Information Theory Notes

ChatGPT, proofread by OM

December 4th 2025

1 Fisher Information: Definition, Intuition, Properties, and Consequences

1.1 Definition

Let $\{p(x;\theta)\}_{\theta\in\Theta}$ be a parametric family of probability density functions (or mass functions) satisfying standard regularity conditions: (i) the support does not depend on θ, (ii) the functions are differentiable in θ, and (iii) differentiation and integration can be interchanged. The *score* is defined as

$$s_\theta(X) := \frac{\partial}{\partial\theta}\log p(X;\theta).$$

The *Fisher information* for a single observation is defined as

$$I(\theta) := \mathbb{E}_\theta\big[s_\theta(X)^2\big] = \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log p(X;\theta)\right)^2\right].$$

We now establish all fundamental facts relating the Fisher information to curvature of the log-likelihood, distinguishability of probability distributions, and estimator precision.

1.2 The Score Has Mean Zero

Lemma 1 *Under the regularity assumptions above,* $\mathbb{E}_\theta[s_\theta(X)] = 0$.

We compute

$$\mathbb{E}_\theta[s_\theta(X)] = \int \frac{p'_\theta(x)}{p(x;\theta)}p(x;\theta)\,dx = \int p'_\theta(x)\,dx = \frac{\partial}{\partial\theta}\int p(x;\theta)\,dx = \frac{\partial}{\partial\theta}1 = 0.$$

This identity is fundamental: the score behaves like a centered *random directional derivative of the log-likelihood*.

1.3 Equality Between the Two Forms of Fisher Information

Theorem 1 *Under the same regularity assumptions,*

$$I(\theta) = \mathbb{E}_\theta\left[s_\theta(X)^2\right] = -\,\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\log p(X;\theta)\right].$$

Differentiate the identity $\mathbb{E}_\theta[s_\theta(X)] = 0$ with respect to θ:

$$0 = \frac{\partial}{\partial\theta}\int s_\theta(x)\,p(x;\theta)\,dx = \int \frac{\partial}{\partial\theta}\left(s_\theta(x)p(x;\theta)\right)dx.$$

Using the product rule and the fact that $p'_\theta = p\,s_\theta$,

$$0 = \int \partial_\theta s_\theta(x)\,p(x;\theta)\,dx \;+\; \int s_\theta(x)\,p(x;\theta)\,s_\theta(x)\,dx.$$

Thus

$$\mathbb{E}_\theta[s_\theta(X)^2] = -\,\mathbb{E}_\theta[\partial_\theta s_\theta(X)] = -\,\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\log p(X;\theta)\right].$$

This identity reveals that Fisher information is the *expected curvature* of the log-likelihood around the true parameter.

1.4 Fisher Information Measures Sensitivity of the Distribution

The score $s_\theta(X)$ measures how sensitively the log-density reacts to infinitesimal perturbations of the parameter θ. Large magnitudes of the score correspond to samples X for which the likelihood surface changes sharply with θ. Because $I(\theta) = \mathrm{Var}[s_\theta(X)]$, the Fisher information captures how variable (and therefore how informative) this sensitivity is across samples.

More formally, when the Fisher information is large, small changes in θ cause the distribution $p(\cdot;\theta)$ to change rapidly, which makes it easier to infer θ from observations. When the Fisher information is small, the distribution changes only weakly with θ, making the parameter intrinsically difficult to estimate.

1.5 Fisher Information as Curvature of the Log-Likelihood

Let $X_1, \ldots, X_n$ be i.i.d. with log-likelihood $\ell(\theta) = \sum_{i=1}^n \log p(X_i;\theta)$. Using the identity above,

$$\mathbb{E}_\theta[-\ell''(\theta)] = nI(\theta).$$

Thus:

- A large $I(\theta)$ means the likelihood is *sharply peaked* around the true parameter, yielding a precise estimator.

- A small $I(\theta)$ means the likelihood is relatively flat, making all nearby parameter values almost equally plausible.

This connects Fisher information to the geometry of likelihood-based inference.

1.6 Fisher Information as Local Distinguishability

The next result formally connects Fisher information with the *statistical distinguishability* of nearby probability distributions.

Theorem 2 *For sufficiently smooth families,*

$$\mathrm{KL}(p_\theta \,\|\, p_{\theta+\Delta\theta}) = \frac{1}{2} I(\theta)(\Delta\theta)^2 + o\big((\Delta\theta)^2\big).$$

Expand $\log p(x;\theta+\Delta\theta)$ to second order:

$$\log p(x;\theta+\Delta\theta) = \log p(x;\theta) + s_\theta(x)\,\Delta\theta + \frac{1}{2}\frac{\partial^2}{\partial\theta^2}\log p(x;\theta)(\Delta\theta)^2 + o((\Delta\theta)^2).$$

Plugging this into the KL divergence definition and using $\mathbb{E}[s_\theta(X)] = 0$ and

$$I(\theta) = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log p(X;\theta)\right],$$

one obtains

$$\mathrm{KL}(p_\theta \,\|\, p_{\theta+\Delta\theta}) = \tfrac{1}{2} I(\theta)(\Delta\theta)^2 + o((\Delta\theta)^2).$$

This shows that Fisher information equals the *second-order growth rate* of KL divergence between nearby distributions. A large $I(\theta)$ means the distributions diverge rapidly as θ changes—they are easy to distinguish statistically. A small $I(\theta)$ means they remain close—they are hard to tell apart.

