Figure 4.1: Maximum likelihood estimator

$w = Fv$, we get:

$$v^T \mathbb{F}\mathbb{C}\mathbb{F}v \cdot v^T \mathbb{F}v \geq (v^T \mathbb{F}v)^2, \tag{4.22}$$

$$v^T \mathbb{F}\mathbb{C}\mathbb{F}v \geq v^T \mathbb{F}v. \tag{4.23}$$

Since the above inequality is valid for arbitrary v, this implies

$$\mathbb{F}\mathbb{C}\mathbb{F} \geq \mathbb{F} \Rightarrow \mathbb{C} \geq \mathbb{F}^{-1}, \tag{4.24}$$

where the final result we have obtained by acting on both sides with $\mathbb{F}^{-1}$. ■

Remark. From the derived bound it follows in particular that: $\Delta^2\tilde{\theta}_i \geq (\mathbb{F}^{-1})_{ii} \geq (\mathbb{F}_{ii})^{-1}$, and the last inequality is in general strict if $\mathbb{F}$ contains nonzero off-diagonal elements. To see this consider: $1 = e_i^T \sqrt{\mathbb{F}}\sqrt{\mathbb{F}^{-1}} e_i \overset{\text{C-S}}{\leq} e_i^T \mathbb{F} e_o e_i^T \mathbb{F}^{-1} e_i$, where e_i is the basis vector with 1 at i-th position and zeros elsewhere. This inequality leads to $(\mathbb{F}^{-1})_{ii} \geq 1/\mathbb{F}_{ii}$.

4.4 Maximum likelihood estimator

Typically, we will encounter situations when there is no unbiased estimator that strictly saturates the CR bound for the parameter we want to estimate. We are therefore looking for some universally applicable recipe to find a good estimator.

Definition 4.3 (Maximum likelihood (ML) estimator)**.** Given a probabilistic model, $p_\theta(x)$, the ML estimator is defined as:

$$\tilde{\theta}_{\mathrm{ML}}(x) = \mathrm{argmax}_\theta[l_x(\theta)], \tag{4.25}$$

where $l_x(\theta) = p_\theta(x)$ is the likelihood function, for which θ is the argument.

In other words, given observed event x we look for such a parameter θ for which probability $p_\theta(x)$ (or equivalently the likelihood $l_x(\theta)$) is maximal—the event is most likely, see Fig. **??**. The position of the maximum corresponds to $\tilde{\theta}_{\mathrm{ML}}(x)$.

Remark. In practice, since $l_x(\theta)$ will often be represented as product of many terms (as in e.g. repeated experiment scenarios), it is much more efficient and stable numerically to maximize $\log[l_x(\theta)]$ (the log-likelihood function), as products will turn into sums, and since the log function is monotonic the position of the maximum will remain unchanged.

Example 4.2 Consider our Gaussian example,

$$l_{\boldsymbol{x}}(\theta) = p_\theta(\boldsymbol{x}) = \frac{1}{\sqrt{2\pi\sigma^2}^N} \prod_{i=1}^N e^{-(x_i-\theta)/\sigma^2}. \tag{4.26}$$

For a given $\boldsymbol{x}$ we look for the maximum of l. The condition

$$\frac{\mathrm{d}}{\mathrm{d}\theta} p_\theta(\boldsymbol{x}) = 0, \text{ or equivalently } \frac{\mathrm{d}}{\mathrm{d}\theta} \log p_\theta(\boldsymbol{x}) = 0 \tag{4.27}$$

implies:

$$\tilde{\theta}_{\mathrm{ML}}(\boldsymbol{x}) = \frac{1}{N} \sum_i x_i. \tag{4.28}$$

We see that the ML estimator is actually the same simple estimator we have proven before to be efficient.

The above apparent coincidence of the ML estimator and the efficient estimator is a general feature. Note that the CR bound saturability condition (4.16) implies that if we take θ such that $\frac{\mathrm{d}}{\mathrm{d}\theta} \log p_\theta(\boldsymbol{x}) = 0$, i.e. extremum of the log likelihood function, then the actual efficient estimator $\tilde{\theta}(\boldsymbol{x}) = \theta = \tilde{\theta}_{\mathrm{ML}}(\boldsymbol{x})$ (unless $\lambda(\theta) = 0$, which corresponds to a trivial case of FI equal to zero), so indeed the ML estimator is the efficient estimator.

