Fisher Information and Its Application

Patrick Sylvestre
Tsinghua University
February 13, 2023

Abstract

本文主要介绍 Fisher 信息的定义、性质以及其计算与应用。

让我们先从一个问题开始。假定我们有来自某分布的若干样本，我们并不知道真实的分布是如何的，不过根据数据特点我们可以假设其分布形式（现在分布形式已知），但分布中的参数未知。我们需要根据现有的样本去估计分布的参数（参考数理统计中的估计理论），与此同时产生的另一个问题是：在给定样本的情况下，估计参数的难度应该如何度量?

我们以高斯过程为例来辅助回答这个问题：假定已知某随机变量 X 服从高斯分布 $\mathcal{N}(\mu, \sigma^2)$，但仅知方差 σ^2 而不知均值 μ，我们的目标是通过样本估计均值 μ。首先我们要问的一个问题是，对于不同的参数值来说，估计的难度是相同的吗? 如图1所示，直观来看，对于不同的方差，参数估计的难度是不一样的。对于方差最小的高斯分布，其均值是最容易估计的。因为当方差较小时，随机样本比方差较大时更可能接近均值[1]。抽象地说，样本包含的参数信息越多，就越容易估计。相反，样本中包含的参数信息越少，就越难以估计。在这里，样本的方差越小，其携带关于均值的信息越多。其次，我们应该用什么样的工具来衡量样本携带的关于参数的信息呢? Fisher 信息[2]是衡量随机样本包含未知参数多少信息量最广泛的解决方案（并不唯一）。原因如下：第一，它的定义具有信息量 [6] 的特征；第二，它的性质优良。接下来，我们介绍 Fisher 信息的正式定义。

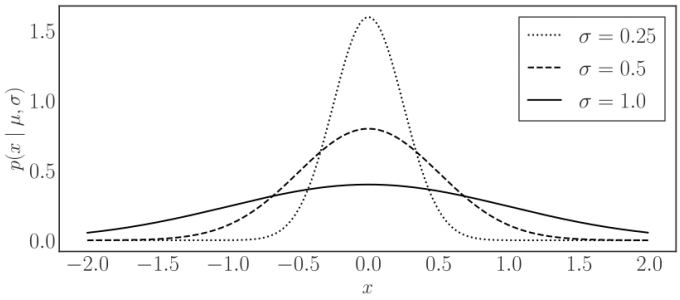

Figure 1: 三个不同方差的高斯分布 $p(x|\mu, \sigma)$. 随着方差的增加, 样本携带的关于分布均值的信息越来越少.

1 Fisher Information and Its Properties

1.1 Univariate Case

若关于随机变量 X 的条件分布 $p(x|\theta)$ （θ 是分布的参数）满足下列条件 [10]：

- 参数空间 Θ 是 $\mathbb{R}$ 中的开区间；

[1]这个表述在此是不严谨的，但是此处只需要唯象的理解。
[2]由剑桥大学 Ronald Aylmer Fisher 爵士提出。

- 支撑集 $S = \{x|p(x|\theta) > 0\}$ 与 θ 无关；
- $\frac{\mathrm{d}p(x|\theta)}{\mathrm{d}\theta}$ 对 $\theta \in \Theta$ 几乎处处存在；
- 关于 $p(x|\theta)$ 的微分与积分运算可交换次序，i.e.

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\int_{-\infty}^{\infty} p(x|\theta)\,\mathrm{d}x = \int_{-\infty}^{\infty} \frac{\mathrm{d}}{\mathrm{d}\theta} p(x|\theta)\,\mathrm{d}x \tag{1}$$

- 期望 $\mathbb{E}_x\left[\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\ln p(x|\theta)\right)^2\right]$ 存在。

那么我们可以定义 Fisher 信息如下：

定义 1 (Fisher Information)**.** 对于一元随机变量，样本点 x 在待估计参数 θ 处的 *Fisher* 信息定义为：

$$I_X(\theta) \triangleq \mathbb{E}_x\left[\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\ln p(x|\theta)\right)^2\right]. \tag{2}$$

其中，条件分布 $p(x|\theta)$ 称为**似然函数**[3][4] *(likelihood function)*，有时也记为 $p_\theta(x)$，$\ell(\theta) \triangleq \ln p(x|\theta)$ 称为**对数似然函数** *(log-likelihood function)*，对数似然函数的导数 $\ell'(\theta)$ 称为**品质函数** *(score function)*。那么 *Fisher* 信息定义可以改写为[5]：

$$I_X(\theta) \triangleq \mathbb{E}_x\left[(\ell'(\theta))^2\right].$$

Fisher 信息试图量化随机变量 X 对参数 θ 值的敏感性。如果 θ 的较小变化导致 X 的可能值的较大变化（i.e. 条件分布 $p(x|\theta)$ 的较大变化），那么我们很容易从样本推断 θ 的真实值，换句话说观察到的样本告诉了我们很多关于 θ 的信息。反之，则不然。这与我们在图1得到的直观印象是吻合的。

