

---

Often we are interested in the way in which some parameter in the world, but we can't measure it directly. Rather, we have access to measurements whose distribution changes according to the value of the parameter.

- We have a coin and we want to know if it's fair
- We want to measure the firing rate of a neuron
- We want to measure the expected height in a population

We shall denote the parameter of interest by  $\theta$  and the distribution by  $p(x|\theta)$ . This is called a **parameterized family of distributions**.

Our goal will be to estimate  $\theta$  from observations. Typically we will have  $n$  IID observations from this distribution. Our strategy will depend on how we evaluate our performance. Some examples of parametric families

- Bernoulli distribution. For every  $\theta$  define:

$$\begin{aligned}p(x = 1|\theta) &= \theta \\p(x = 0|\theta) &= 1 - \theta\end{aligned}$$

Here  $\Omega = \{\theta | 0 \leq \theta \leq 1\}$ .

- Gaussian (or normal) distributions:

$$p(x|\mu, \sigma) = \mathcal{N}(\mu, \sigma^2)$$

- Poisson
- Categorical. Parameters are  $\theta_1, \dots, \theta_k$  (non-negative) such that  $\sum_i \theta_i = 1$  where  $p(x = i|\theta) = \theta_i$ .

Another important example is where we make measurements of pairs  $(x_i, y_i)$  (for example  $x$  could be intelligence and  $y$  could be grades) and we want to model the relation between  $x$  and  $y$ . This is referred to as the problem of regression. Assume we have a model for  $p(y_i|x_i; \theta)$ , for example:

$$p(y|x; \theta) = \mathcal{N}(\theta_1 x + \theta_2, 1) \tag{1}$$

We will want to estimate the parameters  $\theta$  which specify the linear relation between  $x$  and  $y$ .

## 1 Bayesian Parameter Estimation

We can treat this is a standard decision problem where  $\theta$  is the hidden state of the world. Our action will be to return an estimate of  $\theta$  which we will denote by  $\hat{\theta}(x^n)$  (so analogous to  $\alpha(x)$  before). Assume we have a prior  $p(\theta)$ .

We just need a loss function. Below we review several choices of loss and the resulting estimators.

### 1.1 Squared error - L2 loss

One possibility is the **squared error** loss:

$$\lambda(\hat{\theta}|\theta) = (\hat{\theta} - \theta)^2 \quad (2)$$

It makes sense for continuous parameters since if there is a small difference between  $\theta$  and  $\hat{\theta}$  we don't want to pay much. As we show next the optimal Bayesian strategy is the following estimator:

$$\hat{\theta}_{L2}(x^n) = E_{p(\theta|x^n)}(\theta) = \int \theta p(\theta|x^n) d\theta \quad (3)$$

We use the subscript  $L2$  to denote this is the optimal estimator with respect to the  $L2$  loss. What this means is you should calculate the distribution  $p(\theta|x^n)$  (i.e., the posterior of the parameter given the observations) and take its mean. This is sometimes called the posterior mean.

**Proof:** Bayesian decision theory tell us that the optimal action  $\alpha$  is the one that minimizes the risk  $R(\alpha|x)$ . Since in our case  $\alpha$  corresponds to  $\hat{\theta}$  we write:

$$R[\hat{\theta}|x^n] = \int (\hat{\theta} - \theta)^2 p(\theta|x^n) d\theta \quad (4)$$

Take derivative w.r.t.  $\hat{\theta}$  to get:

$$\begin{aligned} \int (\hat{\theta} - \theta) p(\theta|x^n) d\theta &= 0 \\ \hat{\theta} \int p(\theta|x^n) d\theta &= \int \theta p(\theta|x^n) d\theta \\ \hat{\theta} &= \int \theta p(\theta|x^n) d\theta \end{aligned}$$

**Example (Gaussian family and prior):** consider the case where  $\theta$  is the mean of a Gaussian distribution:

$$p(x|\theta) = \mathcal{N}(x, 1) \quad (5)$$

And the prior is  $p(\theta) = \mathcal{N}(0, 1)$ . Given a set of IID observations, the posterior is then:

$$p(\theta|x^n) = \frac{p(x^n|\theta)p(\theta)}{p(x^n)} \propto p(x^n|\theta)p(\theta) \quad (6)$$

