Chapter 8

Fisher Information

Having explored the definitions associated with exponential families and their robustness properties, we now turn to a study of somewhat more general parameterized distributions, developing connections between divergence measures and other geometric ideas such as the Fisher information. After this, we illustrate a few consequences of Fisher information for optimal estimators, which gives a small taste of the deep connections between information geometry, Fisher information, exponential family models. In the coming chapters, we show how Fisher information measures come to play a central role in sequential (universal) prediction problems.

8.1 Fisher information: definitions and examples

We begin by defining the Fisher information. Let $\{P_\theta\}_{\theta \in \Theta}$ denote a parametric family of distributions on a space $\mathcal{X}$, each where $\theta \in \Theta \subset \mathbb{R}^d$ indexes the distribution. Throughout this lecture and the next, we assume (with no real loss of generality) that each P_θ has a density given by p_θ. Then the *Fisher information* associated with the model is the matrix given by

$$I_\theta := \mathbb{E}_\theta \left[\nabla_\theta \log p_\theta(X) \nabla \log p_\theta(X)^\top \right] = \mathbb{E}_\theta[\dot{\ell}_\theta \dot{\ell}_\theta^\top], \tag{8.1.1}$$

where the score function $\dot{\ell}_\theta = \nabla_\theta \log p_\theta(x)$ is the gradient of the log likelihood at θ (implicitly depending on X) and the expectation $\mathbb{E}_\theta$ denotes expectation taken with respect to P_θ. Intuitively, the Fisher information captures the variability of the gradient $\nabla \log p_\theta$; in a family of distributions for which the score function $\dot{\ell}_\theta$ has high variability, we intuitively expect estimation of the parameter θ to be easier—different θ change the behavior of $\dot{\ell}_\theta$—though the log-likelihood functional $\theta \mapsto \mathbb{E}_{\theta_0}[\log p_\theta(X)]$ varies more in θ.

Under suitable smoothness conditions on the densities p_θ (roughly, that derivatives pass through expectations; see Remark 8.1 at the end of this chapter), there are a variety of alternate definitions of Fisher information. These smoothness conditions hold for exponential families, so at least in the exponential family case, everything in this chapter is rigorous. (We note in passing that there are more general definitions of Fisher information for more general families under quadratic mean differentiability; see, for example, van der Vaart [4].) First, we note that the score function has

mean zero under P_θ: we have

$$\begin{aligned}\mathbb{E}_\theta[\dot{\ell}_\theta] &= \int p_\theta(x) \nabla_\theta \log p_\theta(x) dx = \int \frac{\nabla p_\theta(x)}{p_\theta(x)} p_\theta(x) dx \\ &= \int \nabla p_\theta(x) dx \stackrel{(\star)}{=} \nabla \int p_\theta(x) dx = \nabla 1 = 0,\end{aligned}$$

where in equality $(\star)$ we have assumed that integration and derivation may be exchanged. Under similar conditions, we thus attain an alternate definition of Fisher information as the negative expected hessian of $\log p_\theta(X)$. Indeed,

$$\nabla^2 \log p_\theta(x) = \frac{\nabla^2 p_\theta(x)}{p_\theta(x)} - \frac{\nabla p_\theta(x) \nabla p_\theta(x)^\top}{p_\theta(x)^2} = \frac{\nabla^2 p_\theta(x)}{p_\theta(x)} - \dot{\ell}_\theta \dot{\ell}_\theta^\top,$$

so we have that the Fisher information is equal to

$$\begin{aligned}I_\theta &= \mathbb{E}_\theta[\dot{\ell}_\theta \dot{\ell}_\theta^\top] = -\int p_\theta(x) \nabla^2 \log p_\theta(x) dx + \int \nabla^2 p_\theta(x) dx \\ &= -\mathbb{E}[\nabla^2 \log p_\theta(x)] + \nabla^2 \underbrace{\int p_\theta(x) dx}_{=1} = -\mathbb{E}[\nabla^2 \log p_\theta(x)]. \qquad (8.1.2)\end{aligned}$$

Summarizing, we have that

$$I_\theta = \mathbb{E}_\theta[\dot{\ell}_\theta \dot{\ell}_\theta] = -\mathbb{E}_\theta[\nabla^2 \log p_\theta(X)].$$

This representation also makes clear the additional fact that, if we have n i.i.d. observations from the model P_θ, then the information content similarly grows linearly, as $\log p_\theta(X_1^n) = \sum_{i=1}^n \log p_\theta(X_i)$.