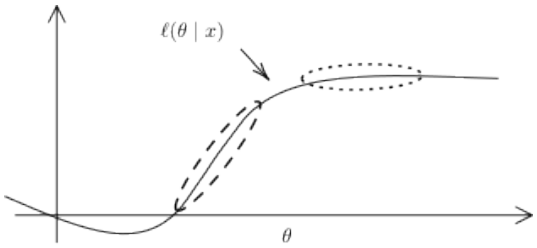

Figure 2: 对于某样本点 x，其对应的对数似然函数 $\ell(\theta|x)$ 的示例. 显然，对于长虚线包围的部分相比于短虚线包围的部分的具有更高的关于 θ 的信息量.

例 1 (Fisher Information of Fixed Variance Gaussian Distribution)**.** 考虑高斯分布 $X \sim \mathcal{N}(\mu, \sigma^2)$，但仅知方差 σ^2 而不知均值 μ。其中：

$$p(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right].$$

那么其对数似然函数和 *score function* 直接有：

$$\ell(\mu|x,\sigma) = \ln p(x|\mu,\sigma) = -\ln(\sqrt{2\pi}\sigma) - \frac{(x-\mu)^2}{2\sigma^2}$$
$$\ell'(\mu|x,\sigma) = \frac{\mathrm{d}\ell(\mu|x,\sigma)}{\mathrm{d}\mu} = \frac{(x-\mu)}{\sigma^2}.$$

[3]从这个角度看，Fisher 信息描述了当下的似然函数能有多大的潜力寻找出最优参数。

[4]为什么与似然概念产生联系?

[5]可以证明，Fisher 信息为 score 的方差，所以 $I_X(\theta) \geq 0$。Fisher 信息越高，其 score 函数的绝对值越大。

由定义知，高斯分布在待估计参数 μ 处的 *Fisher* 信息为：

$$I_X(\mu) = \mathbb{E}_x\left[(\ell'(\theta))^2\right] = \mathbb{E}_x\left[\left(\frac{(x-\mu)}{\sigma^2}\right)^2\right] = \frac{1}{\sigma^4}\mathbb{E}_x\left[(x-\mu)^2\right] = \frac{1}{\sigma^2}.$$

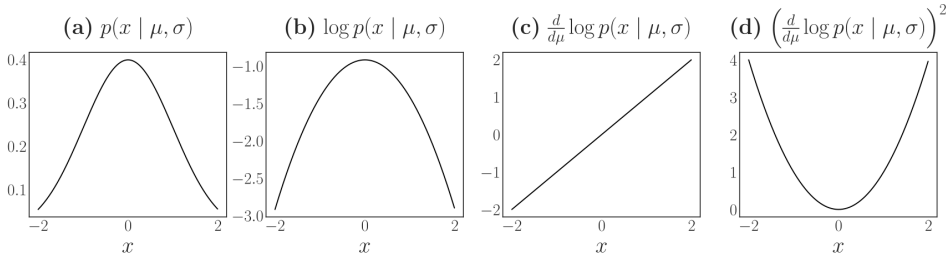

Figure 3: 高斯分布 Fisher 信息量计算的步骤图例.

例 2 (Fisher Information for Poisson Distribution)**.** 考虑泊松分布，其中：

$$p(x|\lambda) = \frac{\lambda^x}{x!}e^{-\lambda} \qquad x \in \mathbb{N}^+ \tag{3}$$

那么有：

$$\begin{aligned} \ell(\lambda) &= \ln p(x|\lambda) = x\ln\lambda - \lambda - \ln(x!) \\ \ell'(\lambda) &= \frac{\mathrm{d}}{\mathrm{d}\lambda}\ln p(x|\lambda) = \frac{x}{\lambda} - 1 \end{aligned} \tag{4}$$

于是

$$I_X(\lambda) = \mathbb{E}\left[(\frac{x}{\lambda} - 1)^2\right] = \frac{1}{\lambda} \tag{5}$$

例 3 (Fisher Information for Exponential Distribution)**.** 考虑指数分布，其概率密度函数为：

$$p(x|\theta) = \frac{1}{\theta}e^{-\frac{x}{\theta}} \qquad x > 0, \theta > 0 \tag{6}$$

那么有：

$$\begin{aligned} \ell(\theta) &= \ln p(x|\theta) = -\ln\theta - \frac{x}{\theta} \\ \ell'(\theta) &= -\frac{1}{\theta} + \frac{x}{\theta^2} \end{aligned} \tag{7}$$