Thus:

$$p(\theta|x^n) = \frac{1}{\sqrt{(2\pi)^n}} e^{-\frac{1}{2}\sum_i(x_i-\theta)^2} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\theta^2} \quad (7)$$

Ignoring the constants for a moment we have:

$$\begin{aligned} p(\theta|x^n) &\propto e^{-\frac{1}{2}\sum_i(x_i-\theta)^2} e^{-\frac{1}{2}\theta^2} \\ &\propto e^{-\frac{1}{2}((n+1)\theta^2 - 2\theta\sum_i x_i)} \\ &\propto e^{-\frac{n+1}{2}(\theta - \frac{1}{n+1}\sum_i x_i)^2} = \mathcal{N}\left(\frac{1}{n+1}\sum_i x_i, \frac{1}{n+1}\right) \end{aligned}$$

The Bayesian estimator is the mean of the posterior and thus we have:

$$\hat{\theta}_{L2}(x^n) = \frac{1}{n+1} \sum_i x_i \quad (8)$$

Some observations about this:

- The estimator behaves as if we had another observation  $x_{n+1} = 0$  (although we only have  $n$  observations). This *dummy* observation is the outcome of using the prior (which in this case is indeed has mean zero).
- The variance of the posterior shrinks with  $n$ . This implies that it will be more strongly peaked around its mean, and hence more *confident* of the estimate. It makes sense since we are getting more observations.
- The posterior is also a Gaussian (same as the prior and the parametric family). When a posterior has the same form as the prior for a given family, we say the prior is a conjugate prior for that family. Thus a Gaussian prior is a conjugate prior for the Gaussian family.

## 1.2 MAP Estimator

What if we want our estimator  $\hat{\theta}$  to equal  $\theta$  exactly. In other words we want our loss to be zero if  $\hat{\theta} = \theta$  and one otherwise. We already encountered this loss when talking about Bayesian decision theory,

$$\lambda(\hat{\theta}|\theta) = \begin{cases} 0 & \hat{\theta} = \theta \\ 1 & \hat{\theta} \neq \theta \end{cases} \quad (9)$$

Alternatively for the continuous case  $\lambda(\hat{\theta}|\theta) = 1 - \delta(\hat{\theta} - \theta)$  where  $\delta$  is the Dirac delta function.

We already know that the optimal strategy in his case:

$$\hat{\theta}_{MAP}(x^n) = \arg \max_{\theta} p(\theta|x^n) \quad (10)$$

Where MAP stands for maximum-aposteriori. It is interesting to see how this behaves as the number of examples grows:

$$\begin{aligned} p(\theta|x^n) &\propto p(\theta)p(x^n|\theta) \\ \log p(\theta|x^n) &= \log p(\theta) + \log p(x^n|\theta) \\ \log p(\theta|x^n) &= \log p(\theta) + \sum_{i=1}^n \log p(x_i|\theta) \end{aligned}$$

It is clear that as the number of samples grows the term  $\log p(\theta)$  has little effect on the posterior. This implies that the prior is negligible and we may think of considering maximizing only the second term. This is indeed a very important approach, called maximum likelihood. Note that it is non-Bayesian in that it does not assume a prior.

$$\hat{\theta}_{ML}(x^n) = \arg \max_{\theta} p(x^n|\theta) \quad (11)$$

The function  $p(x^n|\theta)$  is called the likelihood. It's often simple to maximize it's log instead.

**Example:** In the Gaussian example above, the posterior is a Gaussian:  $p(\theta|x^n) = \mathcal{N}(\frac{1}{n+1} \sum_i x_i, \frac{1}{n+1})$ . The  $\theta$  that maximizes  $p(\theta|x^n)$  is its means and thus  $\hat{\theta}_{MAP}(x^n) = \frac{1}{n+1} \sum_i x_i$  (same as the L2 estimator).

