Decision making under uncertainty

Michael U. Gutmann

Probabilistic Machine Learning (INFR11298)
School of Informatics, The University of Edinburgh

Autumn Semester 2026

Recap

- We modelled actions as interventions in the data generating process.
- For DAGs, intervening on a node removes all incoming arrows into the node.
- We discussed how to compute/predict the effect of interventions.
- However, we have not yet discussed which action to choose in the face of uncertainty.
- This is the topic of decision theory, discussed here.

Program

1. Brief introduction to decision theory
2. Its use in statistics and machine learning

Program

1. Brief introduction to decision theory
   - Loss and decision principles
   - Connection to causality
   - Mild cognitive impairment example
2. Its use in statistics and machine learning

Loss

- We frame decision making under uncertainty as a game. The rules are known in advance: the possible actions $\mathbf{a}$, the possible values of a hidden quantity $\mathbf{h}$, and the loss $\ell(\mathbf{h}, \mathbf{a})$.
- You choose an action $\mathbf{a}$ without knowing the value of $\mathbf{h}$. Then $\mathbf{h}$ is revealed and you incur the loss $\ell(\mathbf{h}, \mathbf{a})$.
- Cognitive impairment example:
  - Hidden/unobserved quantity: cognitive impairment
  - Action: possible life style changes and cognitive training
- Classification example:
  - Hidden/unobserved quantity: true class label
  - Action: estimate of the class label
- Parameter estimation example:
  - Hidden/unobserved quantity: true parameter value of a model
  - Action: estimate of the parameter

Decision principles

- Popular decision principles are the minimax and the expected loss (utility) principle.
- Minimax principle:
  - Choose the action that minimises the maximum loss $\max_{\mathbf{h}} \ell(\mathbf{h}, \mathbf{a})$.
  - Pro: Useful in case of adversaries. Does not depend on the distribution of $\mathbf{h}$.
  - Con: Often too pessimistic, resulting in overly conservative actions
- Expected loss principle:
  - Choose the action that minimises the expected loss $\mathbb{E}_{\mathbf{h}}[\ell(\mathbf{h}, \mathbf{a})]$, called the risk.
  - Pro: Broadly useful, combines impact of action (loss) with chance of occurrence (expectation).
  - Con: Action depends on the distribution of $\mathbf{h}$.
- We here focus on the expected loss principle.

Expected loss

- Risk $\mathbb{E}_{\mathbf{h}}[\ell(\mathbf{h}, \mathbf{a})]$ depends on the distribution of $\mathbf{h}$.
- Provides considerable flexibility:
  - It can be past data on $\mathbf{h}$ that we have collected.
  - It can be our personal/prior belief of what $\mathbf{h}$ may be.
  - It can be derived from some indirect evidence about $\mathbf{h}$ in the form of $\mathbf{x} \sim p(\mathbf{x}|\mathbf{h})$.
- Leads to different schools of decision making.
- In the last case, we compute the expectation with respect to the conditional

$$p(\mathbf{h}|\mathbf{x}) = \frac{p(\mathbf{x}|\mathbf{h})p(\mathbf{h})}{p(\mathbf{x})} \tag{1}$$

- Resulting expected loss

$$R(\mathbf{a}|\mathbf{x}) = \mathbb{E}_{p(\mathbf{h}|\mathbf{x})}[\ell(\mathbf{h}, \mathbf{a})] \tag{2}$$

is sometimes called the posterior expected loss / posterior risk.

From action to policy

- Denote by $f(\mathbf{h})$ the distribution of $\mathbf{h}$ that we average over and by $R(\mathbf{a}; f)$ the corresponding risk.
- The optimal action is

$$\mathbf{a}^*(f) = \operatorname*{argmin}_{\mathbf{a}} R(\mathbf{a}; f) \tag{3}$$

$$= \operatorname*{argmin}_{\mathbf{a}} \mathbb{E}_{f(\mathbf{h})} [\ell(\mathbf{h}, \mathbf{a})] \tag{4}$$

- In case of posterior risk, the optimal action depends on $\mathbf{x}$

$$\mathbf{a}^*(p(\mathbf{h}|\mathbf{x})) = \operatorname*{argmin}_{\mathbf{a}} R(\mathbf{a}; p(\mathbf{h}|\mathbf{x})) \tag{5}$$

$$= \operatorname*{argmin}_{\mathbf{a}} \mathbb{E}_{p(\mathbf{h}|\mathbf{x})} [\ell(\mathbf{h}, \mathbf{a})] \tag{6}$$

- We will overload notation and denote $\mathbf{a}^*(p(\mathbf{h}|\mathbf{x}))$ by $\mathbf{a}^*(\mathbf{x})$.
- Called a policy, mapping evidence $\mathbf{x}$ to actions.