于是

$$I_X(\theta) = \mathbb{E}\left[(\frac{x-\theta}{\theta^2})^2\right] = \frac{1}{\theta^2} \tag{8}$$

例 4 (Fisher Information for Binomial Distribution)**.** 考虑二项分布 $X \sim B(N, p)$，其分布律为：

$$f(x|p) = \binom{N}{x}p^x(1-p)^{N-x} \qquad x = 0, 1, \ldots, N \tag{9}$$

那么其 *Fisher* 信息为：

$$I_X(p) = \frac{1}{p^2(1-p)^2}\sum_{x=0}^{N}(x-Np)^2\binom{N}{x}p^x(1-p)^{N-x} \tag{10}$$

此结论需要用到在离散情况下 *Fisher Information* 定义的一个恒等变形，在此留作练习。

类似于熵，Fisher 信息也有链式法则，结论如下：

性质 1 (Chain Rule of Fisher Information[8])**.** *若 X, Y 为两个随机变量，且存在联合分布。那么对于同一个待估计参数 θ，有：*

$$I_{X,Y}(\theta) = I_X(\theta) + I_{Y|X}(\theta). \tag{11}$$

其中：$I_{Y|X}(\theta) = \mathbb{E}_x\left[I_{Y|X=x}(\theta)\right]$ 为条件分布的 Fisher 信息。特别地，如果 X, Y 独立，那么有：

$$I_{X,Y}(\theta) = I_X(\theta) + I_Y(\theta). \tag{12}$$

这直接说明，两个独立的随机变量含有的信息量是它们分别含有的信息量的总和。特别地，n 个独立同分布样本观测值的信息是样本量大小为 1 的观测值包含的信息的 n 倍。

Proof. 注意到联合分布 $p_\theta(x, y) = p_\theta(x)p_\theta(y|x)$，那么：

$$\begin{aligned}
I_{X,Y}(\theta) &= \mathbb{E}_{x,y}\left[\left(\frac{\mathrm{d}\ln p_\theta(x,y)}{\mathrm{d}\theta}\right)^2\right] = \mathbb{E}_{x,y}\left[\left(\frac{\mathrm{d}\ln p_\theta(x)}{\mathrm{d}\theta} + \frac{\mathrm{d}\ln p_\theta(y|x)}{\mathrm{d}\theta}\right)^2\right] \\
&= \mathbb{E}_x\left[\left(\frac{\mathrm{d}\ln p_\theta(x)}{\mathrm{d}\theta}\right)^2\right] + \mathbb{E}_{y|x}\left[\left(\frac{\mathrm{d}\ln p_\theta(y|x)}{\mathrm{d}\theta}\right)^2\right] + 2\mathbb{E}_{x,y}\left[\frac{\mathrm{d}\ln p_\theta(x)}{\mathrm{d}\theta}\frac{\mathrm{d}\ln p_\theta(y|x)}{\mathrm{d}\theta}\right] \\
&= I_X(\theta) + I_{Y|X}(\theta) + 2\mathbb{E}_x\left[\frac{\mathrm{d}\ln p_\theta(x)}{\mathrm{d}\theta}\mathbb{E}_y\left[\frac{\mathrm{d}\ln p_\theta(y|x)}{\mathrm{d}\theta}\middle| X = x\right]\right] \\
&\overset{\text{交换微分积分号}}{=} I_X(\theta) + I_{Y|X}(\theta).
\end{aligned}$$

□

性质 2 (Data Processing Inequality[9][6])**.** *对任意随机变量 X 和任意函数 $f(X)$，那么对于同一个待估计参数 θ，有：*

$$I_{f(X)}(\theta) \le I_X(\theta). \tag{13}$$

iff. 在 $f(x) = x$ 取等。

Proof.

$$I_{f(X)}(\theta) \overset{\text{性质1}}{=} I_{f(X),X}(\theta) - I_{X|f(X)}(\theta) \le I_{f(X),X}(\theta) = I_X(\theta).^7$$

□

1.2 Multivariable Case

对于多个参数的分布，Fisher 信息变为 Fisher 信息矩阵。该推广是自然的。

定义 2 (Fisher Information Matrix)**.** *考虑参数分布簇 $\{p_{\boldsymbol{\theta}}\}$, $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^n$, 记随机向量:*

$$\nabla_{\boldsymbol{\theta}} \ln p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \left(\frac{\partial \ln p_{\boldsymbol{\theta}}(\boldsymbol{x})}{\partial\theta_1}, \cdots, \frac{\partial \ln p_{\boldsymbol{\theta}}(\boldsymbol{x})}{\partial\theta_n}\right)^{\mathrm{T}}.$$

