Foundations of Machine Learning II
Course 4*

Guillaume Charpiat & Gaétan Marceau Caron

This course is about information geometry and Fisher information.

Fisher information Let $\mathcal{M}\theta)$ be a model with parameters θ and then, we define the Fisher information:

$$J(\theta) := \mathbb{E}_{x\sim P_\theta} \frac{\partial^2 - \log p_\theta(x)}{\partial\theta^2} \quad \text{(Average Hessian under law } x \sim p_\theta) \tag{1}$$

$$= \mathbb{E}_{x\sim P_\theta} \frac{\partial - \log p_\theta(x)}{\partial\theta} \frac{\partial - \log p_\theta(x)}{\partial\theta} \quad \text{(covariance of the gradient)} \tag{2}$$

$$\tag{3}$$

To prove the equality, we look at the gradient:

$$\frac{\partial \log p_\theta(x)}{\partial\theta}\delta\theta = \frac{1}{p_\theta(x)} \frac{\partial p_\theta(x)}{\partial\theta}\delta\theta \tag{4}$$

and the hessian:

$$\frac{\partial^2 \log p_\theta(x)}{\partial\theta^2}(\delta\theta)(\delta\theta) = -\frac{\delta\theta}{p_\theta(x)^2} \frac{\partial p_\theta(x)}{\partial\theta} \frac{\partial p_\theta(x)}{\partial\theta}^\top \delta\theta + \frac{1}{p_\theta(x)} \frac{\partial^2 \log p_\theta(x)}{\partial\theta^2} \tag{5}$$

then, by doing a Taylor expansion and commuting the integral with the partial derivative, the terms of order superior to 2 equal zero.

Geometry For two probability distribution P_θ and $P_{\theta'}$, we define the KL divergence for defining the notion of metric and distances. To prove the link between KL and $J(\theta)$, we need the following identities:

1. $\ln(1+z) = z - \frac{1}{2}z^2 + \mathcal{O}(z^3)$
2. $\frac{\partial \ln f(z)}{\partial z} = \frac{1}{f(z)} \frac{\partial f(z)}{\partial z}$

3. $\int_x p_\theta(x) = 1 \forall\theta \int \frac{\partial_\theta^k p_\theta(x)}{\partial\theta^k} dx = 0 \mathrm{P}_{\theta+\delta\theta}(x) = p_\theta(x) + \delta\theta \frac{\partial p_\theta}{\partial\theta}(x) + \frac{1}{2}\delta\theta \frac{\partial p_\theta(x)}{\partial\theta^2}\delta\theta + \mathcal{O}(\delta\theta^3)$ Now, we look at

$$\mathrm{KL}((\,\|\,p)_{\theta+\delta\theta}\,||p_\theta) = \int_x p_{\theta+\delta\theta} \ln \frac{p_{\theta+\delta\theta}(x)}{p_\theta(x)} dx \tag{6}$$

*https://www.lri.fr/~gcharpia/machinelearningcourse/

If you develop the expression at most order 3, then we obtain naturally the Fisher metric and its interpretation as a measure of curvature (entropy curvature). Suppose that you have a dataset (x_i), we want to find θ s.t. $x \sim p_\theta$. Then, we can approximate the expectation with the dataset.

Cramer-Rao bound Suppose that we have observations $(\mathrm{x}_i), we want to find the best \theta$. Let define some properties: For any unbiased estimator : $\mathbb{E}_{x \sim p_\theta}\left[\hat{\theta}\right] = \theta$, we have the Cramer-Rao bound:

$$Var(\text{any unbiased estimator}) \geqslant \frac{1}{J(\theta)} \tag{7}$$

The proof of the Cramer-Rao bound uses the Cauchy-Schwartz inequality. Let $T = \hat{\theta} - \theta$, then we have:

$$\mathbb{E}_{x \sim p_\theta}\left[\frac{\partial \log p_\theta}{\partial \theta} T\right]^2 \leqslant \mathbb{E}_{x \sim p_\theta}\left[\frac{\partial \log p_\theta}{\partial \theta}\right]^2 \mathbb{E}_{x \sim p_\theta}\left[T^2\right] \tag{8}$$

There is a bug with the proof.

