Statistical Decision Problems

Bernstein workshop on Sensing and Deciding in Time

Christos Dimitrakakis

Frankfurt Institute for Advanced Studies, Goethe University, Germany

November 4, 2010

1 Statistical decision problems
- Rewards that depend on the outcome of an experiment
- Concavity of the Bayes risk

2 Specific decision problems
- Model estimation
- Time-domain models

1 Statistical decision problems
- Rewards that depend on the outcome of an experiment
- Concavity of the Bayes risk

2 Specific decision problems
- Model estimation
- Time-domain models

Statistical decision problems

- Deciding whether or not to take the umbrella.
- Choosing a treatment or a diagnostic test for a patient.
- Determining the position of a moving object.
- Estimating the parameters of a model.
- Choosing whether to stop sampling.
- Deciding between alternative hypotheses.
- Planning in an unknown stochastic environment.

Belief + decision space + utility ⇒ decision

Decisions and outcomes

- Consider an experiment with possible outcomes $w \in \Omega$.

Decisions and outcomes

- Consider an experiment with possible outcomes $w \in \Omega$.
- We must select a decision $d \in D$ *before* knowing the outcome of the experiment.

Decisions and outcomes

- Consider an experiment with possible outcomes $w \in \Omega$.
- We must select a decision $d \in D$ *before* knowing the outcome of the experiment.
- We will obtain a reward $r \in R$ which depends on both w and d.

Decisions and outcomes

- Consider an experiment with possible outcomes $w \in \Omega$.
- We must select a decision $d \in D$ *before* knowing the outcome of the experiment.
- We will obtain a reward $r \in R$ which depends on both w and d.
- Our utility function U, assigns values to rewards.

Decisions and outcomes

- Consider an experiment with possible outcomes $w \in \Omega$.
- We must select a decision $d \in D$ *before* knowing the outcome of the experiment.
- We will obtain a reward $r \in R$ which depends on both w and d.
- Our utility function U, assigns values to rewards.
- How should we choose d?

Decisions and outcomes

- Consider an experiment with possible outcomes $w \in \Omega$.
- We must select a decision $d \in D$ *before* knowing the outcome of the experiment.
- We will obtain a reward $r \in R$ which depends on both w and d.
- Our utility function U, assigns values to rewards.
- How should we choose d?

Example (Taking the umbrella)

We must decide whether or not to take an umbrella to work. Our reward is a combination of whether we get wet and the amount of objects that we carry. We would rather not get wet and not carry too many things. The only events of interest are whether it rains or not.

Decisions and outcomes

- Consider an experiment with possible outcomes $w \in \Omega$.
- We must select a decision $d \in D$ *before* knowing the outcome of the experiment.
- We will obtain a reward $r \in R$ which depends on both w and d.
- Our utility function U, assigns values to rewards.
- How should we choose d?

Example (Taking the umbrella)

We must decide whether or not to take an umbrella to work. Our reward is a combination of whether we get wet and the amount of objects that we carry. We would rather not get wet and not carry too many things. The only events of interest are whether it rains or not.

r	Rain	No Rain
Umbrella	Burdened, Dry	Burdened, Dry
No Umbrella	Wet	Dry.

Rewards and utility

Definition (Reward function)

There exists a function $\sigma : \Omega \times D \to R$, such that if we select $d \in D$ and the experimental outcome is $w \in \Omega$, we obtain reward

$$r = \sigma(w, d) \tag{1.1}$$

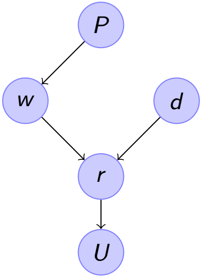

Figure: The dependency structure of the decision problem.

Rewards and utility

Definition (Reward function)

There exists a function $\sigma : \Omega \times D \to R$, such that if we select $d \in D$ and the experimental outcome is $w \in \Omega$, we obtain reward

$$r = \sigma(w, d) \tag{1.1}$$

Definition (Utility function)

The utility function $U : R \to \mathbb{R}$ gives a value to each reward, such that: We prefer r_1 to r_2, if and only if $U(r_1) > U(r_2)$.