满足:

1. *$\forall \boldsymbol{\theta} \in \boldsymbol{\Theta}$, $\nabla_{\boldsymbol{\theta}} \ln p_{\boldsymbol{\theta}}(\boldsymbol{x})$ 有定义;*
2. *对 $p_{\boldsymbol{\theta}}$ 微分记积分可交换且支撑集 $\{\boldsymbol{x} : p_{\boldsymbol{\theta}}(\boldsymbol{x}) > 0\}$ 与 $\boldsymbol{\theta}$ 无关[8];*
3. *其二阶矩存在: $\mathbb{E}_{\boldsymbol{x}}[\|\nabla_{\boldsymbol{\theta}} \ln p_{\boldsymbol{\theta}}(\boldsymbol{x})\|^2] < \infty$;*

则 Fisher Information Matrix 定义如下:

$$I_{\boldsymbol{X}}(\boldsymbol{\theta}) \triangleq \mathbb{E}_{\boldsymbol{x}}[\nabla_{\boldsymbol{\theta}} \ln p_{\boldsymbol{\theta}}(\boldsymbol{x}) \nabla_{\boldsymbol{\theta}} \ln p_{\boldsymbol{\theta}}(\boldsymbol{x})^{\mathrm{T}}] \in \mathbb{R}^{n\times n}. \tag{14}$$

性质 3. *Fisher Information Matrix 有如下相关性质与等价形式[9]:*

[6]注意此处有别于信息论中的数据处理不等式。
[7]此处需要简单用到 Data Refinement Inequality。
[8]指数分布簇一般均满足该条件.
[9]对于一元的情况，有相似的结论，请自行证明。

1. $\mathbb{E}_{\boldsymbol{x}}[\nabla_{\boldsymbol{\theta}} \ln p_{\boldsymbol{\theta}}(\boldsymbol{x})] = \mathbf{0}$, $I_{\boldsymbol{X}}(\boldsymbol{\theta}) = \mathrm{Cov}_{\boldsymbol{x}}[\nabla_{\boldsymbol{\theta}} \ln p_{\boldsymbol{\theta}}(\boldsymbol{x})]$;
2. 若 $p_{\boldsymbol{\theta}}$ 二阶可导，则 $I_{\boldsymbol{X}}(\boldsymbol{\theta}) = -\mathbb{E}_{\boldsymbol{x}}[\nabla^2_{\boldsymbol{\theta}} \ln p_{\boldsymbol{\theta}}(\boldsymbol{x})]$;
3. $I_{\boldsymbol{X}}(\boldsymbol{\theta})$ 对称且半正定;

Proof. 分别证明如下.

1. 考察:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{x}}[\nabla_{\boldsymbol{\theta}} \ln p_{\boldsymbol{\theta}}(\boldsymbol{x})] &= \int p_{\boldsymbol{\theta}}(\boldsymbol{x}) \nabla_{\boldsymbol{\theta}} \ln p_{\boldsymbol{\theta}}(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x} \\ &= \int p_{\boldsymbol{\theta}}(\boldsymbol{x}) \frac{\nabla_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}}(\boldsymbol{x})}{p_{\boldsymbol{\theta}}(\boldsymbol{x})} \,\mathrm{d}\boldsymbol{x} \\ &\overset{\text{定义2性质 2}}{=} \nabla_{\boldsymbol{\theta}} \int p_{\boldsymbol{\theta}}(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x} = \nabla_{\boldsymbol{\theta}} 1 = \mathbf{0}.\end{aligned}$$

   再由协方差定义知:

$$\begin{aligned}I_{\boldsymbol{X}}(\boldsymbol{\theta}) &= \mathbb{E}_{\boldsymbol{x}}[\nabla_{\boldsymbol{\theta}} \ln p_{\boldsymbol{\theta}}(\boldsymbol{x}) \nabla_{\boldsymbol{\theta}} \ln p_{\boldsymbol{\theta}}(\boldsymbol{x})^{\mathrm{T}}] \\ &= \mathbb{E}_{\boldsymbol{x}}[(\nabla_{\boldsymbol{\theta}} \ln p_{\boldsymbol{\theta}}(\boldsymbol{x}) - \mathbb{E}_{\boldsymbol{x}}[\nabla_{\boldsymbol{\theta}} \ln p_{\boldsymbol{\theta}}(\boldsymbol{x})])(\nabla_{\boldsymbol{\theta}} \ln p_{\boldsymbol{\theta}}(\boldsymbol{x}) - \mathbb{E}_{\boldsymbol{x}}[\nabla_{\boldsymbol{\theta}} \ln p_{\boldsymbol{\theta}}(\boldsymbol{x})])^{\mathrm{T}}] \\ &= \mathrm{Cov}_{\boldsymbol{x}}[\nabla_{\boldsymbol{\theta}} \ln p_{\boldsymbol{\theta}}(\boldsymbol{x})].\end{aligned}$$