We now give two examples of Fisher information, the first somewhat abstract and the second more concrete.

Example 8.1 (Canonical exponential family)**:** In a canonical exponential family model, we have $\log p_\theta(x) = \langle \theta, \phi(x) \rangle - A(\theta)$, where ϕ is the sufficient statistic and A is the log-partition function. Because $\dot{\ell}_\theta = \phi(x) - \nabla A(\theta)$ and $\nabla^2 \log p_\theta(x) = -\nabla^2 A(\theta)$ is a constant, we obtain

$$I_\theta = \nabla^2 A(\theta).$$

♣

Example 8.2 (Two parameterizations of a Bernoulli)**:** In the canonical parameterization of a Bernoulli as an exponential family model (Example 6.1), we had $p_\theta(x) = \exp(\theta x - \log(1 + e^\theta))$ for $x \in \{0, 1\}$, so by the preceding example the associated Fisher information is $\frac{e^\theta}{1+e^\theta} \frac{1}{1+e^\theta}$. If we make the change of variables $p = P_\theta(X = 1) = e^\theta/(1 + e^\theta)$, or $\theta = \log \frac{p}{1-p}$, we have $I_\theta = p(1 - p)$. On the other hand, if $P(X = x) = p^x (1 - p)^{1-x}$ for $p \in [0, 1]$, the standard formulation of the Bernoulli, then $\nabla \log P(X = x) = \frac{x}{p} - \frac{1-x}{1-p}$, so that

$$I_p = \mathbb{E}_p \left[\left(\frac{X}{p} - \frac{1 - X}{1 - p} \right)^2 \right] = \frac{1}{p} + \frac{1}{1 - p} = \frac{1}{p(1 - p)}.$$

That is, the parameterization can change the Fisher information. ♣

8.2 Estimation and Fisher information: elementary considerations

The Fisher information has intimate connections to estimation, both in terms of classical estimation and the information games that we discuss subsequently. As a motivating calculation, we consider estimation of the mean of a Bernoulli(p) random variable, where $p \in [0, 1]$, from a sample $X_1^n \stackrel{\text{i.i.d.}}{\sim} \text{Bernoulli}(p)$. The sample mean $\widehat{p}$ satisfies

$$\mathbb{E}[(\widehat{p} - p)^2] = \frac{1}{n} \text{Var}(X) = \frac{p(1-p)}{n} = \frac{1}{I_p} \cdot \frac{1}{n},$$

where I_p is the Fisher information for the single observation Bernoulli(p) family as in Example 8.2. In fact, this inverse dependence on Fisher information is unavoidable, as made clear by the Cramér Rao Bound, which provides lower bounds on the mean squared error of all unbiased estimators.

Proposition 8.3 (Cramér Rao Bound)**.** *Let $\phi : \mathbb{R}^d \to \mathbb{R}$ be an arbitrary differentiable function and assume that the random function (estimator) T is unbiased for $\phi(\theta)$ under P_θ. Then*

$$\text{Var}(T) \geq \nabla \phi(\theta)^\top I_\theta^{-1} \nabla \phi(\theta).$$

As an immediate corollary to Proposition 8.3, we may take $\phi(\theta) = \langle \lambda, \theta \rangle$ for $\lambda \in \mathbb{R}^d$. Then varying λ over all of $\mathbb{R}^d$, and we obtain that for any unbiased estimator T for the parameter $\theta \in \mathbb{R}^d$, we have $\text{Var}(\langle \lambda, T \rangle) \geq \lambda^\top I_\theta^{-1} \lambda$. That is, we have

Corollary 8.4. *Let T be unbiased for the parameter θ under the distribution P_θ. Then the covariance of T has lower bound*

$$\text{Cov}(T) \succeq I_\theta^{-1}.$$

In fact, the Cramér-Rao bound and Corollary 8.4 hold, in an asymptotic sense, for substantially more general settings (without the unbiasedness requirement). For example, see the books of van der Vaart [4] or Le Cam and Yang [3, Chapters 6 & 7], which show that under appropriate conditions (known variously as quadratic mean differentiability and local asymptotic normality) that no estimator can have smaller mean squared error than Fisher information in any uniform sense.

We now prove the proposition, where, as usual, we assume that it is possible to exchange differentiation and integration.

Proof Throughout this proof, all expectations and variances are computed with respect to P_θ. The idea of the proof is to choose $\lambda \in \mathbb{R}^d$ to minimize the variance

$$\text{Var}(T - \langle \lambda, \dot{\ell}_\theta \rangle) \geq 0,$$

then use this λ to provide a lower bound on $\text{Var}(T)$.