Rewards and utility

Definition (Reward function)

There exists a function $\sigma : \Omega \times D \to R$, such that if we select $d \in D$ and the experimental outcome is $w \in \Omega$, we obtain reward

$$r = \sigma(w, d) \tag{1.1}$$

Definition (Utility function)

The utility function $U : R \to \mathbb{R}$ gives a value to each reward, such that: We prefer r_1 to r_2, if and only if $U(r_1) > U(r_2)$.

Example

When R is the space of monetary rewards, frequently $U(r) = \log(r)$.

Rewards and utility

Definition (Reward function)

There exists a function $\sigma : \Omega \times D \to R$, such that if we select $d \in D$ and the experimental outcome is $w \in \Omega$, we obtain reward

$$r = \sigma(w, d) \tag{1.1}$$

Definition (Utility function)

The utility function $U : R \to \mathbb{R}$ gives a value to each reward, such that: We prefer r_1 to r_2, if and only if $U(r_1) > U(r_2)$.

Example (Umbrella example continued)

U	Rain	No Rain
Umbrella	0	0
No Umbrella	-9	1.

Rewards and utility

Definition (Reward function)

There exists a function $\sigma : \Omega \times D \to R$, such that if we select $d \in D$ and the experimental outcome is $w \in \Omega$, we obtain reward

$$r = \sigma(w, d) \tag{1.1}$$

Definition (Utility function)

The utility function $U : R \to \mathbb{R}$ gives a value to each reward, such that: We prefer r_1 to r_2, if and only if $U(r_1) > U(r_2)$.

Expected utility

Given the above definitions, the optimal decision maximises:

$$\mathbb{E}(U \mid d) = \int_\Omega U[\sigma(w, d)] \, \mathrm{d}P(w) \tag{1.2}$$

The rewards and utilities are part of the problem definition

- The reward set specifies which outcomes we care to differentiate.
- The utility function elucidates our preference for different outcomes.

Together, they define our goal.

Loss and risk

- The unknown outcome of the experiment W is called a parameter

Loss and risk

- The unknown outcome of the experiment W is called a parameter
- The set of outcomes Ω is called the parameter space.

Loss and risk

- The unknown outcome of the experiment W is called a parameter
- The set of outcomes Ω is called the parameter space.

Definition (Loss)

$$L(w, d) = -U[\sigma(w, d)]. \tag{1.3}$$

Loss and risk

- The unknown outcome of the experiment W is called a parameter
- The set of outcomes Ω is called the parameter space.

Definition (Loss)

$$L(w, d) = -U[\sigma(w, d)]. \tag{1.3}$$

Definition (Risk)

$$\rho(P, d) = \int_\Omega L(w, d) \, \mathrm{d}P(w). \tag{1.4}$$

Loss and risk

- The unknown outcome of the experiment W is called a parameter
- The set of outcomes Ω is called the parameter space.

Definition (Loss)

$$L(w, d) = -U[\sigma(w, d)]. \tag{1.3}$$

Definition (Risk)

$$\rho(P, d) = \int_\Omega L(w, d) \,\mathrm{d}P(w). \tag{1.4}$$

Definition (Bayes risk)

$$\rho^*(P) = \inf_{d \in D} \rho(P, d) \tag{1.5}$$

Example

Let $\Omega = \{0, 1\}$ and $D = [0, 1]$. We define the loss $L : \Omega \times D \to \mathbb{R}$ as

$$L(w, d) = |w - d|^{\alpha}, \tag{1.6}$$

$\alpha \geq 1$. The distribution of outcomes is

$$\mathbb{P}(W = 0) = u \qquad \mathbb{P}(W = 1) = 1 - u. \tag{1.7}$$

For $\alpha = 1$:

$$\rho(P, d) = L(0, d)u + L(1, d)(1 - u) = du + (1 - d)(1 - u), \tag{1.8}$$

so, if $u > 1/2$, then the risk is minimised by setting $d = 0$.

