Lecture 8 - Bayesian Experimental Design

OBJECTIVE: We address the problem of myopic Bayesian experimental design. We demonstrate the approach on a simple linear regression problem, where the design problem is to choose an input point and request a label for the chosen input. This problem is also known as active learning or value of information. We start with linear models with Gaussian distributions because these have analytical solutions. These solutions will however be applicable to nonlinear models using Bayesian kernel methods known as Gaussian processes.

◇ BAYESIAN MYOPIC DESIGNS

Formally, the decision problem consists of the following elements:

- **Experiments e**: These are the actions. Which experiment should we conduct? Which question should we ask?

- **Parameters $\boldsymbol{\theta}$**: These are the states. That is, the parameters of the model we are using to represent reality. In Bayesian modelling, we assume that we have a prior distribution on the states.
- **Data y**: In the Bayesian setting, we no longer have access to the states. Instead we make observations after conducting an experiment and use these to update our beliefs about the states. A priori, we don't know which observations our experiment will produce.
- **Utilities** u: These are our standard rewards.

After conducting an experiment, we make observations and use these to update our prior $p(\boldsymbol{\theta})$ into the posterior beliefs $p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{e})$. This posterior becomes the updated prior for the next decision stage.

Schematically, we can view our problem with a decision tree on a sequential game:

⋆

The rational way of choosing the best experiment, given that we don't know what data it will produce and that the states are hidden, is to compute the maximum expected utility of the experiment:

$$u^\star(\mathbf{e}) = \max_{\mathbf{e}} \int \int u(\mathbf{y}, \mathbf{e}) p(\mathbf{y}|\boldsymbol{\theta}, \mathbf{e}) p(\boldsymbol{\theta}) d\boldsymbol{\theta} d\mathbf{y}$$

The term $u(\mathbf{y}, \mathbf{e})$ is the expected utility over the posterior parameters $\boldsymbol{\theta}'$:

$$u(\mathbf{y}, \mathbf{e}) = \int u(\mathbf{y}, \mathbf{e}, \boldsymbol{\theta}') p(\boldsymbol{\theta}'|\mathbf{y}, \mathbf{e}) d\boldsymbol{\theta}'$$

Hence,

$$u^\star(\mathbf{e}) = \max_{\mathbf{e}} \int \int \int u(\mathbf{y}, \mathbf{e}, \boldsymbol{\theta}') p(\boldsymbol{\theta}'|\mathbf{y}, \mathbf{e}) p(\mathbf{y}|\boldsymbol{\theta}, \mathbf{e}) p(\boldsymbol{\theta}) d\boldsymbol{\theta} d\mathbf{y} d\boldsymbol{\theta}'$$

This utility can also be written as follows:

$$u^\star(\mathbf{e}) = \max_{\mathbf{e}} \int \int u(\mathbf{y}, \mathbf{e}, \boldsymbol{\theta}') p(\boldsymbol{\theta}'|\mathbf{y}, \mathbf{e}) p(\mathbf{y}|\mathbf{e}) d\mathbf{y} d\boldsymbol{\theta}'$$

⋆ Proof:

What are sensible choices of $u(\mathbf{y}, \mathbf{e})$?

◇ SHANNON ENTROPY

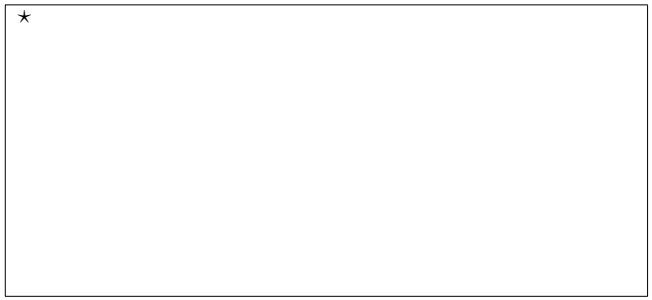
Entropy is an information theory measure of uncertainty. The entropy of a distribution $p(\mathbf{x})$ is defined as follows:

$$H(x) = -\int [\log p(\mathbf{x})] p(\mathbf{x}) d\mathbf{x}$$

Information is defined as the negative entropy. As an example, consider a possibly biased coin with $p(x = tails) = \boldsymbol{\theta}$. Then,

$$H(x) = -\boldsymbol{\theta} \log \boldsymbol{\theta} - (1 - \boldsymbol{\theta}) \log(1 - \boldsymbol{\theta})$$

and its plot looks like:



◇ BAYESIAN D-DESIGNS

We can use the entropy idea to choose which question will result in us learning a better posterior. That is, we choose to optimize the Shannon information of the posterior distribution:

$$u^{\star}(\mathbf{e}) = \max_{\mathbf{e}} \int \int \int \left[\log p(\boldsymbol{\theta}'|\mathbf{y}, \mathbf{e})\right] p(\boldsymbol{\theta}'|\mathbf{y}, \mathbf{e}) d\boldsymbol{\theta}' p(\mathbf{y}, \boldsymbol{\theta}|\mathbf{e}) d\boldsymbol{\theta} d\mathbf{y}$$

We will see later, in the linear model, that this criterion says that we will learn the most by asking questions where we know the least.

◇ BAYESIAN A-DESIGNS

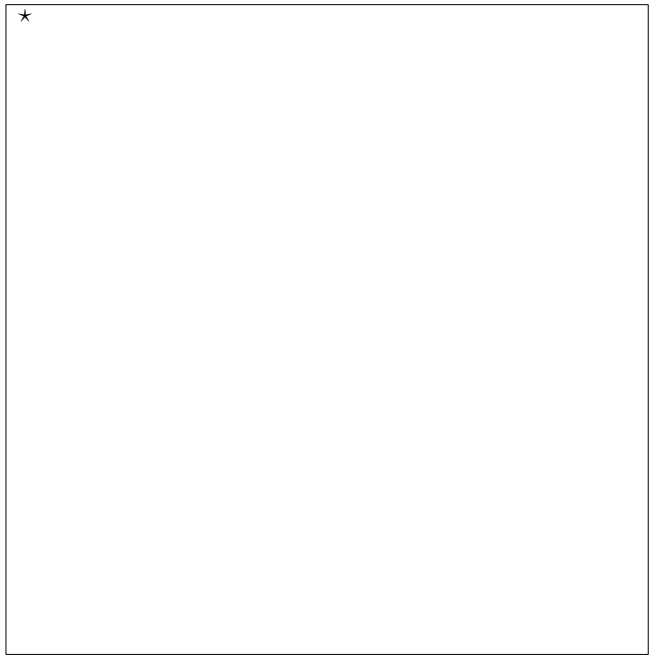
A-optimality is a popular strategy for gathering information to identify the parameters of the model. In particular, we can choose to optimize the following quadratic utility:

$$u^\star(\mathbf{e}) = \max_{\mathbf{e}} - \int\int\int (\boldsymbol{\theta}'-\boldsymbol{\theta})^T A(\boldsymbol{\theta}'-\boldsymbol{\theta}) p(\boldsymbol{\theta}'|\mathbf{y},\mathbf{e}) d\boldsymbol{\theta}' p(\mathbf{y},\boldsymbol{\theta}|\mathbf{e}) d\boldsymbol{\theta} d\mathbf{y}$$

where A is a symmetric positive matrix that emphasizes which parts of the parameter space we consider more relevant. For each possible world (weighted by the prior), the goal is to make a good prediction of $\boldsymbol{\theta}$ in squared-error loss. That is $\boldsymbol{\theta}'$ can be thought of as our identification or guess of the hidden possible world $\boldsymbol{\theta}$. The aim is to improve this identification.

◇ LINEAR SUPERVISED REGRESSION

Given the data $\{x_{1:n}, y_{1:n}\}$, with $x_i \in \mathbb{R}^q$ and $y_i \in \mathbb{R}$, we want to fit a hyper-plane that maps features of the input $f(x) \in \mathbb{R}^d$ to y.