**Example:** The Beta distribution over variable  $0 \leq \theta \leq 1$  is defined as:

$$p(\theta; \alpha, \beta) = \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1} \quad (12)$$

where  $B(\alpha, \beta)$  is a normalization constant known as the Beta function. Its mean is  $\frac{\alpha}{\alpha+\beta}$ . We will next see that it is the conjugate prior of the Bernouli distribution, where the Bernouli family  $p(x; \theta)$  is defined for  $x \in \{0, 1\}$  and

$$p(x|\theta) = \begin{cases} \theta & x = 1 \\ 1 - \theta & x = 0 \end{cases} \quad (13)$$

Say we want to estimate  $\theta$  given  $x$  with a Bernouli family and Beta prior. Lets see what the posterior is:

$$\begin{aligned} p(\theta|x) &\propto \theta^{\alpha-1} (1-\theta)^{\beta-1} \theta^x (1-\theta)^{1-x} \\ &= \theta^{\alpha+x-1} (1-\theta)^{\beta+1-x-1} \end{aligned}$$

Thus we have that  $p(\theta|x) = B(\alpha+x, \beta+1-x)$ . The posterior mean is  $\frac{\alpha+x}{\alpha+\beta+1}$ . Repeating this  $n$  times we will have after  $n$  observations:  $p(\theta|x^n) =$

$B(\alpha + \sum_i x_i, \beta + 1 - \sum_i x_i)$ . The posterior mean is thus:  $\frac{\alpha + \sum_i x_i}{\alpha + \beta + n}$ . This has the following nice interpretation: assume that originally we had  $\alpha$  successes ( $x = 1$ ) and  $\beta$  failures ( $x = 0$ ). We now add  $n$  data points to this initial data, and the resulting estimate is exactly the proportion of successes in this new data.

The maximum likelihood in this case is as follows. The likelihood function is:

$$p(x^n|\theta) = \prod_i p(x_i|\theta) = \prod_i \theta^{x_i} (1 - \theta)^{1-x_i}$$

Take the log to get:

$$\sum_i [x_i \log \theta + (1 - x_i) \log(1 - \theta)] \quad (14)$$

Now take the derivative w.r.t.  $\theta$  and equate to zero:

$$\begin{aligned} \frac{1}{\theta} \sum_i x_i + \frac{1}{1-\theta} \sum_i x_i - \frac{n}{1-\theta} &= 0 \\ \frac{1}{\theta(1-\theta)} \sum_i x_i &= \frac{n}{1-\theta} \\ \hat{\theta}(x^n) &= \frac{1}{n} \sum_i x_i \end{aligned}$$

**Example: Maximum likelihood for Gaussian with unknown mean and variance.** Assume our parametric family is:

$$p(x|\theta_1, \theta_2) = \mathcal{N}(\theta_1, \theta_2) \quad (15)$$

What is the ML estimate for these two given a sample  $x^n$ . Write the likelihood:

$$p(x^n|\theta_1, \theta_2) = \frac{1}{(\sqrt{2\pi\theta_2})^n} e^{-\frac{1}{2\theta_2} \sum_i (x_i - \theta_1)^2} \quad (16)$$

We want to find the  $\theta_1, \theta_2$  that maximize it. We can equivalently maximize its log given by:

$$\log p(x^n|\theta_1, \theta_2) = -\frac{n}{2} \log(2\pi\theta_2) - \frac{1}{2\theta_2} \sum_i (x_i - \theta_1)^2 \quad (17)$$

Take the derivative w.r.t.  $\theta_1$  to get:

$$\frac{\partial \log p(x^n|\theta_1, \theta_2)}{\partial \theta_1} = -\frac{1}{\theta_2} \sum_i (x_i - \theta_1) = 0 \quad (18)$$

From which we conclude:

$$\hat{\theta}_1(x^n) = \frac{1}{n} \sum_i x_i \quad (19)$$

This gives us  $\theta_1$ . What about  $\theta_2$ . Take the derivative w.r.t.  $\theta_2$ :

$$\frac{\partial \log p(x^n | \theta_1, \theta_2)}{\partial \theta_1} = -\frac{n}{2\theta_2} + \frac{1}{2\theta_2^2} \sum_i (x_i - \hat{\theta}_1(x^n))^2 = 0 \quad (20)$$

From which we conclude:

$$\hat{\theta}_2(x^n) = \frac{1}{n} \sum_i (x_i - \hat{\theta}_1(x^n))^2 \quad (21)$$

