§ 1. Fisher Information Matrix.

Assume we have a distribution $p(x|\theta)$ parameterized by θ. We consider the gradient of its log likelihood

$$S(\theta) = \nabla_\theta \log p(x|\theta)$$

as the score function.

Proposition 1. The expectation of the score function w.r.t. our model is 0.

Proof.

$$\begin{aligned}
&\mathbb{E}_{x\sim p(\cdot|\theta)}[\nabla_\theta \log p(x|\theta)] \\
&= \int_{\mathcal{X}} p(x|\theta)\cdot \nabla_\theta \log p(x|\theta)\, dx \\
&= \int_{\mathcal{X}} p(x|\theta)\cdot \frac{\nabla_\theta p(x|\theta)}{p(x|\theta)}\, dx \\
&= \int_{\mathcal{X}} \nabla_\theta p(x|\theta)\, dx \\
&= \nabla_\theta \int_{\mathcal{X}} p(x|\theta)\, dx \\
&= \nabla_\theta 1 \\
&= 0.
\end{aligned}$$

∎

Definition 2. The covariance of the score of our model is defined as the Fisher Information Matrix:

$$F = \mathbb{E}_{x\sim p(\cdot|\theta)}[\nabla_\theta \log p(x|\theta)\cdot \nabla_\theta \log p(x|\theta)^T].$$

Remark. When it is untractable to compute the expectation,

F could be replaced by its empirical version,

$$\hat{F} = \frac{1}{N} \cdot \sum_{i=1}^{N} \nabla_\theta \cdot \log p(x_i|\theta) \cdot \nabla_\theta \cdot \log p(x_i|\theta)^T,$$

which is given by our training data $X = \{x_1, x_2, \ldots, x_N\}$.

Proposition 3. The negative expected Hessian matrix of the log likelihood is the FIM F.

Proof. $H_{\log p(x|\theta)} = J(\nabla_\theta \cdot \log p(x|\theta))$

$$= J\left(\frac{\nabla_\theta \cdot p(x|\theta)}{p(x|\theta)}\right)$$

$$= \frac{(\nabla_\theta^2 \cdot p(x|\theta)) \cdot p(x|\theta) - \nabla_\theta \cdot p(x|\theta) \cdot \nabla_\theta \cdot p(x|\theta)^T}{p^2(x|\theta)}$$

$$= \frac{H_{p(x|\theta)}}{p(x|\theta)} - \left(\frac{\nabla_\theta \cdot p(x|\theta)}{p(x|\theta)}\right)\left(\frac{\nabla_\theta \cdot p(x|\theta)}{p(x|\theta)}\right)^T.$$

Taking expectation on both sides leads to

$$\mathbb{E}_{x\sim p(\cdot|\theta)}[H_{\log p(x|\theta)}]$$

$$= \mathbb{E}_{x\sim p(\cdot|\theta)}\left[\frac{H_{p(x|\theta)}}{p(x|\theta)} - \left(\frac{\nabla_\theta \cdot p(x|\theta)}{p(x|\theta)}\right)\left(\frac{\nabla_\theta \cdot p(x|\theta)}{p(x|\theta)}\right)^T\right]$$

$$= \int_{\mathcal{X}} p(x|\theta) \cdot \frac{H_{p(x|\theta)}}{p(x|\theta)}\, dx - \mathbb{E}_{x\sim p(\cdot|\theta)}[\nabla_\theta \cdot \log p(x|\theta) \cdot \nabla_\theta \cdot \log p(x|\theta)^T]$$

$$= H \cdot \int_{\mathcal{X}} p(x|\theta)\, dx - F$$

$$= H \cdot 1 - F$$

$= -F.$

□

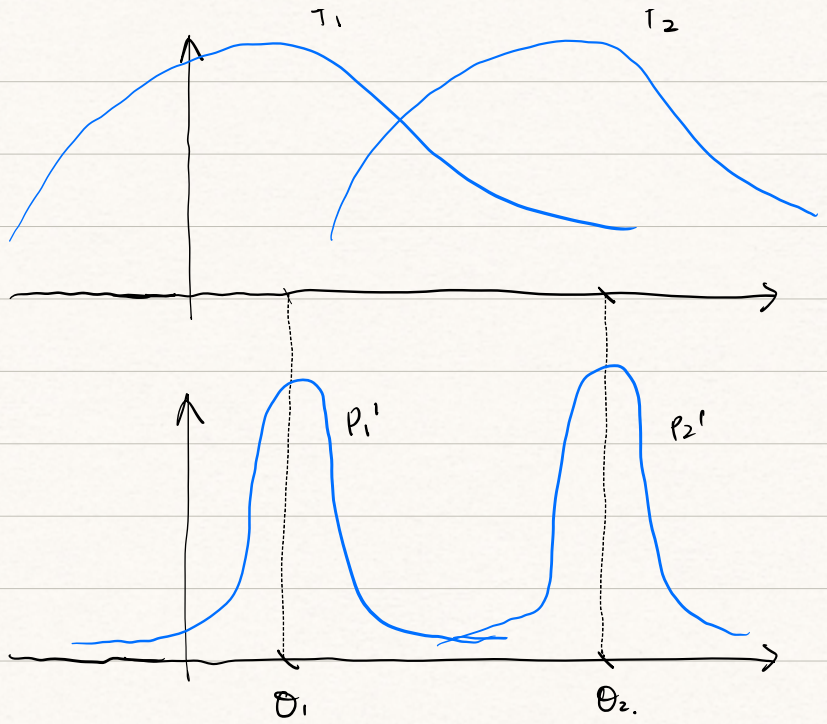
§2. Distribution Space

We would like to maximize the likelihood function to find the most likely parameter θ. Equivalent formulation would be to minimize the loss function $L(\theta)$, which is the negative log likelihood function.