Mathematically, the linear model prediction is expressed as follows:

$$\widehat{\mathbf{y}} = \mathbf{X}\boldsymbol{\theta}$$

$$\begin{bmatrix} \widehat{y}_1 \\ \vdots \\ \widehat{y}_n \end{bmatrix} = \begin{bmatrix} f_1(\mathbf{x}_1) & \cdots & f_d(\mathbf{x}_1) \\ \vdots & \vdots & \vdots \\ f_1(\mathbf{x}_n) & \cdots & f_d(\mathbf{x}_n) \end{bmatrix} \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_d \end{bmatrix}$$

If we have several outputs $y_i \in \mathbb{R}^c$, our linear regression expression becomes:

⋆

◇ MAXIMUM LIKELIHOOD

If our errors are Gaussian distributed, we can use the model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \mathcal{N}(0, \sigma^2 I)$$

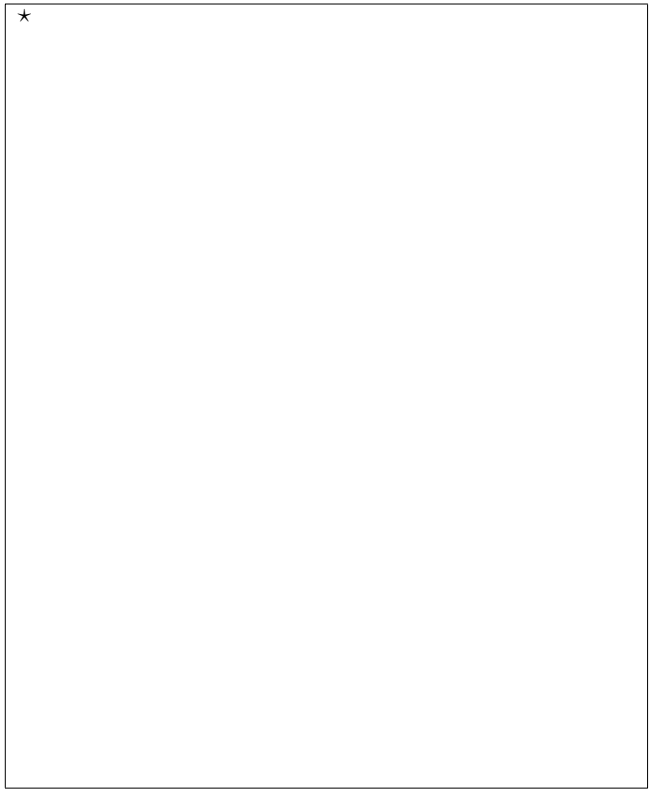
Note that the mean of $\mathbf{y}$ is $\mathbf{X}\boldsymbol{\theta}$ and that its covariance is $\sigma^2 I$, where we assume that σ^2 is known.

We can equivalently write this expression using the probability density of $\mathbf{y}$ **given** $\mathbf{X}$ and $\boldsymbol{\theta}$:

$$p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) = \left(2\pi\sigma^2\right)^{-n/2} e^{-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\theta})^T(\mathbf{y}-\mathbf{X}\boldsymbol{\theta})}$$

The maximum likelihood (ML) estimate of $\boldsymbol{\theta}$ is obtained by taking the derivative of the log-likelihood, $\log p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})$, and equating to zero. The idea of maximum likelihood learning is to maximise the likelihood of seeing some data $\mathbf{y}$ by mod-

ifying the parameters ($\boldsymbol{\theta}$).



◇ BAYESIAN LEARNING FOR LINEAR-GAUSSIAN MODELS

In Bayesian learning we incorporate our prior beliefs or **preferences** into the model. We are interested in the posterior distribution:

$$p(\boldsymbol{\theta}|\mathbf{X},\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{X},\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y}|\mathbf{X})}$$

If we choose a Gaussian prior $\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\theta}_0, \sigma^2 R^{-1})$. Then, the posterior is proportional to:

$$p(\boldsymbol{\theta}|\mathbf{X},\mathbf{y}) \propto \left(2\pi\sigma^2\right)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\theta})^T(\mathbf{y}-\mathbf{X}\boldsymbol{\theta})} \left|2\pi\sigma^2 R^{-1}\right|^{-\frac{1}{2}} e^{-\frac{1}{2\sigma^2}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^T R(\boldsymbol{\theta}-\boldsymbol{\theta}_0)}$$

Our task is to rearrange terms in the exponents in order to obtain a simple analytical expression for the posterior distribution.