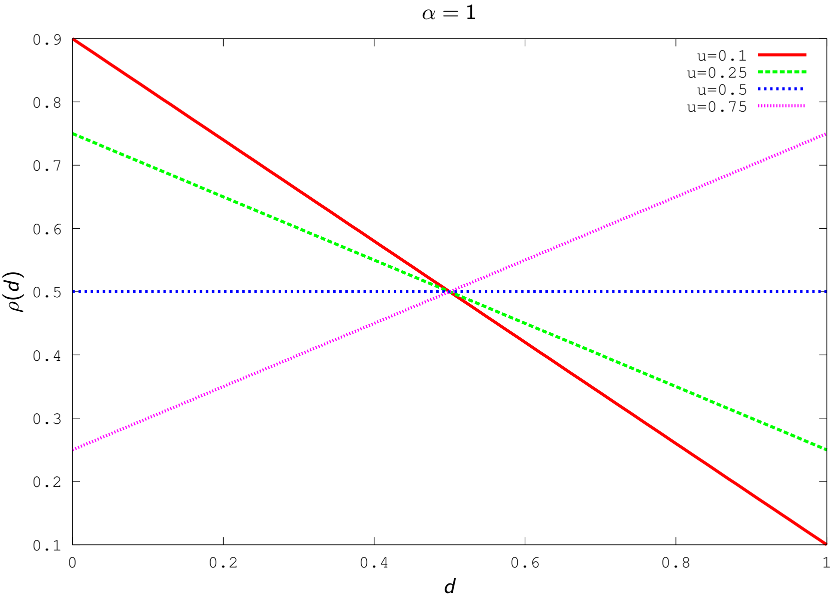

Figure: Risk for four different distributions with absolute loss

Example

Let $\Omega = \{0, 1\}$ and $D = [0, 1]$. We define the loss $L : \Omega \times D \to \mathbb{R}$ as

$$L(w, d) = |w - d|^{\alpha}, \tag{1.6}$$

$\alpha \geq 1$. The distribution of outcomes is

$$\mathbb{P}(W = 0) = u \qquad \mathbb{P}(W = 1) = 1 - u. \tag{1.7}$$

For $\alpha > 1$:

$$\rho(P, d) = d^{\alpha} u + (1 - d)^{\alpha}(1 - u), \tag{1.8}$$

and by differentiating:

$$d^* = \left[1 + \left(\frac{1}{1/u - 1}\right)^{\frac{1}{\alpha - 1}}\right]^{-1}.$$

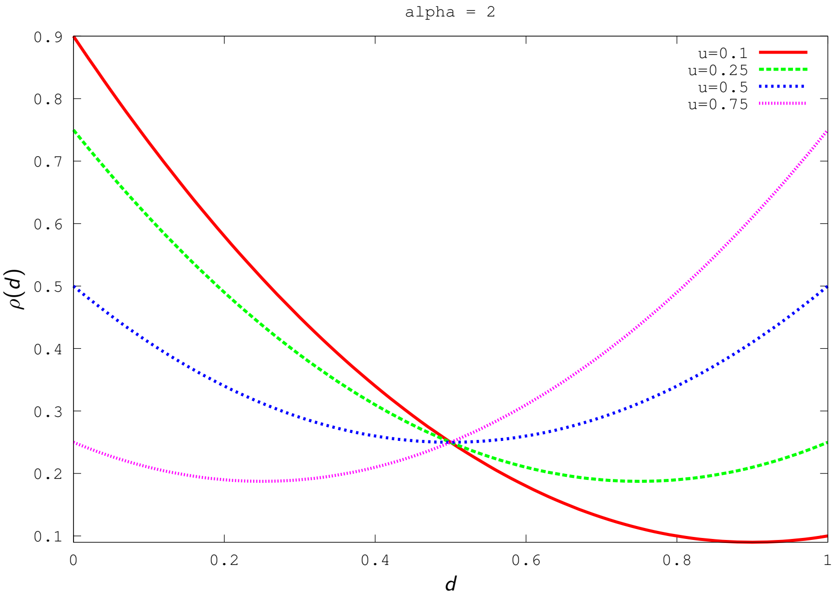

Figure: Risk for four different distributions with quadratic loss.

Estimation under quadratic loss

Now consider $w \in \mathbb{R}$ with $w \sim P$ and $d \in \mathbb{R}$. We define the loss as

$$L(w, d) = |w - d|^2. \tag{1.8}$$

The optimal decision minimises

$$\mathbb{E}(L \mid d) = \int_{\mathbb{R}} |w - d|^2 \, \mathrm{d}P(w)$$

Then, as long as $\frac{\partial}{\partial d}|w - d|^2$ is measurable

$$\begin{aligned}
\frac{\partial}{\partial d} \int_{\mathbb{R}} |w - d|^2 \, \mathrm{d}P(w) &= \int_{\mathbb{R}} \frac{\partial}{\partial d} |w - d|^2 \, \mathrm{d}P(w) && (1.9)\\
&= 2 \int_{\mathbb{R}} (w - d) \, \mathrm{d}P(w) && (1.10)\\
&= 2 \int_{\mathbb{R}} w \, \mathrm{d}P(w) - 2 \int_{\mathbb{R}} d \, \mathrm{d}P(w) && (1.11)\\
&= 2 \, \mathbb{E} \, w - 2d, && (1.12)
\end{aligned}$$

so the cost is minimised for $d = \mathbb{E} \, w$.