We will now prove the most important theorem of classical estimation theory, namely the asymptotic efficiency of the ML estimator, which means that the ML estimator will asymptotically saturate the CR bound in a model with large number of identical and independent repetitions of the experiment. We will need the following Lemma:

Lemma 4.1. Let $p_1(x)$, $p_2(x)$ be two probability distributions, then

$$D(p_1|p_2) = \int \mathrm{d}\, x p_1(x) \log \frac{p_1(x)}{p_2(x)} \geq 0, \tag{4.29}$$

where $D(p_1|p_2)$ is called the relative entropy.

Proof. The log function is concave, which means that $\log(\sum_i w_i t_i) \geq \sum_i w_i t_i$, for $w_i \geq 0$, $\sum_i w_i = 1$. Setting $t_x = \frac{p_2(x)}{p_1(x)}$, $w_x = p_1(x)$ and utilizing the concavity of the log function we get

$$\int \mathrm{d}x\, p_1(x) \log \frac{p_2(x)}{p_1(x)} \leq \log \int \mathrm{d}x\, p_2(x) = 0, \tag{4.30}$$

which ends the proof. □

We are now ready to prove the main theorem.

Theorem 4.3. Let

$$p_\theta(\boldsymbol{x}) = p_\theta(x_1) \cdot \cdots \cdot p_\theta(x_N), \tag{4.31}$$

represent the joint probability distribution for N independent repetitions of an experiment. The ML estimator will be asymptotically unbiased and efficient in the limit $N \to \infty$, which formally means:

$$\tilde{\theta}_{\mathrm{ML}} \sim \mathcal{N}\left(\theta_0, \frac{1}{FN}\right), \tag{4.32}$$

where θ_0 is the true value of the parameter and F is the FI corresponding to a single experiment $p_\theta(x_i)$ at $\theta = \theta_0$.

Proof. We start by making some technical assumptions concerning the regularity of $p_\theta(x)$. We assume that that $\log p_\theta(x)$ has derivatives up to order 2 and $\langle \partial_\theta \log p_\theta(x) \rangle = 0$. The proof consists of two parts. First we prove asymptotic unbiasedness and then efficiency.

Asymptotic unbiasedness. Let $\tilde{\theta}$ be an estimator. Let us divide the log-likelihood function at $\tilde{\theta}$ by N:

$$\frac{1}{N} l_{\boldsymbol{x}}(\tilde{\theta}) = \frac{1}{N} \log p_{\tilde{\theta}}(\boldsymbol{x}) = \frac{1}{N} \sum_i \log p_{\tilde{\theta}}(x_i). \tag{4.33}$$

By the law of large numbers, for almost every sequence $\boldsymbol{x}$, we get

$$\frac{1}{N} l_{\boldsymbol{x}}(\tilde{\theta}) \xrightarrow{N\to\infty} \int \mathrm{d}x\, p_{\theta_0}(x) \log p_{\tilde{\theta}}(x) \tag{4.34}$$

where θ_0 is the true value. Using Lemma 4.1 we get

$$\int \mathrm{d}x\, p_{\theta_0}(x) \log p_{\tilde{\theta}}(x) \leq \int \mathrm{d}x\, p_{\theta_0}(x) \log p_{\theta_0}(x). \tag{4.35}$$

This shows that the argument $\tilde{\theta}$ for which we obtain the maximum of $l_{\boldsymbol{x}}(\theta)$, i.e. the ML estimator, in the asymptotic limit $N \to \infty$ will correspond to the true value.

Asymptotic efficiency. We will start by invoking the mean value theorem, which states that assuming $\theta_0 < \tilde{\theta}$ (the order here is not important) there exist $\theta_0 \leq \bar{\theta} \leq \tilde{\theta}$ such that:

$$\frac{\frac{\mathrm{d}\log p_\theta(\boldsymbol{x})}{\mathrm{d}\theta}\Big|_{\theta=\tilde{\theta}} - \frac{\mathrm{d}\log p_\theta(\boldsymbol{x})}{\mathrm{d}\theta}\Big|_{\theta=\theta_0}}{\tilde{\theta} - \theta_0} = \frac{\mathrm{d}^2 \log p_\theta(\boldsymbol{x})}{\mathrm{d}\theta^2}\Big|_{\theta=\bar{\theta}}. \tag{4.36}$$

