous distance measure between functions is called Kullback-Leibler divergence or in short, KL-*divergence*. Let's assume we have two functions $f(x)$ and $g(x)$ defined on the same domain $\mathcal{X}$. The distance between these two functions using KL-*divergence* is,

$$\mathrm{KL}(f||g) = \int_{x\in\mathcal{X}} f \log \frac{f}{g} dx$$

it should be easy to see how to use this definition for discrete domains

$$\mathrm{KL}(f||g) = \sum_{x\in\mathcal{X}} f \log \frac{f}{g}$$

one could check that the above measure has the two following properties,

$$\mathrm{KL}(f||g) \geq 0 \tag{1.1a}$$

$$\mathrm{KL}(f||g) = 0 \iff f(x) = g(x), \forall x \in \mathcal{X} \tag{1.1b}$$

The second property shows that this measure isn't symmetric, i.e. $\mathrm{KL}(p||q) \neq KL(q||p)$; this it might not be an exact *distance* measure. You can check the main paper, **?** to see more details on KL-*divergence*.

1.2.2 α-divergence

α-divergence or alpha-divergence is a generalization of KL-divergence, which is defined as followiing:

$$D_\alpha(p||q) = \frac{\int \alpha p(\mathbf{x}) + (1-\alpha)q(\mathbf{x}) - p^\alpha(\mathbf{x})q^{1-\alpha}(\mathbf{x})d\mathbf{x}}{\alpha(1-\alpha)}$$

where $\alpha \in \mathbb{R}$. Some properties of this divergence are,

1. The α-divergence is convex with respect to $p(.)$ and $q(.)$
2. $D_\alpha(p||q) \geq 0$
3. $D_\alpha(p||q) = 0$ iff $p = q$. (This is easy to check from the definition)

For different values of α the above divergence has very interesting properties. Some of these these speciall cases are listed here:

1. $\lim_{\alpha\to 0} \mathrm{D}_\alpha(p||q) = \mathrm{KL}(q||p)$
2. $\lim_{\alpha\to 1} \mathrm{D}_\alpha(p||q) = \mathrm{KL}(p||q)$
3. $\mathrm{D}_{-1}(p||q) = \frac{1}{2}\int \frac{(q(\mathbf{x})-p(\mathbf{x}))^2}{q(\mathbf{x})} d\mathbf{x}$
4. $\mathrm{D}_2(p||q) = \frac{1}{2}\int \frac{(q(\mathbf{x})-p(\mathbf{x}))^2}{p(\mathbf{x})} d\mathbf{x}$
5. $\mathrm{D}_{\frac{1}{2}}(p||q) = 2\int \left(\sqrt{p(\mathbf{x})} - \sqrt{q(\mathbf{x})}\right)^2 d\mathbf{x}$
   It is easy to see that in this case when $\alpha = 0.5$ the divergence is symmetric. This is known as Hellinger distance. The square root of this divergence satisfies the triangle inequality.

1.3 Maximum entropy principle

The exponential distribution also arises naturally from the *maximum entropy principle*. The maximum entropy procedure consists of seeking the probability distribution which maximizes information entropy, subject to the constraints of the information. If we constrain the expected values of the sufficient statistics to be mean of the empirical mean of them under the sampled data, the resulting distribution which maximizes the entropy is the exponential family. To make the statement more accurate, let's express it in terms of formula. Given iid random variables, $\mathbf{x}_1, \ldots, \mathbf{x}_n \sim p^*$, where p^* is the underlying *unknown* distribution. Define a set of functions $\{\phi_i(.) : \mathcal{X} \to \mathbb{R}\}_{i\in\mathcal{I}}$, and consider the empirical mean of the sampled data under these functions,

$$\hat{\mu}_j = \frac{1}{n}\sum_{i=1}^{n} \phi_j(\mathbf{x}_i),\ \forall j \in \mathcal{I}$$

The problem is that we are looking for a distribution, $p(.)$ which has the same set of expected values for the defined functions as their empirical distribution, i.e.

$$\mathbb{E}_p\left[\phi_j(\mathbf{x})\right] = \int_{\mathcal{X}} \phi_j(\mathbf{x}) p(\mathbf{x}) \nu(d\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} \phi_j(\mathbf{x}_i) = \hat{\mu}_j,$$

and at the same time maximizes the entropy under this distribution,

$$H(p) = \int_{\mathcal{X}} -p(\mathbf{x}) \log p(\mathbf{x}) \nu(d\mathbf{x}).$$

This constrained optimization will result in the parametric form of the exponential family of distributions defined here. In other words,

$$\begin{cases} \max_p H(p) = \max_p \int_{\mathcal{X}} -p(\mathbf{x}) \log p(\mathbf{x}) \nu(d\mathbf{x}), \\ \text{such that, } \mathbb{E}_p\left[\phi_j(\mathbf{x})\right] = \hat{\mu}_j \end{cases}$$

The answer to the this constrained optimization problem is an exponential form,

$$\hat{p}(\mathbf{x};\boldsymbol{\theta}) \propto \exp\left\{\sum_{\alpha\in\mathcal{I}} \theta_\alpha \phi_\alpha(\mathbf{x})\right\}$$

■ **Lemma 1.1 — From Maximum Entropy to Exponential Families.** Finding maximum entropy (least informative) distribution with given expectation with respect to a set of functions, $\{\Phi(\mathbf{x})\}_i$, will result in the exponential families,

$$\begin{cases} \max E = \max \left\{ - \int_{\mathcal{X}} p(\mathbf{x}) \log p(\mathbf{x}) d\mathbf{x} \right\} \\ \mathbb{E}_p \left[\phi(\mathbf{x}) \right] = \mu \end{cases}$$

is equivalent to,

$$\max \left\{ \langle \mu, \theta \rangle - \mathcal{Z}(\theta) \right\}$$

where $\mathcal{Z}(\theta) = \int_{\mathcal{X}} \exp \left(\phi(\mathbf{x}), \theta \right) d\mathbf{x}$

Proof. This can be shown using Lagrange dual

$$\begin{aligned} L(p, \theta, \alpha, \beta) = & - \int_{\mathcal{X}} p(\mathbf{x}) \log p(\mathbf{x}) d\mathbf{x} \\ & + \alpha^\top \left(1 - \int_{\mathcal{X}} \exp \left(\phi(\mathbf{x}), \theta \right) d\mathbf{x} \right) \\ & + \beta^\top \left(\mu - \int_{\mathcal{X}} \phi(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} \right) \end{aligned}$$

Taking functional derivative with respect to $p(.)$,

$$\frac{\partial L}{\partial p} = - \log p(\mathbf{x}) + \beta + 1 + \theta^\top \phi(\mathbf{x}) = 0$$

$$\Rightarrow p(\mathbf{x}) = \exp \left(\theta^\top \phi(\mathbf{x}) + \beta + 1 \right)$$

$$\Rightarrow p(\mathbf{x}) = \frac{1}{\mathcal{Z}(\theta)} \exp \left(\theta^\top \phi(\mathbf{x}) \right), \ \mathcal{Z}(\theta) = \int_{\mathcal{X}} \exp \left(\phi(\mathbf{x}), \theta \right) d\mathbf{x}$$

■