A mixture of distributions

Consider two probability measures P, Q on Ω, such that $P(A)$ and $Q(A)$ is the probability of A under each distribution.

A mixture of distributions

Consider two probability measures P, Q on Ω, such that $P(A)$ and $Q(A)$ is the probability of A under each distribution.

Definition

For any P, Q and $\alpha \in [0, 1]$,

$$\alpha P + (1 - \alpha) Q$$

denotes the probability measure such that

$$\alpha P(A) + (1 - \alpha) Q(A)$$

for any $A \subset \Omega$.

Concavity of the Bayes risk

Theorem

For probability measures P, Q on Ω and any $\alpha \in [0,1]$

$$\rho^*[\alpha P + (1-\alpha)Q] \geq \alpha \rho^*(P) + (1-\alpha)\rho^*(Q). \tag{1.13}$$

Proof.

From the definition of risk (1.4), for any decision $d \in D$:

$$\rho[\alpha P + (1-\alpha)Q, d] = \alpha \rho(P, d) + (1-\alpha)\rho(Q, d)$$

And so from the Bayes risk (1.5)

$$\begin{aligned}\rho^*[\alpha P + (1-\alpha)Q] &= \inf_{d \in D} \rho[\alpha P + (1-\alpha)Q, d] \\ &= \inf_{d \in D} [\alpha \rho(P, d) + (1-\alpha)\rho(Q, d)].\end{aligned}$$

□

Concavity of the Bayes risk

Theorem

For probability measures P, Q *on* Ω *and any* $\alpha \in [0, 1]$

$$\rho^*[\alpha P + (1-\alpha)Q] \geq \alpha\rho^*(P) + (1-\alpha)\rho^*(Q). \tag{1.13}$$

Proof.

$$\rho^*[\alpha P + (1-\alpha)Q] = \inf_{d \in D}[\alpha\rho(P, d) + (1-\alpha)\rho(Q, d)].$$

□

Concavity of the Bayes risk

Theorem

For probability measures P, Q on Ω and any $\alpha \in [0,1]$

$$\rho^*[\alpha P + (1-\alpha)Q] \geq \alpha \rho^*(P) + (1-\alpha)\rho^*(Q). \tag{1.13}$$

Proof.

$$\rho^*[\alpha P + (1-\alpha)Q] = \inf_{d \in D} [\alpha \rho(P,d) + (1-\alpha)\rho(Q,d)].$$

Since $\inf_x [f(x) + g(x)] \geq \inf_x f(x) + \inf_x g(x)$,

$$\begin{aligned} \rho^*[\alpha P + (1-\alpha)Q] &\geq \alpha \inf_{d \in D} \rho(P,d) + (1-\alpha) \inf_{d \in D} \rho(Q,d) \\ &= \alpha \rho^*(P) + (1-\alpha)\rho^*(Q) \end{aligned}$$

□

The risk function for quadratic loss

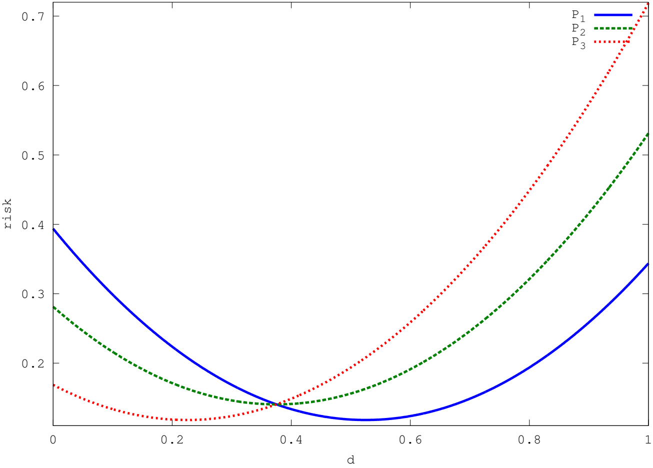

Figure: Fixed distribution, varying decision. The decision risk under three different distributions.