**Example (exponential distribution with exponential prior):** Say your family is  $p(x|\theta) = \theta e^{-\theta x}$  (the exponential distribution). And the prior is  $p(\theta) = e^{-\theta}$ . Note that the mean of the exponential distribution is  $\frac{1}{\theta}$ . Lets see what the MAP and ML estimators are The posterior is:

$$p(\theta|x^n) \propto e^{-\theta n} e^{-\theta \sum_i x_i} \quad (22)$$

In log it is:

$$\log p(\theta|x^n) = -\theta + n \log \theta - \theta \sum_i x_i + c \quad (23)$$

Set derivative to zero to get the MAP:

$$\begin{aligned} -1 + \frac{n}{\theta} - \sum_i x_i &= 0 \\ \hat{\theta}_{MAP}(x^n) &= \frac{n}{\sum_i x_i + 1} \end{aligned}$$

If we just wanted to maximize the likelihood we would get:

$$\hat{\theta}_{ML}(x^n) = \frac{n}{\sum_i x_i}$$

This is exactly the inverse of the average so it makes sense.

## 2 The non-Bayesian view

When we are not “allowed” to consider priors, we can only talk about the performance of the method conditioned on values of  $\theta$ . So, assume we have an estimator  $\hat{\theta}(x^n)$ . Now:

- Fix  $\theta$

- Generate samples from  $p(x^n|\theta)$ .
- Calculate the estimator  $\hat{\theta}(x^n)$ .

We will get different values for this and they will have some distribution. Example: Say our estimator is  $\hat{\theta}(x^n) = \frac{1}{n} \sum_i x_i$  of  $\mu$  in a normal distribution with unknown mean  $\mu$  (i.e.,  $\mathcal{N}(\mu, \sigma^2)$  with known variance  $\sigma^2$ ). It will be distributed as  $\mathcal{N}(\mu, \sigma^2/n)$ .

Ideally we would like the estimator to be always “close” to  $\hat{\theta}$ . But since we want to support multiple possible  $\theta$  values, we will have to make some errors...

We now consider several properties we would like our estimator to have.

## 2.1 Consistency

**Reminder:** As we get more and more data, we would expect  $\hat{\theta}(x^n)$  to get close to  $\theta$  regardless of the value of the latter. If an estimator has this property it is said to be consistent. Formally: An estimator is consistent if for every value of  $\theta$  the following holds in probability for samples drawn from  $p(x^n|\theta)$ :

$$\hat{\theta}(x^n) \rightarrow \theta \quad (24)$$

We now prove that maximum likelihood is consistent. This holds under several regularity conditions (things being differentiable bounded etc). But something that certainly should hold is that parameters uniquely define the distribution. For example if our parametric family had  $\mathcal{N}(\mu, \sigma^2)$  and  $\mathcal{N}(2\mu, \sigma^2)$  there would be two values of  $\mu$  that give the same distribution. Lets assume this can't happen (the model is identifiable). In other words  $D_{KL}[p(x|\theta_1)|p(x|\theta_2)] = 0$  iff  $\theta_1 = \theta_2$ . Assume  $\theta$  generate

$$\begin{aligned} \lim_{n \rightarrow \infty} \hat{\theta}(x_n) &= \lim_{n \rightarrow \infty} \arg \max_{\theta'} \frac{1}{n} \log p(x^n|\theta') \\ &= \arg \max_{\theta'} \lim_{n \rightarrow \infty} \frac{1}{n} \log p(x^n|\theta') \\ &= \arg \max_{\theta'} \lim_{n \rightarrow \infty} \frac{1}{n} \log p(x^n|\theta') = \arg \max_{\theta'} \lim_{n \rightarrow \infty} \frac{1}{n} \sum_i p(x_i|\theta') \\ &= \arg \max_{\theta'} \sum_x p(x|\theta) \log p(x|\theta') \\ &= \arg \max_{\theta'} \left[ -D_{KL}[p(x|\theta)|p(x|\theta')] + \sum_x p(x|\theta) \log p(x|\theta) \right] \\ &= \arg \max_{\theta'} [-D_{KL}[p(x|\theta)|p(x|\theta')]] \\ &= \theta \end{aligned}$$

This gives another interesting insight on maximum likelihood:

$$\begin{aligned} \frac{1}{n} \log p(x^n | \theta') &= \frac{1}{n} \sum_i \log p(x^i | \theta') \\ &= \sum_i p_e(x) \log p(x | \theta') \\ &= -D_{KL}[p_e(x) | p(x | \theta')] + \sum_x p_e(x) \log p_e(x) \end{aligned}$$

So the  $\theta'$  that maximizes the above minimizes  $D_{KL}[p_e(x) | p(x | \theta')]$ . In other words, maximum likelihood is the distribution in the family that is closely (in the KL sense) to the empirical distribution.