2. 考察:

$$\begin{aligned}\nabla^2_{\boldsymbol{\theta}} \ln p_{\boldsymbol{\theta}}(\boldsymbol{x}) &= \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}) \nabla^2_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}}(\boldsymbol{x}) - \nabla_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}}(\boldsymbol{x}) \nabla_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}}(\boldsymbol{x})^{\mathrm{T}}}{p^2_{\boldsymbol{\theta}}(\boldsymbol{x})} \\ &= \frac{\nabla^2_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}}(\boldsymbol{x})}{p_{\boldsymbol{\theta}}(\boldsymbol{x})} - \nabla_{\boldsymbol{\theta}} \ln p_{\boldsymbol{\theta}}(\boldsymbol{x}) \nabla_{\boldsymbol{\theta}} \ln p_{\boldsymbol{\theta}}(\boldsymbol{x})^{\mathrm{T}}.\end{aligned}$$

   对上式求期望:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{x}}[\nabla^2_{\boldsymbol{\theta}} \ln p_{\boldsymbol{\theta}}(\boldsymbol{x})] &= \int \nabla^2_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}}(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x} - \mathbb{E}_{\boldsymbol{x}}[\nabla_{\boldsymbol{\theta}} \ln p_{\boldsymbol{\theta}}(\boldsymbol{x}) \nabla_{\boldsymbol{\theta}} \ln p_{\boldsymbol{\theta}}(\boldsymbol{x})^{\mathrm{T}}] \\ &\overset{\text{定义2性质 2}}{=} \nabla^2_{\boldsymbol{\theta}} \int p_{\boldsymbol{\theta}}(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x} - I_{\boldsymbol{X}}(\boldsymbol{\theta}) \\ &= -I_{\boldsymbol{X}}(\boldsymbol{\theta}).\end{aligned}$$

3. 由定义是显然的.

□

Fisher 信息是数理统计中的一个基本概念，很多统计结果都与 Fisher 信息有关。比如最大似然估计的渐进方差；后验模态的渐进分布等。Cramér-Rao 不等式与 Fisher 信息量也有极为密切的关系。具体来说，C-R 不等式中有一个被称为 C-R 下界的统计量，它是有关参数的无偏估计的方差的下界。如果达到了这个下界，那么它就是 MVU 估计[10]。我们现在来介绍 C-R 下界。

定理 1 (Cramér-Rao Bound, CRB[11])**.** *考虑待估计的参数向量 $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^n$，记 $\boldsymbol{\theta}$ 估计量 $\hat{\boldsymbol{\theta}} = \boldsymbol{T}(X) = (T_1(X), \cdots, T_n(X))^{\mathrm{T}}$，其期望记为 $\varphi(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{x}}\left[\hat{\boldsymbol{\theta}}\right]$。那么估计量 $\hat{\boldsymbol{\theta}}$ 的协方差矩阵满足[12]:*

$$\mathrm{Cov}_{\boldsymbol{x}}[\hat{\boldsymbol{\theta}}] \geq \left(\frac{\partial \varphi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right) I^{-1}_{\boldsymbol{X}}(\boldsymbol{\theta}) \left(\frac{\partial \varphi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^{\mathrm{T}}. \tag{15}$$

[10] MVU 是 Minimum Variance Unbiased 的缩写。

[11] 以统计学家 Harald Cramér 和 C. R. Rao 的名字命名，但是也被 Maurice Fréchet、Georges Darmois、Alexander Aitken 和 Harold Silverstone 独立得出。

[12] 方阵 $\boldsymbol{A} \geq \boldsymbol{B}$ 表示方阵 $\boldsymbol{A} - \boldsymbol{B}$ 为半正定阵。

其中 $I_{\boldsymbol{X}}(\boldsymbol{\theta})$ 为 *Fisher* 信息矩阵，由定义2明确。

特别地，若 $\varphi(\boldsymbol{\theta})=\boldsymbol{\theta}$ 则称估计量 $\hat{\boldsymbol{\theta}}$ 为**无偏估计量** *(unbiased estimator)*。那么此时：

$$\frac{\partial\varphi(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}=\boldsymbol{I}_{n\times n},$$

退化为单位矩阵，因此：

$$\mathrm{Cov}_{\boldsymbol{x}}[\hat{\boldsymbol{\theta}}]\geq I_{\boldsymbol{X}}^{-1}(\boldsymbol{\theta}). \tag{16}$$