Concavity of the Bayes risk

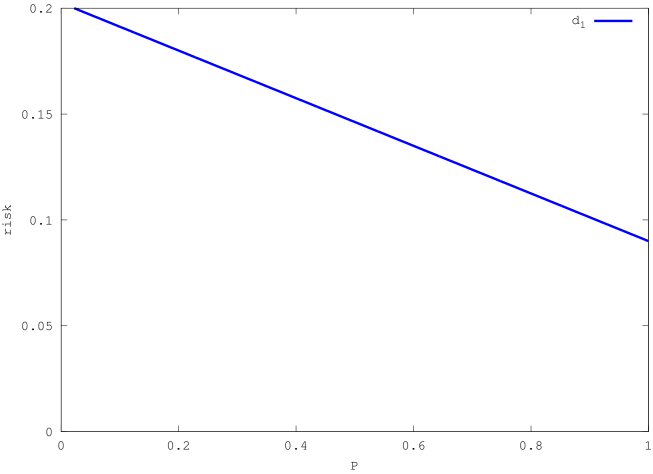

Figure: Fixed decision, varying distribution. The risk of a fixed decision is a linear function of P

Concavity of the Bayes risk

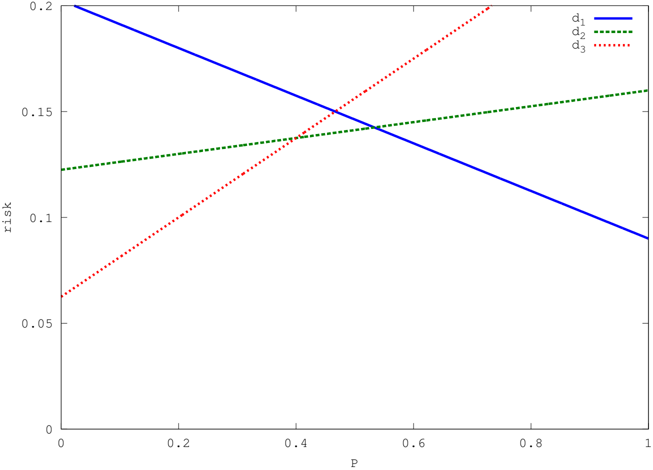

Figure: The risk of a few decisions as P varies. Each decision corresponds to one of these lines.

Concavity of the Bayes risk

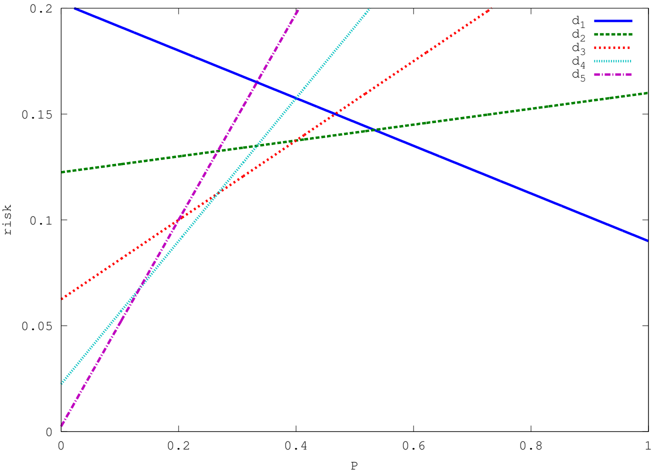

Figure: For each *P*, there is at least one decision minimising the risk.

Concavity of the Bayes risk

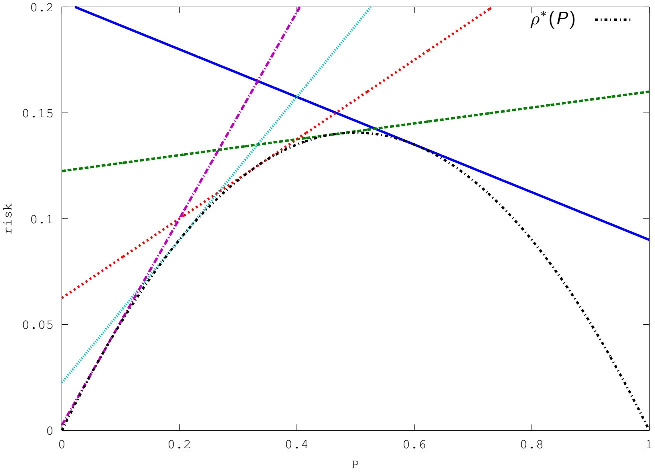

Figure: The Bayes risk is concave and the minimising decision is tangent to it.