## 2.2 Unbiasedness

It might seem like a good idea to have the expected value of  $\hat{\theta}(x^n)$  be  $\theta$ . But note there might be some good estimators that do not satisfy this requirement. Some examples are:

- Bernouli variable:  $\hat{\theta}(x^n) = \frac{1}{n} \sum_i x_i$ . The expected value is the expected value of  $X$  which is  $\theta$ .
- Gaussian with unknown variance:  $\hat{\sigma}^2(x^n) = \frac{1}{n} \sum_i (x_i - \hat{\mu}(x^n))^2$ . This is actually biased (its expected value is  $\frac{n-1}{n} \sigma^2$ ) but it turns out that  $\frac{1}{n-1} \sum_i (x_i - \hat{\mu}(x^n))^2$  is unbiased.

So should we always look for an unbiased estimator? Not necessarily. We could have an estimator that is “close” to  $\theta$  for most samples, even if its average is not exactly correct.

## 2.3 The MSE of an estimator

What we really care about is the expected value of  $(\theta - \hat{\theta}(x^n))^2$  for an estimator.

$$R(\hat{\theta}(x^n) | \theta) = \sum_{x^n} p(x^n | \theta) (\theta - \hat{\theta}(x^n))^2 = E_{p(x^n | \theta)} \left[ (\hat{\theta}(x^n) - \theta)^2 \right] \quad (25)$$

This is also called the mean squared error (MSE). How is this related to the bias? The distribution of the estimator has two key properties in that respect. One is its bias:

$$b(\hat{\theta}(x^n) | \theta) = E_{p(x^n | \theta)} [\hat{\theta}(x^n)] - \theta \quad (26)$$

The other is its variance:

$$V(\hat{\theta}(x^n) | \theta) = E_{p(x^n | \theta)} \left[ (\hat{\theta}(x^n) - E(\hat{\theta}(x^n)))^2 \right] \quad (27)$$

We now show that the MSE nicely decomposes into these two measures.

$$\begin{aligned}
E_{p(x^n|\theta)} \left[ \left( \hat{\theta}(x^n) - \theta \right)^2 \right] &= E_{p(x^n|\theta)} \left[ \left( \hat{\theta}(x^n) - E(\hat{\theta}) + E(\hat{\theta}) - \theta \right)^2 \right] \\
&= E \left[ \left( \hat{\theta}(x^n) - E(\hat{\theta}(x^n)) \right)^2 \right] - 2E \left[ \left( \hat{\theta}(x^n) - E(\hat{\theta}(x^n)) \right) \left( E(\hat{\theta}(x^n)) - \theta \right) \right] + E \left[ \left( E(\hat{\theta}(x^n)) - \theta \right)^2 \right] \\
&= E \left[ \left( \hat{\theta}(x^n) - E(\hat{\theta}(x^n)) \right)^2 \right] + E \left[ \left( E(\hat{\theta}(x^n)) - \theta \right)^2 \right] \\
&= E \left[ \left( \hat{\theta}(x^n) - E(\hat{\theta}(x^n)) \right)^2 \right] + E(\hat{\theta}(x^n) - \theta)^2 \\
&= V(\hat{\theta}(x^n)|\theta) + b^2(\hat{\theta}(x^n)|\theta)
\end{aligned}$$

How can we come up with a strategy that is good across different values of  $\theta$ . It's pretty clear we can't. For example the estimator  $\hat{\theta}(x^n) = 3$  would be perfect (MSE=0) if  $\theta = 3$  but pretty lousy otherwise. Note that it will have zero variance but a bad bias for most values of  $\theta$ . On the other hand, if for a Gaussian case we take  $\hat{\theta}(x^n) = \frac{1}{n} \sum_i x_i$  it will be unbiased but have non-zero variance since it will return different values for different  $x^n$  samples. This is a rough illustration of the no free lunch theorem that says that we can generally not build a perfect estimator.