If $\tilde{\theta} = \tilde{\theta}_{\mathrm{ML}}$ then $\frac{\mathrm{d}\log p_\theta(x)}{\mathrm{d}\theta}\Big|_{\theta=\tilde{\theta}_{\mathrm{ML}}} = 0$ therefore we get

$$\frac{\mathrm{d}\log p_\theta(\boldsymbol{x})}{\mathrm{d}\theta}\Big|_{\theta=\theta_0} = \frac{\mathrm{d}^2 \log p_\theta(\boldsymbol{x})}{\mathrm{d}\theta^2}\Big|_{\theta=\bar{\theta}} (\theta_0 - \tilde{\theta}_{\mathrm{ML}}). \tag{4.37}$$

Let us now consider:

$$\frac{1}{N} \frac{\mathrm{d}^2 \log p_\theta(\boldsymbol{x})}{\mathrm{d}\theta^2}\Big|_{\theta=\bar{\theta}} = \frac{1}{N} \sum_i \frac{\mathrm{d}^2 \log p_\theta(x_i)}{\mathrm{d}\theta^2}\Big|_{\theta=\bar{\theta}}. \tag{4.38}$$

We know that $\tilde{\theta}_{\mathrm{ML}} \xrightarrow{N\to\infty} \theta_0$ and hence $\bar{\theta} \xrightarrow{N\to\infty} \theta_0$. We can therefore write:

$$\frac{1}{N} \frac{\mathrm{d}^2 \log p_\theta(\boldsymbol{x})}{\mathrm{d}\theta^2}\Big|_{\theta=\bar{\theta}} \overset{N\to\infty}{=} \frac{1}{N} \sum_i \frac{\mathrm{d}^2 \log p_\theta(x_i)}{\mathrm{d}\theta^2}\Big|_{\theta=\theta_0} = \left\langle \frac{\mathrm{d}^2 \log p_\theta(x_i)}{\mathrm{d}\theta^2}\Big|_{\theta=\theta_0} \right\rangle = -F. \tag{4.39}$$

Let us define a random variable ξ, which is a sum of N i.i.d variables, as follows:

$$\xi = \frac{1}{\sqrt{N}} \frac{\mathrm{d}\log p_\theta(\boldsymbol{x})}{\mathrm{d}\theta}\Big|_{\theta=\theta_0} = \frac{1}{\sqrt{N}} \sum_i \frac{\mathrm{d}\log p_\theta(x_i)}{\mathrm{d}\theta}\Big|_{\theta=\theta_0}. \tag{4.40}$$

Note that $\langle \xi \rangle = 0$, while the second moment reads:

$$\langle \xi^2 \rangle = \left\langle \frac{1}{N} \left(\sum_i \left. \frac{\mathrm{d} \log p_\theta(x_i)}{\mathrm{d}\theta} \right|_{\theta=\theta_0} \right)^2 \right\rangle = \frac{1}{N} \sum_i \left\langle \left(\left. \frac{\mathrm{d} \log p_\theta(x_i)}{\mathrm{d}\theta} \right|_{\theta=\theta_0} \right)^2 \right\rangle = F. \quad (4.41)$$

By the central limit theorem this implies that $\xi \sim \mathcal{N}(0, F)$. As a result

$$\tilde{\theta}_{\mathrm{ML}} - \theta_0 = \frac{\left. \frac{\mathrm{d} \log p_\theta(\boldsymbol{x})}{\mathrm{d}\theta} \right|_{\theta=\theta_0}}{\left. \frac{\mathrm{d}^2 \log p_\theta(\boldsymbol{x})}{\mathrm{d}\theta^2} \right|_{\theta=\bar{\theta}}} \sim \mathcal{N}\left(0, \frac{NF}{N^2F^2}\right) \quad (4.42)$$

so finally:

$$\tilde{\theta}_{\mathrm{ML}} \sim \mathcal{N}(\theta_0, (NF)^{-1}), \quad (4.43)$$

which shows that asymptotically the maximum likelihood estimator is normally distributed and saturates the CR bound. A priori, it is not clear, however, how large N need to be taken to saturate the bound up to some give accuracy. This depends on the details of the model. □