更特别地，若 $n=1$，且 $\varphi(\theta)=\theta$ 为无偏估计，那么有：

$$\mathrm{Var}_x[\hat{\theta}]\geq\frac{1}{I_X(\theta)}. \tag{17}$$

Proof. 严谨的证明需要使用 Chapman–Robbins bound，这里略去，详细可以参考 [4]。这里我们给出一个特殊情况的简单证明。

当 $n=1$ 时，记估计量 $\hat{\theta}=T(X)$，其期望记为 $\varphi(\theta)=\mathbb{E}_x\left[\hat{\theta}\right]$，考察协方差：

$$\begin{aligned}\mathrm{Cov}(\varphi(\theta),\nabla_\theta\ln p_\theta(x))&=\mathbb{E}_x\left[\varphi(\theta)\nabla_\theta\ln p_\theta(x)\right]-\mathbb{E}_x\left[\varphi(\theta)\right]\underbrace{\mathbb{E}_x\left[\nabla_\theta\ln p_\theta(x)\right]}_{=0(\text{性质3.1})}\\&=\mathbb{E}_x\left[\varphi(\theta)\nabla_\theta\ln p_\theta(x)\right]\\&\overset{\text{交换微分积分号}}{=}\frac{\partial}{\partial\theta}\left[\int T(x)p_\theta(x)\,\mathrm{d}x\right]\\&=\frac{\partial}{\partial\theta}\varphi(\theta)=\varphi'(\theta).\end{aligned}$$

那么由 Cauchy–Schwarz 不等式可以得到：

$$\mathrm{Var}_x[\hat{\theta}]I_X(\theta)=Var_x[\hat{\theta}]\,\mathrm{Var}_x[\nabla_\theta\ln p_\theta(x)]\geq\mathrm{Cov}^2(\varphi(\theta),\nabla_\theta\ln p_\theta(x))=\left(\varphi'(\theta)\right)^2.$$

□

从这个角度讲，C-R 下界表示了无偏估计精度的上限：任何此类估计的精度至多是 Fisher 信息。同时需要指出的是，能够达到 C-R 下界的无偏估计并不多，大多数时候都达不到 C-R 下界，例如，均值为零的正态分布的标准差的无偏估计的方差都大于其 C-R 下界。

2 Application: Adaptive Learning Rate Based on Natural Gradient

在介绍 Natural Gradient 之前，需要补充 Fisher 信息与 KL 散度的简单关系。

对于某分布，有待估计的参数向量 $\boldsymbol{\theta}\in\boldsymbol{\Theta}\subseteq\mathbb{R}^n$。

引理 1. $\forall\boldsymbol{\theta},\boldsymbol{\theta}^*\in\boldsymbol{\Theta}$，其对应的分布为 $p_{\boldsymbol{\theta}}(\boldsymbol{x}),p_{\boldsymbol{\theta}^*}(\boldsymbol{x})$，考察其 *KL* 散度：

$$\mathrm{KL}(p_{\boldsymbol{\theta}^*}(\boldsymbol{x}),p_{\boldsymbol{\theta}}(\boldsymbol{x}))=\int p_{\boldsymbol{\theta}^*}(\boldsymbol{x})\ln\frac{p_{\boldsymbol{\theta}^*}(\boldsymbol{x})}{p_{\boldsymbol{\theta}}(\boldsymbol{x})}\,\mathrm{d}\boldsymbol{x}.$$

那么：

$$\begin{aligned}\nabla_{\boldsymbol{\theta}}\,\mathrm{KL}(p_{\boldsymbol{\theta}^*}(\boldsymbol{x}),p_{\boldsymbol{\theta}}(\boldsymbol{x}))\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}&=\boldsymbol{0}\\\nabla^2_{\boldsymbol{\theta}}\,\mathrm{KL}(p_{\boldsymbol{\theta}^*}(\boldsymbol{x}),p_{\boldsymbol{\theta}}(\boldsymbol{x}))\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}&=I_X(\boldsymbol{\theta}^*).\end{aligned} \tag{18}$$

由 *Taylor* 展开和上式结论可以直接得到，当 $\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2\to0$ 时：

$$\mathrm{KL}(p_{\boldsymbol{\theta}^*}(\boldsymbol{x}),p_{\boldsymbol{\theta}}(\boldsymbol{x}))=\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}^*)^{\mathrm{T}}I_X(\boldsymbol{\theta}^*)(\boldsymbol{\theta}-\boldsymbol{\theta}^*)+o(\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|_2^2). \tag{19}$$