Concavity of the Bayes risk

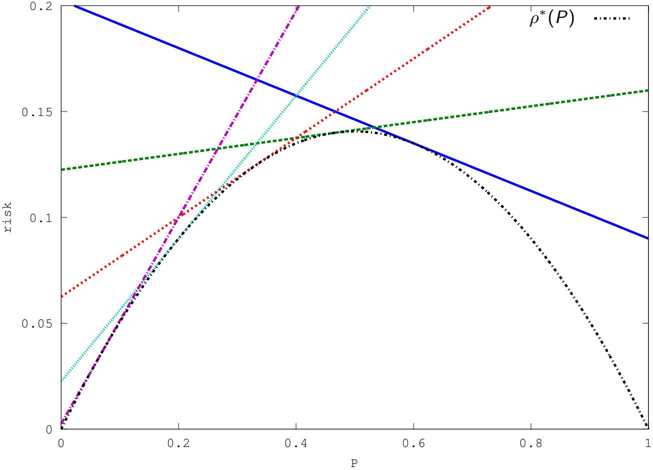

Figure: If we are not very wrong about P, then we are not far from optimal.

Concavity of the Bayes risk

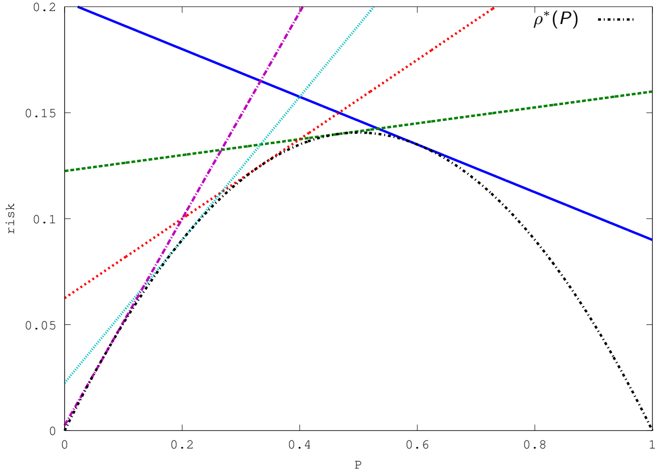

Figure: We can approximate the Bayes risk by taking the minimum of a finite number of decisions.

1 Statistical decision problems
- Rewards that depend on the outcome of an experiment
- Concavity of the Bayes risk

2 Specific decision problems
- Model estimation
- Time-domain models

Statistical decision problems

- So far, we considered some fixed P.
- Now, we wish to construct a P on the basis of prior beliefs and evidence.
- This leads to Bayesian inference.

Model estimation

- Set of models $\mathcal{M} = \{P_\theta : \theta \in \Theta\}$, where Θ is the parameter space.
- Data $x \sim P_{\theta^*}$, where θ^* is the true parameter.
- Our prior belief is a probability measure ξ on Θ. Thus, $\xi(B)$ is our belief that $\theta^* \in B \subset \Theta$.

Model estimation

- Set of models $\mathcal{M} = \{P_\theta : \theta \in \Theta\}$, where Θ is the parameter space.
- Data $x \sim P_{\theta^*}$, where θ^* is the true parameter.
- Our prior belief is a probability measure ξ on Θ. Thus, $\xi(B)$ is our belief that $\theta^* \in B \subset \Theta$.

$$\xi(B \mid x) = \frac{\xi(B, x)}{\xi(x)} = \frac{\int_B P_\theta(x) \, \mathrm{d}\xi(\theta)}{\int_\Theta P_\theta(x) \, \mathrm{d}\xi(\theta)}, \qquad \forall B \subset \Theta \qquad \text{(posterior)}$$

$$\xi(\theta \mid x) = \frac{\xi(\theta, x)}{\xi(x)} = \frac{P_\theta(x)\xi(\theta)}{\sum_{\theta' \in \Theta} P_{\theta'}(x)\xi(\theta')}, \qquad \forall \theta \in \Theta. \qquad \text{(finite case)}$$