To that end, let $\dot{\ell}_{\theta,j} = \frac{\partial}{\partial \theta_j} \log p_\theta(X)$ denote the jth component of the score vector. Because $\mathbb{E}_\theta[\dot{\ell}_\theta] = 0$, we have the covariance equality

$$\begin{aligned}
\text{Cov}(T - \phi(\theta), \dot{\ell}_{\theta,j}) = \mathbb{E}[(T - \phi(\theta))\dot{\ell}_{\theta,j}] = \mathbb{E}[T\dot{\ell}_{\theta,j}] = \int T(x) \frac{\frac{\partial}{\partial \theta_j} p_\theta(x)}{p_\theta(x)} p_\theta(x) dx \\
= \frac{\partial}{\partial \theta_j} \int T(x) p_\theta(x) dx = \frac{\partial}{\partial \theta_j} \phi(\theta),
\end{aligned}$$

where in the final step we used that T is unbiased for $\phi(\theta)$. Using the preceding equality,

$$\mathrm{Var}(T - \langle \lambda, \dot{\ell}_\theta \rangle) = \mathrm{Var}(T) + \lambda^\top I_\theta \lambda - 2\mathbb{E}[(T - \phi(\theta))\langle \lambda, \dot{\ell}_\theta \rangle] = \mathrm{Var}(T) + \lambda^\top I_\theta \lambda - 2\langle \lambda, \nabla \phi(\theta) \rangle.$$

Taking $\lambda = I_\theta^{-1} \nabla \phi(\theta)$ gives $0 \le \mathrm{Var}(T - \langle \lambda, \dot{\ell}_\theta \rangle) = \mathrm{Var}(T) - \nabla \phi(\theta)^\top I_\theta^{-1} \nabla \phi(\theta)$, and rearranging gives the result. □

8.3 Connections between Fisher information and divergence measures

By making connections between Fisher information and certain divergence measures, such as KL-divergence and mutual (Shannon) information, we gain additional insights into the structure of distributions, as well as optimal estimation and encoding procedures. As a consequence of the asymptotic expansions we make here, we see that estimation of 1-dimensional parameters is governed (essentially) by moduli of continuity of the loss function with respect to the metric induced by Fisher information; in short, Fisher information is an unavoidable quantity in estimation. We motivate our subsequent development with the following example.

Example 8.5 (Divergences in exponential families)**:** Consider the exponential family density $p_\theta(x) = h(x) \exp(\langle \theta, \phi(x) \rangle - A(\theta))$. Then a straightforward calculation implies that for any θ_1 and θ_2, the KL-divergence between distributions P_{θ_1} and P_{θ_2} is

$$D_{\rm kl}\left(P_{\theta_1} \| P_{\theta_2}\right) = A(\theta_2) - A(\theta_1) - \langle \nabla A(\theta_1), \theta_2 - \theta_1 \rangle .$$

That is, the divergence is simply the difference between $A(\theta_2)$ and its first order expansion around θ_1. This suggests that we may approximate the KL-divergence via the quadratic remainder in the first order expansion. Indeed, as A is infinitely differentiable (it is an exponential family model), the Taylor expansion becomes

$$\begin{aligned} D_{\rm kl}\left(P_{\theta_1} \| P_{\theta_2}\right) &= \frac{1}{2} \left\langle \theta_1 - \theta_2, \nabla^2 A(\theta_1)(\theta_1 - \theta_2) \right\rangle + O(\|\theta_1 - \theta_2\|^3) \\ &= \frac{1}{2} \left\langle \theta_1 - \theta_2, I_{\theta_1}(\theta_1 - \theta_2) \right\rangle + O(\|\theta_1 - \theta_2\|^3). \end{aligned}$$

♣

In particular, KL-divergence is roughly quadratic for exponential family models, where the quadratic form is given by the Fisher information matrix. We also remark in passing that for a convex function f, the Bregman divergence (associated with f) between points x and y is given by $B_f(x, y) = f(x) - f(y) - \langle \nabla f(y), x - y \rangle$; such divergences are common in convex analysis, optimization, and differential geometry. Making such connections deeper and more rigorous is the goal of the field of information geometry (see the book of Amari and Nagaoka [1] for more).