$\star$

$$p(\boldsymbol{\theta}|\mathbf{X},\mathbf{y}) \;=\; \left|2\pi\sigma^2\Sigma\right|^{-\frac{1}{2}} e^{-\frac{1}{2\sigma^2}(\boldsymbol{\theta}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu})}$$

So the posterior for $\boldsymbol{\theta}$ is Gaussian with **sufficient statistics**:

$$\begin{aligned}\mathbb{E}(\boldsymbol{\theta}|\mathbf{X},\mathbf{y}) &= (\mathbf{X}^T\mathbf{X}+R)^{-1}(\mathbf{X}^T\mathbf{y}+R\boldsymbol{\theta}_0)\\ cov(\boldsymbol{\theta}|\mathbf{X},\mathbf{y}) &= (\mathbf{X}^T\mathbf{X}+R)^{-1}\sigma^2\end{aligned}$$

The maximum a posteriori (MAP) point estimate is:

$$\widehat{\theta}_{MAP} \;=\; (\mathbf{X}^T\mathbf{X}+R)^{-1}(\mathbf{X}^T\mathbf{y}+R\boldsymbol{\theta}_0)$$

A flat ("vague") prior with large variance and zero mean leads to the ML estimate.

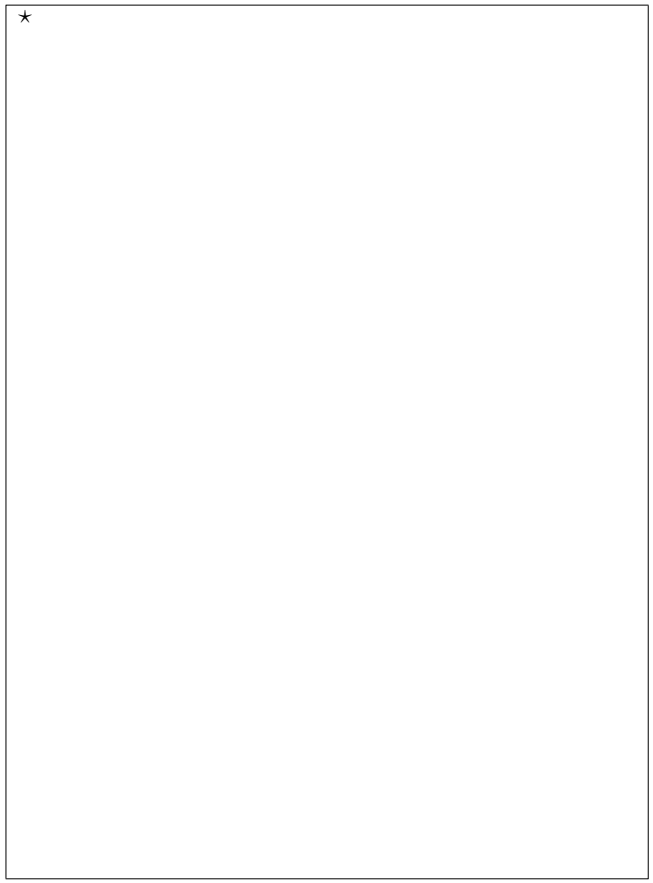
◇ LINEAR EXPERIMENTAL DESIGN

Now that we know how to compute the posterior distribution, we go back to the original problem. We know a finite set of locations $\mathbf{x}$, but we don't know the label $\mathbf{y}$ as these are assumed to be expensive to obtain. Here the experiment is to choose a data point $\mathbf{e} = \mathbf{x}$ that is best to learn a good posterior (one that concentrates its mass on the true value of the parameters).

Using our Gaussian posterior, the optimal D-design simplifies to:

$$u^\star(\mathbf{e}) = \max_{\mathbf{e}} \log |\sigma^{-2}(\mathbf{X}^T\mathbf{X} + R)|$$

That is, we maximize the product of eigenvalues of the information matrix.



The Bayesian A-Design with the Gaussian posterior results in the following criterion:

$$u^{\star}(\mathbf{e}) = \max_{\mathbf{e}} -\sigma^2 trace\left\{A(\mathbf{X}^T\mathbf{X} + R)^{-1}\right\}$$

That is, we minimize the sum of eigenvalues of the posterior covariance (inverse of the posterior information matrix in the Gaussian case), weighted by A.

There are other design criteria. In the following section, we will consider the entropy of the predictive distribution of the data.