Model estimation

- Set of models $\mathcal{M} = \{P_\theta : \theta \in \Theta\}$, where Θ is the parameter space.
- Data $x \sim P_{\theta^*}$, where θ^* is the true parameter.
- Our prior belief is a probability measure ξ on Θ. Thus, $\xi(B)$ is our belief that $\theta^* \in B \subset \Theta$.

$$\xi(B \mid x) = \frac{\xi(B, x)}{\xi(x)} = \frac{\int_B P_\theta(x) \,\mathrm{d}\xi(\theta)}{\int_\Theta P_\theta(x) \,\mathrm{d}\xi(\theta)}, \qquad \forall B \subset \Theta \qquad \text{(posterior)}$$

$$\xi(\theta \mid x) = \frac{\xi(\theta, x)}{\xi(x)} = \frac{P_\theta(x)\xi(\theta)}{\sum_{\theta' \in \Theta} P_{\theta'}(x)\xi(\theta')}, \qquad \forall \theta \in \Theta. \qquad \text{(finite case)}$$

Ideally, we'd like to communicate the *complete posterior* distribution. If this is infeasible we may choose a *single parameter*, or a *credible set of parameters*

Model estimation

- Set of models $\mathcal{M} = \{P_\theta : \theta \in \Theta\}$, where Θ is the parameter space.
- Data $x \sim P_{\theta^*}$, where θ^* is the true parameter.
- Our prior belief is a probability measure ξ on Θ. Thus, $\xi(B)$ is our belief that $\theta^* \in B \subset \Theta$.

$$\xi(B \mid x) = \frac{\xi(B, x)}{\xi(x)} = \frac{\int_B P_\theta(x) \, \mathrm{d}\xi(\theta)}{\int_\Theta P_\theta(x) \, \mathrm{d}\xi(\theta)}, \qquad \forall B \subset \Theta \qquad \text{(posterior)}$$

$$\xi(\theta \mid x) = \frac{\xi(\theta, x)}{\xi(x)} = \frac{P_\theta(x)\xi(\theta)}{\sum_{\theta' \in \Theta} P_{\theta'}(x)\xi(\theta')}, \qquad \forall \theta \in \Theta. \qquad \text{(finite case)}$$

Minimising the Bayes risk

Given a loss function $L : \Theta \times D \to \mathbb{R}$, such that $L(\theta^*, d)$ is the loss of choosing d when the true parameter is θ^*, the optimal choice minimises

$$\mathbb{E}_\xi(L \mid d, x) = \int_\Theta L(\theta, d) \, \mathrm{d}\xi(\theta \mid x). \tag{2.1}$$

Important points

- Model estimation can be formulated as (formally) simple statistical induction.
- Given a prior belief and evidence, we obtain a posterior belief, representing our uncertainty, in a well-understood mathematical framework.
- Ideally, we would like to communicate the complete posterior.
- Alternatively, we can formulate a decision problem whereby we select the decision minimising the Bayes risk. However this is not always easier!
- Nothing changes in the decision problem, other than replacing P with $\xi(\theta \mid x)$.
- The decision and parameter spaces may not be identical. A decision could for example be that $\theta < 0$.

Integrating multiple models

- Consider the problem of sequential prediction.
- Given a sequence $x^t = x_1, \ldots, x_t$, with $x_i \in \mathcal{X}$, predict x_{t+1}.
- This can be formulated as a decision problem where $D = \mathcal{X}$ and the loss function is

$$L(d, x) = \begin{cases} 0, & d = x, \\ 1, & d \neq x. \end{cases}$$

- If d_t is our decision at time t, define $\ell_t = Loss(d_t, x_t)$.
- Assume we have N models out our disposal. We distinguish two cases.
  1. The statistical model case.
  2. The expert case.