We can generalize this example substantially under appropriate smoothness conditions. Indeed, we have

Proposition 8.6. *For appropriately smooth families of distributions $\{P_\theta\}_{\theta \in \Theta}$,*

$$D_{\rm kl}\left(P_{\theta_1} \| P_{\theta_2}\right) = \frac{1}{2} \left\langle \theta_1 - \theta_2, I_{\theta_1}(\theta_1 - \theta_2) \right\rangle + o(\|\theta_1 - \theta_2\|^2). \tag{8.3.1}$$

We only sketch the proof, as making it fully rigorous requires measure-theoretic arguments and Lebesgue's dominated convergence theorem.

Sketch of Proof By a Taylor expansion of the log density $\log p_{\theta_2}(x)$ about θ_1, we have

$$\begin{aligned}\log p_{\theta_2}(x) = \log p_{\theta_1}(x) + \langle \nabla \log p_{\theta_1}(x), \theta_1 - \theta_2 \rangle \\ + \frac{1}{2}(\theta_1 - \theta_2)^\top \nabla^2 \log p_{\theta_1}(x)(\theta_1 - \theta_2) + R(\theta_1, \theta_2, x),\end{aligned}$$

where $R(\theta_1, \theta_2, x) = O_x(\|\theta_1 - \theta_2\|^3)$ is the remainder term, where O_x denotes a hidden dependence on x. Taking expectations and assuming that we can interchange differentiation and expectation appropriately, we have

$$\begin{aligned}\mathbb{E}_{\theta_1}[\log p_{\theta_2}(X)] &= \mathbb{E}_{\theta_1}[\log p_{\theta_1}(X)] + \left\langle \mathbb{E}_{\theta_1}[\dot{\ell}_{\theta_1}], \theta_1 - \theta_2 \right\rangle \\ &\quad + \frac{1}{2}(\theta_1 - \theta_2)^\top \mathbb{E}_{\theta_1}[\nabla^2 \log p_{\theta_1}(X)](\theta_1 - \theta_2) + \mathbb{E}_{\theta_1}[R(\theta_1, \theta_2, X)] \\ &= \mathbb{E}_{\theta_1}[\log p_{\theta_1}(X)] - \frac{1}{2}(\theta_1 - \theta_2)^\top I_{\theta_1}(\theta_1 - \theta_2) + o(\|\theta_1 - \theta_2\|^2),\end{aligned}$$

where we have assumed that the $O(\|\theta_1 - \theta_2\|^3)$ remainder is uniform enough in X that $\mathbb{E}[R] = o(\|\theta_1 - \theta_2\|^2)$ and used that the score function $\dot{\ell}_\theta$ is mean zero under P_θ. □

We may use Proposition 8.6 to give a somewhat more general version of the Cramér-Rao bound (Proposition 8.3) that applies to more general (sufficiently smooth) estimation problems. Indeed, we will show that Le Cam's method (recall Chapter 2.3) is (roughly) performing a type of discrete second-order approximation to the KL-divergence, then using this to provide lower bounds. More concretely, suppose we are attempting to estimate a parameter θ parameterizing the family $\mathcal{P} = \{P_\theta\}_{\theta \in \Theta}$, and assume that $\Theta \subset \mathbb{R}^d$ and $\theta_0 \in \operatorname{int} \Theta$. Consider the minimax rate of estimation of θ_0 in a neighborhood around θ_0; that is, consider

$$\inf_{\widehat{\theta}} \sup_{\theta = \theta_0 + v \in \Theta} \mathbb{E}_\theta[\|\widehat{\theta}(X_1^n) - \theta\|^2],$$

where the observations X_i are drawn i.i.d. P_θ. Fixing $v \in \mathbb{R}^d$ and setting $\theta = \theta_0 + \delta v$ for some $\delta > 0$, Le Cam's method (2.3.3) then implies that

$$\inf_{\widehat{\theta}} \max_{\theta \in \{\theta_0, \theta + \delta v\}} \mathbb{E}_\theta[\|\widehat{\theta}(X_1^n) - \theta\|^2] \geq \frac{\delta^2 \|v\|^2}{8} \left[1 - \left\|P_{\theta_0}^n - P_{\theta_0 + \delta v}^n\right\|_{\mathrm{TV}}\right].$$

Using Pinsker's inequality that $2\|P - Q\|_{\mathrm{TV}}^2 \leq D_{\mathrm{kl}}(P\|Q)$ and the asymptotic quadratic approximation (8.3.1), we have

$$\left\|P_{\theta_0}^n - P_{\theta_0 + \delta v}^n\right\|_{\mathrm{TV}} \leq \sqrt{\frac{n}{2} D_{\mathrm{kl}}(P_{\theta_0} \| P_{\theta_0 + \delta v})} = \frac{\sqrt{n}}{2}\left(\delta^2 v^\top I_{\theta_0} v + o(\delta^2 \|v\|^2)\right)^{\frac{1}{2}}.$$