1.7 Fisher Information and the Cramér–Rao Lower Bound

The Fisher information plays a crucial role in establishing a universal limit on estimation accuracy.

Theorem 3 (Cramér–Rao Lower Bound) *For any unbiased estimator $\hat{\theta}(X_1^n)$ of θ,*

$$\mathrm{Var}(\hat{\theta}) \;\geq\; \frac{1}{nI(\theta)}.$$

Let $\ell(\theta)$ be the log-likelihood. Using the identity $\mathbb{E}[s_\theta(X)] = 0$ and differentiating with respect to θ, one shows that

$$\mathbb{E}\Big[(\hat{\theta}-\theta)s_\theta(X)\Big] = 1.$$

Applying Cauchy–Schwarz,

$$1^2 \le \mathbb{E}\Big[(\hat{\theta}-\theta)^2\Big] \cdot \mathbb{E}\big[s_\theta(X)^2\big] = \mathrm{Var}(\hat{\theta}) \cdot I(\theta),$$

which rearranges to the stated bound.

Thus a large $I(\theta)$ implies small variance for any unbiased estimator, while small Fisher information fundamentally limits accuracy.

1.8 Example: Fisher Information of a Normal Distribution with Known Variance

Consider $X \sim \mathcal{N}(\mu, \sigma^2)$ with σ^2 known. Then

$$\log p(x;\mu) = -\frac{(x-\mu)^2}{2\sigma^2} + \text{const}, \qquad \frac{\partial}{\partial \mu} \log p(x;\mu) = \frac{x-\mu}{\sigma^2}.$$

Hence

$$I(\mu) = \mathbb{E}\left[\frac{(X-\mu)^2}{\sigma^4}\right] = \frac{1}{\sigma^2}.$$

This perfectly matches the intuition that when the noise variance σ^2 is small, the observations lie tightly around μ and are highly informative; when σ^2 is large, the data are very noisy and provide little information about μ.

1.9 Fisher information for the Bernoulli Model

Let $X \sim \text{Bernoulli}(p)$ with parameter $p \in (0,1)$ and pmf

$$P(X=x) = p^x(1-p)^{1-x}, \qquad x \in \{0,1\}.$$

The support $\{0,1\}$ does not depend on p, and all required regularity conditions hold, so both common representations of Fisher information are valid.

(1) Log-likelihood and score. The log-likelihood for a single observation is

$$\ell(p;x) = \log p(x;p) = x \log p + (1-x)\log(1-p).$$

Differentiate w.r.t. p to obtain the score:

$$s_p(x) = \frac{\partial}{\partial p}\ell(p;x) = \frac{x}{p} - \frac{1-x}{1-p} = \frac{x-p}{p(1-p)}.$$

(2) Expectation of the score (zero mean). Using $\mathbb{E}[X] = p$,

$$\mathbb{E}_p[s_p(X)] = \mathbb{E}\left[\frac{X-p}{p(1-p)}\right] = \frac{\mathbb{E}[X]-p}{p(1-p)} = 0.$$

(3) Fisher information via variance of the score. By definition,

$$I(p) = \mathbb{E}_p[s_p(X)^2] = \mathbb{E}\left[\frac{(X-p)^2}{p^2(1-p)^2}\right] = \frac{\mathrm{Var}(X)}{p^2(1-p)^2}.$$

Since $\mathrm{Var}(X) = p(1-p)$ for a Bernoulli,

$$I(p) = \frac{p(1-p)}{p^2(1-p)^2} = \frac{1}{p(1-p)}.$$

(4) Fisher information via the negative expected second derivative. Differentiate the score to obtain the second derivative of the log-likelihood:

$$\frac{\partial^2}{\partial p^2}\ell(p;x) = \frac{\partial}{\partial p}\left(\frac{x-p}{p(1-p)}\right) = -\frac{x}{p^2} - \frac{1-x}{(1-p)^2}.$$

Taking expectation gives

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

Hence

$$I(p) = -\,\mathbb{E}\left[\frac{\partial^2}{\partial p^2}\ell(p;X)\right] = \frac{1}{p(1-p)},$$

which agrees with the variance-of-score calculation above.

(5) IID sample of size n. For $X_1, \ldots, X_n$ i.i.d. Bernoulli(p), Fisher information is additive:

$$I_n(p) = n\, I(p) = \frac{n}{p(1-p)}.$$