Integrating multiple models

Statistical models

- The i-ith model, after having seen x^t, outputs a complete distribution $p_{i,t}$ on $\mathcal{X}$.
- If we have a prior distribution ξ on the experts, we can write our posterior as

$$\xi(i \mid x^t) = \frac{p_{i,t-1}(x_t)\xi(i \mid x^{t-1})}{\sum_{j=1}^{N} p_{j,t-1}(x_t)\xi(j \mid x^{t-1})}$$

Our decision for x_{t+1} should then minimise

$$\sum_x L(d, x) \sum_i \xi(i \mid x^t) p_{i,t}(x)$$

Integrating multiple models

Experts

- The i-ith expert, after having seen x^t, outputs a prediction $f_{i,t} \in \mathcal{X}$.
- We cannot use the previous approach, as we have no way of assigning probabilities to predictions.
- We also don't know under what loss function the experts make their predictions!
- However, using the exponentially weighted average forecaster

$$\xi(i \mid x^t) = \frac{e^{-\eta \ell_{i,t}} \xi(i \mid x^{t-1})}{\sum_j e^{-\eta \ell_{j,t}} \xi(j \mid x^{t-1})},$$

where $\ell_{i,t} = 1$ if $f_{i,t} = x_t$ and 0 otherwise, we can show that the difference between our loss at time t and the loss of the best expert, is bounded by

$$\sqrt{\frac{\ln N}{t}}$$

Discrete-time Markov process (Markov chain)

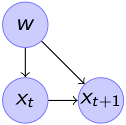

Figure: Markov model

- An alphabet $\mathcal{X}$.
- Parameters w.
- A sequence of observed variables $x^t = x_1, \ldots, x_t$, with $x_i \in \mathcal{X}$:

$$x_{t+1} \perp x^{t-1} \mid x_t = x \sim P_{g(x)}.$$

Discrete case

$\mathcal{X} = \{1, \ldots, N\}$.

$$g(x) = w_x, \qquad w_x \in \mathbb{R}^N, \|w_x\|_1 = 1$$

P_w is the multinomial distribution with parameter w. The posterior distribution of parameters

$$\xi(w \mid x^t)$$

can be estimated efficiently in closed-form if the prior $\xi(w)$ is a Dirichlet product.

Latent discrete-time Markov process

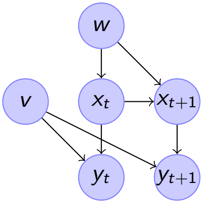

Figure: Latent Markov model

- An alphabet $\mathcal{X}$ and an alphabet $\mathcal{Y}$
- Parameters w, v.
- A sequence of hidden variables $x^t = x_1, \ldots, x_t$, with $x_i \in \mathcal{X}$.

$$x_{t+1} \perp x^{t-1} \mid x_t = x \sim P_{h(x)}.$$

- A sequence of observed variables $y^t = y_1, \ldots, y_t$, with $y_i \in \mathcal{Y}$:

$$y_t \perp y^{t-1}, x^{t-1} \mid x_t = x \sim Q_{g(x)}.$$

- Discrete case: P_i, Q_i are multinomial distributions ⇒ Hidden Markov model.
- Real case: If P_i, Q_i are Gaussian distributions ⇒ Kalman filter.

Latent discrete-time Markov process

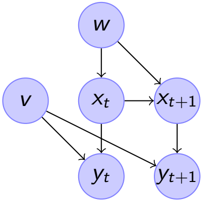

Figure: Latent Markov model

- An alphabet $\mathcal{X}$ and an alphabet $\mathcal{Y}$
- Parameters w, v.
- A sequence of hidden variables $x^t = x_1, \ldots, x_t$, with $x_i \in \mathcal{X}$.

$$x_{t+1} \perp x^{t-1} \mid x_t = x \sim P_{h(x)}.$$

- A sequence of observed variables $y^t = y_1, \ldots, y_t$, with $y_i \in \mathcal{Y}$:

$$y_t \perp y^{t-1}, x^{t-1} \mid x_t = x \sim Q_{g(x)}.$$

- Calculating the posterior $\xi(w, v \mid y^t)$ is not closed-form.
- However $\xi(w, v \mid x^t, y^t)$ can be, leading to efficient (approximate) estimators.
- If we only want to choose a parameter (w, v), various optimisation algorithms can be used to (approximately) minimise the loss.

Summary

- Model estimation is essentially calculation of the parameter distribution.
- If we want to make decisions according to the expected utility principle, we require probabilities of events.
- However, sometimes we can prove that this is not required to obtain good performance.
- In latent processes, we have two problems
  1. Estimation of the model parameters
  2. Estimation of the other latent variables
- Joint estimation is intractable (but interesting special cases exist).
- However, estimating one given the other is tractable for many models of interest.
- You shall see more about how we can make use of that in the next talks.