Natural gradient descent Recall the iteration scheme of the gradient descent algorithm:

$$\theta_{i+1} = \theta_i - \eta_\theta f(\theta) \tag{9}$$

Let's look at the directional derivative $Df(\theta)(\delta\theta) =<_\theta f(\theta)|\delta\theta >_M$ (Riesz representation theorem). Moreover, we have that the norm is defined with the inner product $||\delta\theta||_M^2 =< \delta\theta|\delta\theta >_M$. Consequently, the gradient step is determined by the metric M chosen to represent our parameters. We can generalize the inner product as $< \delta\theta||\delta\theta >_M = \delta\theta M \delta\theta$ for a semi-definite matrix M and then obtain the general gradient:

$$\nabla_M := M^{-1} \nabla_{L_2} \tag{10}$$

For the natural gradient, we take $M = J(\theta)$. We can define the metric as the following:

$$||\delta\theta||_M^2 := ||\delta f||^2 \tag{11}$$

where $f(\theta) = -\log p_\theta(x)$. By comparing two gradients, we can easily show that the natural gradient is invariant.

the case of the exponential family the exponential family is defined as

$$p_\theta(x) = \frac{1}{z_\theta} \exp(\theta \cdot d(x)) \tag{12}$$

Bernoulli and Gaussian distributions belong to the exponential family.

proposition: for exponential family, we have $\nabla_{nat} = Hessian$. (natural gradient is the newton method) One important point is that $\frac{\partial \log p_\theta(x)}{\partial \theta}$ does not depend on x. BUT, Newton is not appropriate for non-convex optimization. Finally, the estimator trained with ∇_{nat} descent can reach the Cramer-Rao bound.

Model Selection: universal coding We have the model p_θ and we want to encode the dataset (x_i). Remember that we have the Kolmogorov complexity:

$$K(\theta, (x_i)) = K(\theta) - \log p_\theta((x_i)) \tag{13}$$

There are at least four possible way to approximate the Kolmogorov complexity:

1. go through all data, find best θ, encode θ and encode $x|\theta$. But can we encode θ cheaply without losing too much
2. θ_0 is predefined, then we observe x_1 and we choose $\theta_1 = argmaxp_\theta(x_1)$, then we repeat for x_2 and we choose $\theta_2 = argmaxp_\theta(x_1, x_2)$. The pros are that there is no encoding of parameter but the cons is that the first parameters are clearly not optimal.
3. we can define the $NML(x) = \frac{p_{\text{best } \theta \text{ for x}}(x)}{\sum_z p_{\text{best } \theta \text{ for z}}(z)}$ (Normalized Maximum Likelihood)

4. Bayesian: p(x) $= \int_\theta p(\theta)p(x|\theta)dx$ NML is not good since we do not have $NML(x, y) \neq NML(y)NML(y|x)$. A better choice is to consider the Bayesian framework since we have

$$p(x) = \int_\theta p(\theta)p(x|\theta)d\theta \tag{14}$$

$$\geqslant p(\theta^*)p(x|\theta*) \tag{15}$$

where θ^* is the best parameter to encode our data.

Parameter precision We want to encode $\theta\varepsilon$ with k bits s.t. $\theta 2^{-k}$ From Kolmogorov, we have

$$K(\theta, x) = K(\theta) - \log p_\theta(x) \tag{16}$$

$$= -\log epsilon - \log p_{\theta^*+\varepsilon}(x) \tag{17}$$

$$= -\log \varepsilon - \log p_{\theta^*}(x) - \frac{1}{2}\varepsilon^2 \frac{\partial^2 \log p_\theta}{\partial \theta^2}_{|\theta=\theta^*} \tag{18}$$

By taking the derivative w.r.t. ε equals zero, we obtain $\varepsilon = 1\sqrt{J(\theta)}$ For a whole dataset of size n, we have $\varepsilon = 1\sqrt{n}1\sqrt{J(\theta)}$. This can be related to the Bayesian Information Criterion (BIC).

Prior by default: Jeffrey's prior If we do not have any idea on the prior, we can naturally choose the uniform prior. But, if the parameters lie on the whole $\mathbb{R}$, this is not defined. According to the previous result on ε, we should sample more over regions where the model varies a lot when the paramters move: $q(\theta)\sqrt{I(\theta)}$. As an example, for a Bernoulli distribution, we have $I(\theta) = \theta(1-\theta)$ and so, the Jeffrey's prior is $q(\theta) = \frac{1}{\pi}\frac{1}{\sqrt{\theta(1-\theta)}}$.

Context Tree Weighting For text prediction, the CTW gives a probability for all possible Markov chain orders.

$$\sum_{\text{Markov Chain}} 2^{-\text{order}} \int_\theta p_{\text{Jeffrey}}(\theta) p(x|\theta) \tag{19}$$