Proof. 留做习题。 □

考虑可微损失函数 $\mathcal{L}(\boldsymbol{\theta}) : \mathbb{R}^n \mapsto \mathbb{R}$，我们考察在以 KL 散度为度量的分布空间中的类似于以欧几里得度量的最速下降法。为此，我们的最小化目标：

$$\boldsymbol{v}* = \underset{\boldsymbol{v}, \mathrm{KL}(p_{\boldsymbol{\theta}^*}(\boldsymbol{x}), p_{\boldsymbol{\theta}+\boldsymbol{v}}(\boldsymbol{x}))=c}{\operatorname{argmin}} \mathcal{L}(\boldsymbol{\theta}+\boldsymbol{v}). \tag{20}$$

其中 $c = \frac{1}{2}\|\nabla_\theta \mathcal{L}(\boldsymbol{\theta})\|^2$。在以 KL 散度为度量的空间中研究优化问题的好处是，算法对模型的参数更加鲁棒，也即算法不关心模型是具体的参数数值，而只关心参数的分布。将 KL 散度固定为某个常数以内的目的是确保优化目标以恒定速度上界沿着空间移动，而不管曲率如何 [3]。

由于 $\mathcal{L}(\boldsymbol{\theta})$ 可微，使用 Lagrange 乘子法求解理论模型 $\boldsymbol{v}*$。拉式量：

$$\begin{aligned} W &= \mathcal{L}(\boldsymbol{\theta}+\boldsymbol{v}) + \lambda(\mathrm{KL}(p_{\boldsymbol{\theta}^*}(\boldsymbol{x}), p_{\boldsymbol{\theta}+\boldsymbol{v}}(\boldsymbol{x})) - c) \\ &\overset{\text{引理1}}{\approx} \mathcal{L}(\boldsymbol{\theta}) + \nabla_\theta \mathcal{L}(\boldsymbol{\theta})^{\mathrm{T}}\boldsymbol{v} + \frac{1}{2}\lambda \boldsymbol{v}^{\mathrm{T}} I_X(\boldsymbol{\theta})\boldsymbol{v} - \lambda c. \end{aligned}$$

求导得：

$$\begin{aligned} \mathbf{0} &= \frac{\partial}{\partial \boldsymbol{v}}\left[\mathcal{L}(\boldsymbol{\theta}) + \nabla_\theta \mathcal{L}(\boldsymbol{\theta})^{\mathrm{T}}\boldsymbol{v} + \frac{1}{2}\lambda \boldsymbol{v}^{\mathrm{T}} I_X(\boldsymbol{\theta})\boldsymbol{v} - \lambda c\right] \\ \mathbf{0} &= \nabla_\theta \mathcal{L}(\boldsymbol{\theta}) + \lambda I_X(\boldsymbol{\theta})\boldsymbol{v} \\ \boldsymbol{v} &= -\frac{\nabla_\theta \mathcal{L}(\boldsymbol{\theta})}{\lambda I_X(\boldsymbol{\theta})}. \end{aligned}$$

不考虑常数项，我们得到了 (更自然的) 优化方向。将 $I_X^{-1}(\boldsymbol{\theta})$ 吸收进梯度项，我们得到了如下定义。

定义 3 (Natural Gradient)**.** *对于可微损失函数 $\mathcal{L}(\boldsymbol{\theta}) : \mathbb{R}^n \mapsto \mathbb{R}$，可以定义自然梯度：*

$$\tilde{\nabla}_\theta \mathcal{L}(\boldsymbol{\theta}) = I_X^{-1}(\boldsymbol{\theta})\nabla_\theta \mathcal{L}(\boldsymbol{\theta}). \tag{21}$$

由上述定义可以直接给出一个参数优化策略。

Algorithm 1: Natural Gradient based Optimization Policy

while Convergence **do**
- Do forward pass on our model and compute loss$\mathcal{L}(\boldsymbol{\theta})$;
- Compute the gradient$\nabla_\theta \mathcal{L}(\boldsymbol{\theta})$;
- Compute the Fisher Information Matrix$I_X(\boldsymbol{\theta})$, or its empirical version (based on training data);
- Compute the natural gradient$\tilde{\nabla}_\theta \mathcal{L}(\boldsymbol{\theta}) = I_X^{-1}(\boldsymbol{\theta})\nabla_\theta \mathcal{L}(\boldsymbol{\theta})$;
- Update the parameter:$\boldsymbol{\theta} = \boldsymbol{\theta} - \alpha\tilde{\nabla}_\theta \mathcal{L}(\boldsymbol{\theta})$, where $a\alpha$ is the learning rate;

诚然，该方法基本可以实现学习率的自适应调整。但是其代价是巨大的：本质上该算法是一个二阶优化算法，其所需的计算复杂度在参数量巨大的情况下是不可接受的。

解决这个问题的一种方法是使用近似 Fisher 信息阵。像 Adam[1] 这样的方法计算梯度的一阶和二阶矩的滑动平均值。其一阶矩视为动量，而二阶矩近似于 Fisher 信息矩阵，但将其约束为对角矩阵。在实践中，Adam 算法效果一般较为理想，目前是优化深度神经网络的事实标准。但是 Adam 算法不是自然梯度 [7]。