Causal decision theory

- A more recent decision principle is framed in terms of causality (Pearl, Causality, Ch 4).
- Choose the action that minimises

$$\mathcal{R}(\mathbf{a}) = \mathbb{E}_{p(\mathbf{y};do(\mathbf{a}))}\left[\ell_o(\mathbf{y})\right] \tag{7}$$

  where $\ell_o(\mathbf{y})$ is the loss for *outcome* $\mathbf{y}$.
- Note the use of the postinterventional distribution $p(\mathbf{y}; do(\mathbf{a}))$.
- It may also be conditioned on evidence $\mathbf{x}$, i.e. $p(\mathbf{y}|\mathbf{x}; do(\mathbf{a}))$.
- We may include an action-dependency in the loss, so that we have $\ell_o(\mathbf{y}, \mathbf{a})$
- In some cases, we can relate $\mathcal{R}(\mathbf{a})$ to the expected loss.

Rewriting it in terms of expected loss (optional, not examinable)

- If outcome $\mathbf{y}$ depends on unobserved variables $\mathbf{h}$, we have

$$p(\mathbf{y}; do(\mathbf{a})) = \int p(\mathbf{h}; do(\mathbf{a}))p(\mathbf{y}|\mathbf{h}; do(\mathbf{a}))\mathrm{d}\mathbf{h} \tag{8}$$

$$= \mathbb{E}_{p(\mathbf{h}; do(\mathbf{a}))}\left[p(\mathbf{y}|\mathbf{h}; do(\mathbf{a}))\right] \tag{9}$$

- Assume that the intervention happens after $\mathbf{h}$ is generated ($\mathbf{h}$ is not a descendant of $\mathbf{a}$). Then $p(\mathbf{h}; do(\mathbf{a})) = p(\mathbf{h})(*)$.
- Under this assumption

$$\mathcal{R}(\mathbf{a}) = \mathbb{E}_{p(\mathbf{y},\mathbf{h}; do(\mathbf{a}))}\left[\ell_o(\mathbf{y}, \mathbf{a})\right] \tag{10}$$

$$= \mathbb{E}_{p(\mathbf{h}; do(\mathbf{a}))}\mathbb{E}_{p(\mathbf{y}|\mathbf{h}; do(\mathbf{a}))}\left[\ell_o(\mathbf{y}, \mathbf{a})\right] \tag{11}$$

$$\overset{(*)}{=} \mathbb{E}_{p(\mathbf{h})}\mathbb{E}_{p(\mathbf{y}|\mathbf{h}; do(\mathbf{a}))}\left[\ell_o(\mathbf{y}, \mathbf{a})\right] \tag{12}$$

- This has the form of the expected loss with

$$\ell(\mathbf{h}, \mathbf{a}) = \mathbb{E}_{p(\mathbf{y}|\mathbf{h}; do(\mathbf{a}))}\left[\ell_o(\mathbf{y}, \mathbf{a})\right] \tag{13}$$

Mild cognitive impairment (MCI) example

- 70 year old man: prior MCI probability 5/45, posterior MCI probability 2/3 after the online test.
- Loss function $L(h, a)$ with $h \in \{0, 1\}$ indicating whether he has MCI or not and $a \in \{0, 1\}$ whether he takes action or not.
- Assume taking action means life-style changes (less alcohol, more exercise) and attending cognitive training sessions for four months.
- Loss is measured in terms of "quality-adjusted life years" (QALY).
- 1 QALY means one year in perfect health. 0.5 QALY can mean one year which is only half enjoyable, e.g. due to pain.

MCI example: loss function

- Loss function as a table

Action/MCI	$h = 1$	$h = 0$
$a = 1$	c+(1-e)H	c
$a = 0$	H	0

- H is the harm of having MCI unmanaged over the decision horizon (1 year). Assume its 14 low quality days $\rightarrow H = 14/365 = 0.038$ QALY.
- e is the reduction of the harm. Assume $e = 0.25$.
- c is the loss in quality time due to taking the action (e.g. travel to training sessions, attending them instead of doing something fun, social stigma etc). Assume 0.002 QALY.