Some notes about bias-variance:

- Generally restricting the size of the possible  $\theta$  set will tend to increase variance but reduce bias.
- Generally increasing the number of samples using the same estimator will reduce variance so that we can use a lower bias estimator.
- Biased estimators may not be consistent asymptotically but can work better than consistent estimators for finite sample sizes.
- This analysis is when averaging over all possible samples. We would prefer to be able to say something about the generalization ability of a specific sample. We did it for consistency, but would like to say more. We will do this later in the course.

So, it seems like if we constrain the values of  $\theta$  that we consider, we will lower the variance but probably increase bias. If our constraint is met in reality we will have lowered both bias and variance. To illustrate this further, consider a set of pairs  $(t_i, x_i)$  where  $x_i$  is distributed as  $\mathcal{N}(\theta(t_i), 1)$  for some  $\theta(t)$  we don't know. We would like to recover  $\theta(t_i)$  via some estimator  $\hat{\theta}(x^n; t_i)$ . Note we do not have the *IID* assumption for  $x$  anymore. Assume the  $t$  are fixed throughout.

it can be seen that:

$$\sum_i E_{p(x^n|\theta(t_i))} (\hat{\theta}(x^n; t_i) - \theta(t_i))^2 = \sum_i \left[ V(\hat{\theta}(x^n; t_i)|\theta(t)) + b^2(\hat{\theta}(x^n; t_i)|\theta(t)) \right] \tag{28}$$

Now, assume  $\hat{\theta}(x^n; t_i)$  is obtained via a model:

$$x_i \sim \mathcal{N}\left(\sum_{k=0}^d \beta_k t_i^k, \sigma^2\right)$$

So  $\hat{\theta}(x^n; t_i)$  is calculated as follows:

- Get the ML estimates for  $\hat{\beta}_k$ .
- Calculate  $\hat{\theta}(x^n; t_i) = \sum_{k=0}^d \hat{\beta}_k t_i^k$

You will show in the exercise that in this case the total variance is given by:

$$\sum_i V(\hat{\theta}(x^n; t_i) | \theta(t)) = (d-1)\sigma^2 \quad (29)$$

So as you take more dimensions your variance grows.

### 3 Minimum Variance Unbiased Estimators - Fisher Information

Suppose we stick with unbiased estimators (note that there is no general real reason to, but it is an interesting case to study). The bias-variance decomposition implies that with bias zero, minimizing MSE is equivalent to minimizing the variance of the estimator.

We say that an estimator is a minimum variance unbiased estimator (MVUB) if it unbiased and its variance is lower than or equal to the variance of any other unbiased estimator for any value of  $\theta$ . Do such estimators exist? Yes, but for specific forms of distributions.

Our plan is as follows:

- Provide a lower bound on  $V(\hat{\theta}(x^n) | \theta)$  that holds for all **unbiased** estimators.
- Show estimators that attain it (and say for which distribution families this is possible)

The minimum variance of an estimator should be somehow related to the way  $p(x|\theta)$  changes with  $\theta$ . We begin with a scalar  $\theta$ .

Given a parametric family  $p(x|\theta)$ , define the **Score function**,  $S(\theta)$  as follows:

$$S(x, \theta) = \frac{\partial \log p(x|\theta)}{\partial \theta} \quad (30)$$

### 3 MINIMUM VARIANCE UNBIASED ESTIMATORS - FISHER INFORMATION

---

This measures how much the likelihood at  $x$  changes when  $\theta$  is changed infinitesimally. As  $\theta$  changes some  $x$  will have higher probabilities and some lower. Overall the change will be zero, as the following lemma states:

$$E_{p(x|\theta)} [S(x, \theta)] = 0 \quad \forall \theta \quad (31)$$

**Proof:**

$$\begin{aligned} E_{p(x|\theta)} [S(x, \theta)] &= \int \frac{\partial \log p(x|\theta)}{\partial \theta} p(x|\theta) dx \\ &= \int \frac{1}{p(x|\theta)} \frac{\partial p(x|\theta)}{\partial \theta} p(x|\theta) dx \\ &= \int \frac{\partial p(x|\theta)}{\partial \theta} dx \\ &= \frac{\partial}{\partial \theta} \int p(x|\theta) dx = 0 \end{aligned}$$

For the next to last step, we assume that integration and differentiation can be interchanged. This is true under appropriate smoothness conditions.