除此之外，自然梯度在强化学习中的 Markov 过程决策中的 Policy Gradient 中 (TRPO) 占据重要的作用 [5]，决策的状态空间一般远远小于 DNN 的参数数量。

3 Application: Overcome Catastrophic Forgetting in Neural Networks

在持续学习 (continual learning) 领域，考虑两个学习任务 A 和 B。一个经常发生的问题是，任务 A 的数据量足够，而任务 B 的数据量不足。所以一般的我们可能先学习任务 A 再学习

任务 B。在学习第二个任务 B 时，神经网络倾向于忘记之前学习的前一个任务的知识 (A)。该学习过程的可以被写为 (对应图4中的"no penalty" 项):

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \mathcal{L}_B(\boldsymbol{\theta}).$$

为了避免忘记任务 A 中学到的知识，一个简单的技巧是我们可以最小化 $\boldsymbol{\theta}$ 和 $\boldsymbol{\theta}_A^*$ 两者之间的距离。因此学习过程变为 (对应图4中的"L_2" 项):

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \mathcal{L}_B(\boldsymbol{\theta}) + \frac{1}{2}\alpha\|\boldsymbol{\theta} - \boldsymbol{\theta}_A^*\|_2^2.$$

其中 α 为标量表明旧任务相比于新任务的重要性。

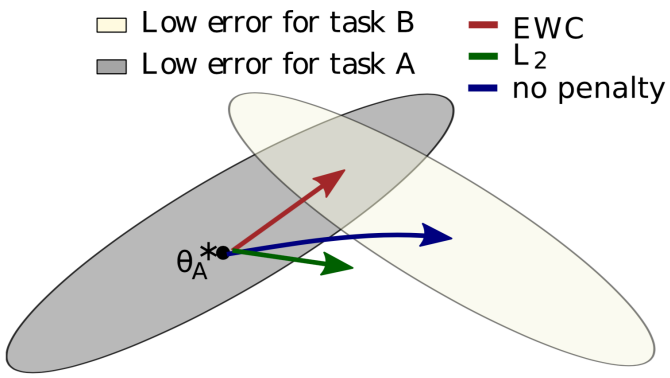

Figure 4: 任务 B 在任务 A 之后的学习过程示意图. 图摘自 [2].

事实证明，L_2 约束是如此强大，以至于它可能会阻碍任务 B 的学习过程。在这里，我们还有一个观察到：在神经网络中，我们经常过度参数化 (over-parametrized) 模型。可能有一些参数不太有用，而另一些参数更有价值。在 L_2 约束情况下，每个参数都被平等对待。在这里，我们想使用 Fisher 信息矩阵中的对角线分量来识别哪些参数对任务 A 更重要，并对它们施加更高的权重，该方案称为 Elastic Weight Consolidation(EWC)[2]。该学习过程的可以被写为 (对应图4中的"EWC" 项):

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \mathcal{L}_B(\boldsymbol{\theta}) + \frac{1}{2}\alpha \sum_{i=1}^{n} [I_X(\boldsymbol{\theta}_A^*)]_{ii}(\theta_i - \theta_{A,i}^*)^2.$$

其中 $[I_X(\boldsymbol{\theta}_A^*)]_{ii}$ 为 Fisher 信息矩阵的对角元素，其大小代表了参数 θ_i 对于任务 A 的重要性。

References

[1] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

[2] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. Proceedings of the national academy of sciences, 114(13):3521–3526, 2017.

[3] Razvan Pascanu and Yoshua Bengio. Revisiting natural gradient for deep networks. arXiv preprint arXiv:1301.3584, 2013.

[4] Yury Polyanskiy and Yihong Wu. Lecture notes on information theory. Lecture Notes for ECE563 (UIUC) and, 6(2012-2016):7, 2014.

[5] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In International conference on machine learning, pages 1889–1897. PMLR, 2015.

[6] MK Steven. Fundamentals of statistical processing: Estimation theory. Prectice Hall, 1993.

[7] Karl Stratos. Fisher information and policy gradient methods. 2020.

[8] Harry L Van Trees. Detection, estimation, and modulation theory, part I: detection, estimation, and linear modulation theory. John Wiley & Sons, 2004.

[9] Ram Zamir. A proof of the fisher information inequality via a data processing argument. IEEE Transactions on Information Theory, 44(3):1246–1250, 1998.

[10] 茆诗松, 程依明, 濮晓龙, 查看清, and 单书目. 概率论与数理统计教程: 第二版. 北京: 高等教育出版社, 2011.