MCI example: optimal action

- Loss function

Action/MCI	$h = 1$	$h = 0$
$a = 1$	0.0305	0.002
$a = 0$	0.038	0

- Prior risk (units of QALY):

$$R(a = 1; p(h)) = 5/45 \cdot 0.0305 + 40/45 \cdot 0.002 = 0.0052$$
$$R(a = 0; p(h)) = 5/45 \cdot 0.038 + 40/45 \cdot 0 = 0.0042$$

No action $a = 0$ has lower risk and is thus better.

- Posterior risk (units of QALY):

$$R(a = 1; p(h|x)) = 2/3 \cdot 0.0305 + 1/3 \cdot 0.002 = 0.0210$$
$$R(a = 0; p(h|x)) = 2/3 \cdot 0.038 + 1/3 \cdot 0 = 0.0253$$

Taking action $a = 1$ has lower risk and is thus better.

Program

1. Brief introduction to decision theory
   - Loss and decision principles
   - Connection to causality
   - Mild cognitive impairment example
2. Its use in statistics and machine learning

Program

1. Brief introduction to decision theory

2. Its use in statistics and machine learning
   - Common loss functions
   - Learning as a decision under the 0-1 loss
   - Learning as a decision under the log-loss

Common loss functions

- Loss functions $\ell(\mathbf{h}, \mathbf{a})$ are best tailored to the problem at hand.
- There are, however, a number of "standard" loss functions that are widely used in statistics and machine learning.
- They include
  - Quadratic loss
  - Absolute error loss
  - Zero-one loss
  - Log-loss
- For each loss, we next derive the optimal action/policy.
- We denote the distribution of $\mathbf{h}$ by f.

Quadratic loss

- For simplicity, assume h and a are one-dimensional.
- The quadratic loss is

$$\ell(h, a) = (h - a)^2 \tag{14}$$

- Optimal action

$$a^*(f) = \operatorname*{argmin}_a \mathbb{E}_{f(h)} \left[(h - a)^2 \right] \tag{15}$$

- May be continuous-valued even if h is discrete.

Quadratic loss

- To solve the optimisation problem, let $m = \mathbb{E}_f[h]$ be the expected value of h under f. The expected loss is

$$\begin{aligned}
\mathbb{E}_{f(h)}\left[(h-a)^2\right] &= \mathbb{E}_{f(h)}\left[(h-m+m-a)^2\right] && (16)\\
&= \mathbb{E}_{f(h)}\Big[(h-m)^2 + 2(h-m)(m-a) \\
&\quad + (m-a)^2\Big] && (17)\\
&= \mathbb{E}_{f(h)}\left[(h-m)^2\right] + 0 + (m-a)^2 && (18)\\
&= \mathbb{V}(h) + (m-a)^2 && (19)
\end{aligned}$$

- We thus have

$$a^*(f) = \operatorname*{argmin}_a (m-a)^2 = m$$

- The optimal action is the expected value of h wrt f: $\mathbb{E}_f[h]$.
- Generalises to multidimensional case: $\mathbf{a}^*(f) = \mathbb{E}_f[\mathbf{h}]$.

Absolute error loss

- We assume that h and a are one-dimensional.
- We here consider a loss that increases linearly as a deviates from h, rather than quadratically: $\ell(h, a) = |h - a|$.
- Optimal action

$$a^*(f) = \operatorname*{argmin}_a \mathbb{E}_{f(h)}[|h - a|] \tag{20}$$

- We can show that $a^*(f)$ is the median of $f(h)$, i.e. any point at which probability mass is equally split between the left and the right.

  More formally, let $F(\alpha) = \mathbb{P}(h \leq \alpha)$ be the cumulative distribution function of f. The median is any number m that satisfies $F(m^-) \leq 0.5 \leq F(m)$ where m^- is the lower limit of F at m. Not unique in case of discrete random variables.

Absolute error loss (proof, not examinable)

- We consider the more general case

$$\ell(h, a) = \begin{cases} k_1(h-a) & \text{if } h \geq a \\ k_2(a-h) & \text{if } h < a \end{cases} \quad \text{with } k_1, k_2 > 0 \qquad (21)$$

  This loss is called the quantile or pinball loss.