Usual way to solve this optimization is to use gradient descent. In this case, we take a step in the steepest descent around the local neighbourhood of current value of parameter θ, in the parameter space. Formally, we have

$$\frac{-\nabla_\theta L(\theta)}{\|-\nabla_\theta L(\theta)\|} = \lim_{\varepsilon \to 0} \operatorname*{argmin}_{d : \|d\| \le \varepsilon} L(\theta + d).$$

Thus, the optimization in gradient descent depends on the Euclidean geometry of the parameter space. Our purpose is correct, but our approach could be more appropriate, with the fact that two distributions might not be "close" to each other even though their parameters are "close" to each other in the parameter space. Consider the motivating example below:



Though the Euclidean distance between the mean of P_1 and P_2 and the Euclidean distance between the mean of P_1' and P_2', are the same, it is clear that P_1 and P_2 are closer than P_1' and P_2'.

To this end we consider directly optimizing in the distribution space instead of the parameter space. A natural choice is to use KL-divergence as the metric in this distribution space, though KL-divergence is not symmetric and thus not a metric. Actually, it could be shown that as d goes to zero, KL-divergence is asymtotically symmetric.

Proposition 4. FIM $\mathbf{F}$ is the Hessian of KL-divergence between two distributions $p(x|\theta)$ and $p(x|\theta')$, w.r.t. θ', evaluated at $\theta'=\theta$.

Proof. $KL(p(x|\theta), p(x|\theta')) = \mathbb{E}_{p(x|\theta)}[\log p(x|\theta)] - \mathbb{E}_{p(x|\theta)}[\log p(x|\theta')]$

$$\nabla_{\theta'} KL(p(x|\theta), p(x|\theta')) = -\int_{x\sim p(\cdot|\theta)} p(x|\theta) \cdot \nabla_{\theta'} \log p(x|\theta') \cdot dx$$

$$\nabla^2_{\theta'} KL(p(x|\theta), p(x|\theta'))\Big|_{\theta'=\theta} = -\int_{x\sim p(\cdot|\theta)} p(x|\theta) \cdot \nabla^2_{\theta'} \log p(x|\theta')\Big|_{\theta'=\theta} dx$$

$$= -\int_{x\sim p(\cdot|\theta)} p(x|\theta) \cdot H_{\log p(x|\theta)} \cdot dx.$$

$$= F.$$

Before we use the FIM to enhance the gradient descent, we need to derive the Taylor series expansion for KL-divergence around θ.

Proposition 5. Let $d \to 0$. The second order Taylor series expansion of KL-divergence is $KL(p(x|\theta), p(x|\theta+d)) \approx \frac{1}{2}\|d\|_F^2$.

Proof. $KL(p(x|\theta), p(x|\theta+d))$

$$= KL(p(x|\theta), p(x+\theta)) + (\nabla_{\theta'} KL(p(x|\theta), p(x|\theta'))|_{\theta'=\theta})^T d + \frac{d^T(\nabla^2_{\theta'} KL(p(x|\theta), p(x|\theta'))|_{\theta'=\theta})d}{2}$$

$$= 0 + 0 + \frac{1}{2}\|d\|_F^2.$$

$$= \frac{1}{2}\|d\|_F^2.$$

Now we would like to minimize the loss function $\mathcal{L}(\theta)$ in distribution space with the KL-divergence as the metric. Specifically, we do the minimization

$$d^* = \underset{d:\, KL(p(x|\theta),\, p(x|\theta+d))=c}{\operatorname{argmin}} \mathcal{L}(\theta+d). \tag{1}$$

where C is a constant. The purpose of fixing the KL-divergence to some constant is to make the model more robust to the reparameterization of the model, i.e., the algorithm does not care how the model is parameterized, it only cares the distribution induced by the parameter.

If we write Eq(1) in its Lagrangian form, with $\mathcal{L}(\theta+d)$ approximated by the first order Taylor series expansion and KL-divergence approximated by the second order Taylor series expansion, we have

$$d^* = \underset{d}{\operatorname{argmin}}\ \mathcal{L}(\theta+d) + \lambda\cdot\left(KL(p(x|\theta),\, p(x|\theta+d)) - c\right)$$

$$\approx \underset{d}{\operatorname{argmin}}\ \mathcal{L}(\theta) + \nabla_\theta \mathcal{L}(\theta)^T \cdot d + \frac{\lambda}{2}\|d\|_F^2 - \lambda c. \tag{2}$$

Setting the derivative of Eq (2) w.r.t. θ to 0 leads to

$$d = -\frac{1}{\lambda}\cdot F^{-1}\cdot \nabla_\theta \mathcal{L}(\theta),$$

where the constant factor could be absorbed into the learning rate.

Definition. 6. Natural gradient is defined as

$$\tilde{\nabla}_\theta \mathcal{L}(\theta) = I^{-1} \cdot \nabla_\theta \mathcal{L}(\theta).$$