So this is not a good measure of parameter sensitivity. The variance of the score turns out to be very useful.

Define the **Fisher Information**  $J(\theta)$  to be the variance of the score function:

$$J(\theta) = E_{p(x|\theta)} [S^2(x, \theta)] = \int \left[ \frac{\partial \log p(x|\theta)}{\partial \theta} \right]^2 p(x|\theta) dx \quad (32)$$

Another property of Fisher information is that

$$J(\theta) = -E_{p(x|\theta)} \frac{\partial^2 \log p(x|\theta)}{\partial \theta^2} \quad (33)$$

And also  $J_n(\theta) = nJ(\theta)$ .

The **Cramer Rao** theorem (1944) relates the Fisher information to the minimum possible variance of an estimator of  $\theta$ . It says that for any **unbiased** estimator  $\hat{\theta}(x)$  it holds that:

$$V(\hat{\theta}(x^n)|\theta) \geq \frac{1}{J(\theta)} \quad (34)$$

Before proving it we need the following lemma (equivalent to the Cauchy Schwartz inequality): For two random variables  $X, Y$  it holds that

$$E^2(XY) \leq E(X^2)E(Y^2) \quad (35)$$

with equality iff  $X = aY$  for some constant  $a$ . Equivalently

$$\begin{aligned} E^2((X - E(X))(Y - E(Y))) &\leq E((X - E(X))^2)E((Y - E(Y))^2) \\ Cov^2(X, Y) &\leq V(X)V(Y) \end{aligned}$$

(the ratio of the LHS to the RHS is called the correlation coefficient). Note also that  $Cov(X, Y) = E(XY) - E(X)E(Y)$  and that equality holds iff  $Y = aX + b$  for some constants  $a, b$ .

**Proof (Cramer Rao):** Throughout we assume we are given some estimator  $\hat{\theta}(x)$ . Consider the expected value of  $\hat{\theta}(x)S(x, \theta)$ . We are interested in:

$$E_{p(x|\theta)}^2 \left[ \hat{\theta}(x)S(x, \theta) \right] \quad (36)$$

On the one hand:

$$\begin{aligned} E_{p(x|\theta)} \left[ \hat{\theta}(x)S(x, \theta) \right] &= \int \frac{\partial \log p(x|\theta)}{\partial \theta} \hat{\theta}(x) p(x|\theta) dx \\ &= \int \frac{\partial p(x|\theta)}{\partial \theta} \hat{\theta}(x) dx \\ &= \frac{\partial}{\partial \theta} \int p(x|\theta) \hat{\theta}(x) dx = \frac{\partial}{\partial \theta} \theta = 1 \end{aligned}$$

On the other hand:

$$\begin{aligned} E_{p(x|\theta)}^2 \left[ \hat{\theta}(x)S(x, \theta) \right] &= E_{p(x|\theta)} \left[ \hat{\theta}(x)S(x, \theta) - E(\hat{\theta})E(S(x, \theta)) \right]^2 \\ &= E_{p(x|\theta)} \left[ \hat{\theta}(x)S(x, \theta) - E(\hat{\theta})E(S(x, \theta)) \right] \\ &= Cov \left[ \hat{\theta}(x), S(x, \theta) \right] \\ &\leq V(\hat{\theta}(x)|\theta)J(\theta) \end{aligned}$$

Which gives the theorem.

### 3.1 Efficiency and exponential families

An unbiased estimator is said to be **efficient** it attains the Cramer Rao inequality with equality.