- The expected loss is (assuming pdfs)

$$R(a; f) = \int_a^\infty k_1(h-a) f(h) dh + \int_{-\infty}^a k_2(a-h) f(h) dh$$

- Taking derivatives gives (using Leibniz rule)

$$\begin{aligned} \frac{d}{da} R(a; f) &= -k_1 \int_a^\infty f(h) dh + k_2 \int_{-\infty}^a f(h) dh \\ &= -k_1 (1 - \int_{-\infty}^a f(h) dh) + k_2 \int_{-\infty}^a f(h) dh \end{aligned}$$

Absolute error loss (proof, not examinable)

- *continued from previous slide:*

$$\frac{d}{da}R(a;f) = -k_1 + (k_1+k_2)\int_{-\infty}^{a} f(h)dh$$
$$= -k_1 + (k_1+k_2)\mathbb{P}(h \leq a)$$

- Setting the derivative to zero gives the condition

$$\mathbb{P}(h \leq a^*) = \frac{k_1}{k_1+k_2} \tag{22}$$

- This means that $a^*(f)$ is the $k_1/(k_1+k_2)$-th quantile of $f(h)$.
- The second derivative $(k_1+k_2)f(a)$ is non-negative, so that $R(a;f)$ is convex and a^* is a minimum. If $f(a^*) > 0$, the quantile and hence the minimiser is unique. If $f(a) = 0$ in a neighbourhood of a^*, $\mathbb{P}(h \leq a)$ is flat there: the quantile, and hence the minimiser, is not unique (as for the median).
- For $k_1 = k_2$, we obtain the 0.5 quantile, i.e. the median.

Zero-one loss

- The unobserved variable $\mathbf{h}$ and $\mathbf{a}$ may be multivariate.
- For discrete $\mathbf{h}$ and $\mathbf{a}$, the zero-one loss is

$$\ell(\mathbf{h}, \mathbf{a}) = \mathbb{1}(\mathbf{h} \neq \mathbf{a}) = 1 - \mathbb{1}(\mathbf{h} = \mathbf{a}) = \begin{cases} 1 & \text{if } \mathbf{h} \neq \mathbf{a} \\ 0 & \text{if } \mathbf{h} = \mathbf{a} \end{cases} \tag{23}$$

- For continuous $\mathbf{h}$ and $\mathbf{a}$, it is defined as $\ell(\mathbf{h}, \mathbf{a}) = 1 - \delta(\mathbf{h} - \mathbf{a})$ where $\delta(.)$ is the Dirac-delta distribution (picture as Gaussian with vanishingly small variance)
- Optimal action

$$\mathbf{a}^*(f) = \underset{\mathbf{a}}{\operatorname{argmin}} \, \mathbb{E}_{f(\mathbf{h})} \left[\ell(\mathbf{h}, \mathbf{a}) \right] \tag{24}$$

Zero-one loss

- To solve the optimisation problem, note that

$$\mathbb{E}_{f(\mathbf{h})}\left[\ell(\mathbf{h}, \mathbf{a})\right] = 1 - f(\mathbf{h} = \mathbf{a}) \tag{25}$$

  This holds for continuous or discrete quantities.

- Since $\operatorname{argmin}_{\mathbf{a}} 1 - f(\mathbf{h} = \mathbf{a}) = \operatorname{argmax}_{\mathbf{h}} f(\mathbf{h})$, the optimal action is to choose the mode of f:

$$\mathbf{a}^*(f) = \operatorname*{argmax}_{\mathbf{h}} f(\mathbf{h}) \tag{26}$$

$$= \operatorname*{argmax}_{\mathbf{h}} \log f(\mathbf{h}) \tag{27}$$

- The mode corresponds to the most probable states, whose inference we discussed for factor trees: it is the optimal action under the 0-1 loss.

Log-loss

- So far, the action $\mathbf{a}$ was a scalar or vector. Here, the action is a distribution $q(\mathbf{h})$ over $\mathbf{h}$.
- The game is the same as before. The rules are known in advance: you report a distribution q, and the loss is $\ell(\mathbf{h}, q) = \log(1/q(\mathbf{h})) = -\log q(\mathbf{h})$.
- You choose q without knowing the value of $\mathbf{h}$. Then $\mathbf{h}$ is revealed and you incur the loss $-\log q(\mathbf{h})$.
- If q assigns a small probability to the $\mathbf{h}$ that occurs, you incur a large penalty.
- The log loss $\ell(\mathbf{h}, q) = \log(1/q(\mathbf{h}))$ is called the "surprise" in information theory. For pmfs, the surprise is non-negative and 0 for the certain event, i.e. if $q(\mathbf{h}) = 1$.
- First used for pmfs, but later also for pdfs (where it can be negative).