By taking $\delta^2 = (nv^\top I_{\theta_0} v)^{-1}$, for large enough v and n we know that $\theta_0 + \delta v \in \operatorname{int} \Theta$ (so that the distribution $P_{\theta_0 + \delta v}$ exists), and for large n, the remainder term $o(\delta^2 \|v\|^2)$ becomes negligible. Thus we obtain

$$\inf_{\widehat{\theta}} \max_{\theta \in \{\theta_0, \theta + \delta v\}} \mathbb{E}_\theta[\|\widehat{\theta}(X_1^n) - \theta\|^2] \gtrsim \frac{\delta^2 \|v\|^2}{16} = \frac{1}{16} \frac{\|v\|^2}{nv^\top I_{\theta_0} v}. \tag{8.3.2}$$

In particular, in one-dimension, inequality (8.3.2) implies a result generalizing the Cramér-Rao bound. We have the following asymptotic local minimax result:

Corollary 8.7. *Let $\mathcal{P} = \{P_\theta\}_{\theta \in \Theta}$, where $\Theta \subset \mathbb{R}$, be a family of distributions satisfying the quadratic approximation condition of Proposition 8.6. Then there exists a constant $c > 0$ such that*

$$\lim_{v \to \infty} \lim_{n \to \infty} \inf_{\widehat{\theta}_n} \sup_{\theta : |\theta - \theta_0| \le v/\sqrt{n}} \mathbb{E}_\theta \left[(\widehat{\theta}_n(X_1^n) - \theta)^2 \right] \ge c \frac{1}{n} I_{\theta_0}^{-1}.$$

Written differently (and with minor extension), Corollary 8.7 gives a lower bound based on a local modulus of continuity of the loss function with respect to the metric induced by the Fisher information. Indeed, suppose we wish to estimate a parameter θ in the neighborhood of θ_0 (where the neighborhood size decreases as $1/\sqrt{n}$) according to some loss function $\ell : \Theta \times \Theta \to \mathbb{R}$. Then if we define the modulus of continuity of ℓ with respect to the Fisher information metric as

$$\omega_\ell(\delta, \theta_0) := \sup_{v : \|v\| \le 1} \frac{\ell(\theta_0, \theta_0 + \delta v)}{\delta^2 v^\top I_{\theta_0} v},$$

the combination of Corollary 8.7 and inequality (8.3.2) shows that the local minimax rate of estimating $\mathbb{E}_\theta[\ell(\widehat{\theta}_n, \theta)]$ for θ near θ_0 must be at least $\omega_\ell(n^{-1/2}, \theta_0)$. For more on connections between moduli of continuity and estimation, see, for example, Donoho and Liu [2].

Remark 8.1: In order to make all of our exchanges of differentiation and expectation rigorous, we must have some conditions on the densities we consider. One simple condition sufficient to make this work is via Lebesgue's dominated convergence theorem. Let $f : \mathcal{X} \times \Theta \to \mathbb{R}$ be a differentiable function. For a fixed base measure μ assume there exists a function g such that $g(x) \ge \|\nabla_\theta f(x, \theta)\|$ for all θ, where

$$\int_{\mathcal{X}} g(x) d\mu(x) < \infty.$$

Then in this case, we have $\nabla_\theta \int f(x, \theta) d\mu(x) = \int \nabla_\theta f(x, \theta) d\mu(x)$ by the mean-value theorem and definition of a derivative. (Note that for all θ_0 we have $\sup_{v : \|v\|_2 \le \delta} \|\nabla_\theta f(x, \theta)\|_2 \big|_{\theta = \theta_0 + v} \le g(x)$.) More generally, this type of argument can handle absolutely continuous functions, which are differentiable almost everywhere. $\diamondsuit$

Bibliography

[1] S. Amari and H. Nagaoka. *Methods of Information Geometry*, volume 191 of *Translations of Mathematical Monographs*. American Mathematical Society, 2000.

[2] D. L. Donoho and R. C. Liu. Geometrizing rates of convergence I. Technical Report 137, University of California, Berkeley, Department of Statistics, 1987.

[3] L. Le Cam and G. L. Yang. *Asymptotics in Statistics: Some Basic Concepts*. Springer, 2000.

[4] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998. ISBN 0-521-49603-9.