Here we have equality iff

$$\frac{\partial}{\partial \theta} \log p(x|\theta) = S(x, \theta) = B_1(\theta)\hat{\theta}(x) + D_1(\theta) \quad (37)$$

Integrate w.r.t.  $\theta$

$$\begin{aligned} \log p(x|\theta) &= B(\theta)\hat{\theta}(x) + D(\theta) + C(x) \\ p(x|\theta) &= e^{B(\theta)\hat{\theta}(x)+D(\theta)+C(x)} \end{aligned}$$

A family that has this form is called an exponential family of distributions. So we have the theorem: If a family is of the form:

$$p(x|\theta) = e^{A(x)B(\theta)+D(\theta)+C(x)} \quad (38)$$

and  $A(x)$  is an unbiased estimator of  $\theta$  then  $A(x)$  obtains the Cramer Rao bound and is the minimum variance unbiased estimator (the reverse is also true).

### 3.2 Gaussian Example

Assume a Gaussian family with known variance  $\sigma^2$  and unknown mean. Consider the ML estimator  $\hat{\mu} = \frac{1}{n} \sum_i x_i$ . It is unbiased. The variance is  $V(\hat{\mu}|\mu) = \frac{\sigma^2}{n}$ . Lets calculate the Fisher information for it:

$$\frac{\partial^2 \log p(x|\mu)}{\partial \mu^2} = -\frac{1}{\sigma^2} \quad (39)$$

It's independent of  $x$  so the expected value is the same, and we have

$$J_n(\mu) = \frac{n}{\sigma^2} \quad (40)$$

So the variance of the ML estimate is  $\frac{1}{J_n(\mu)}$  and is thus optimal for an unbiased estimator.

### 3.3 Efficiency and Maximum Likelihood

We have already seen that the ML estimator is consistent (i.e., also asymptotically unbiased). Is it efficient? Turns out it asymptotically both normal and efficient, that is as  $n \rightarrow \infty$   $\hat{\theta}_{ML} \sim \mathcal{N}(\theta, \frac{1}{nJ(\theta)})$

Write a Taylor expansion of  $\ell'(\theta)$  around  $\hat{\theta}_{ML}$

$$\begin{aligned} 0 &= \ell'(\hat{\theta}_{ML}) \approx \ell'(\theta) + (\hat{\theta}_{ML} - \theta)\ell''(\theta) \\ \hat{\theta}_{ML} - \theta &= -\frac{\ell'(\theta)}{\ell''(\theta)} \\ \sqrt{n}(\hat{\theta}_{ML} - \theta) &= -\frac{\frac{1}{\sqrt{n}}\ell'(\theta)}{\frac{1}{n}\ell''(\theta)} \end{aligned}$$

The top is:

$$\frac{1}{\sqrt{n}} \sum_i S(x_i, \theta) \quad (41)$$

It thus converges to a normal  $\mathcal{N}(0, J(\theta))$ . The bottom is:

$$-\frac{1}{n} \sum_i \frac{\partial^2 \log p(x|\theta)}{\partial \theta^2} \rightarrow J(\theta) \quad (42)$$

From which it follows that  $\sqrt{n}(\hat{\theta}_{ML} - \theta)$  is approximately distributed as  $\mathcal{N}(0, \frac{1}{J(\theta)})$

## 4 Multivariate version of Carmer Rao

Thus far we looked at a scalar parameter  $\theta$ . What happens if we have multiple parameters  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)$ . Say we have an unbiased estimator of these  $\boldsymbol{\theta}(x^n)$ . What is the minimum variance? In this case we can actually talk about the co-variance. We can bound it via the Fisher information matrix.

$$J_{ij}(\boldsymbol{\theta}) = E_{p(x|\boldsymbol{\theta})} \left[ \frac{\partial \log p(x|\boldsymbol{\theta})}{\partial \theta_i} \frac{\partial \log p(x|\boldsymbol{\theta})}{\partial \theta_j} \right] \quad (43)$$

This is a matrix. For any unbiased estimator  $\hat{\boldsymbol{\theta}}(x)$  of  $\boldsymbol{\theta}$  it can be shown that:

$$\text{Cov}(\hat{\boldsymbol{\theta}}|\boldsymbol{\theta}) \succeq J^{-1}(\boldsymbol{\theta}) \quad (44)$$

Specifically  $V(\hat{\theta}_i|\boldsymbol{\theta}) \geq J_{ii}^{-1}$ .