Log-loss

- The log-loss is said to be local since it only evaluates the quoted q at the value $\mathbf{h}$ that actually occurs.
- Since we do not know the value of $\mathbf{h}$ when asked to report q, we choose q such that the expected loss is minimised,

$$\mathbf{a}^*(f) = \operatorname*{argmin}_q \mathbb{E}_{f(\mathbf{h})}\left[\ell(\mathbf{h}, q)\right] \tag{28}$$

$$= \operatorname*{argmin}_q \mathbb{E}_{f(\mathbf{h})}\left[-\log q(\mathbf{h})\right] \tag{29}$$

- We have two distributions: f is the distribution of $\mathbf{h}$ we average over (e.g. prior, posterior, or data distribution, as before), while q is the reported distribution (the action).
- Note that $\mathbf{a}^*(f)$ is a distribution.
- Expected log loss $\mathbb{E}_{f(\mathbf{h})}\left[-\log q(\mathbf{h})\right]$ is called the cross-entropy of q relative to f.

Log-loss

- To solve the optimisation problem, note that

$$\begin{aligned}\mathbb{E}_{f(\mathbf{h})}\left[-\log q(\mathbf{h})\right] &= \mathbb{E}_{f(\mathbf{h})}\left[-\log q(\mathbf{h}) + \log f(\mathbf{h}) - \log f(\mathbf{h})\right] \\ &= \underbrace{\mathbb{E}_{f(\mathbf{h})}\left[\log \frac{f(\mathbf{h})}{q(\mathbf{h})}\right]}_{\mathrm{KL}(f||q)} \underbrace{-\mathbb{E}_{f(\mathbf{h})}\left[\log f(\mathbf{h})\right]}_{\text{entropy}} \end{aligned} \tag{30}$$

- The Kullback-Leibler (KL) divergence $\mathrm{KL}(f||q)$ is non-negative and zero iff $f = q$ (see later lectures)
- The entropy is the average surprise of f, a measure of the randomness of $\mathbf{h} \sim f(\mathbf{h})$. It does not depend on q.
- The optimal q thus equals f, i.e.

$$\mathbf{a}^*(f) = \operatorname*{argmin}_q \mathbb{E}_{f(\mathbf{h})}\left[-\log q(\mathbf{h})\right] = f(\mathbf{h}) \tag{31}$$

- As before, the optimal action is a property of f, here f itself.

Summary

- We derived the optimal action $\mathbf{a}^*(f) = \operatorname{argmin}_{\mathbf{a}} \mathbb{E}_{f(\mathbf{h})}[\ell(\mathbf{h}, \mathbf{a})]$ for different loss functions $\ell(\mathbf{h}, \mathbf{a})$.
- The optimal actions pick different aspects of f

loss	optimal action
quadratic	expected value of f
absolute error	median of f
pinball	any quantile of f
0-1	mode of f
log	f itself

- Allows us to obtain properties of f, and f itself, by solving an optimisation problem.
- Useful when f is only accessible through samples, e.g. observed data, so that the expectation can be approximated by a sample average.
- Learning arises from decision making (see next slides).

From decision making to learning

- By playing with different choices of $\mathbf{h}$, $f(\mathbf{h})$ and the loss $\ell(\mathbf{h}, \mathbf{a})$, we can frame classification, regression, inference, parameter and density estimation as decision problems. For example:
  - quadratic loss with $f = p(h|\mathbf{x})$ for a continuous label h: regression
  - 0-1 loss with $f = p(h|\mathbf{x})$ for a discrete label h: classification
- We next focus on the 0-1 loss and the log-loss.
- Both lead to maximum likelihood estimation, via MAP and via predictive density estimation, respectively.
- The log-loss also yields the posterior over the model parameters $\boldsymbol{\theta}$.

0-1 loss and MAP estimation

- We have seen that the optimal action for the 0-1 loss is

$$\mathbf{a}^*(f) = \operatorname*{argmax}_{\mathbf{h}} f(\mathbf{h}) \tag{32}$$

- If we use $p(\mathbf{h}|\mathbf{x})$ for $f(\mathbf{h})$, we obtain

$$\mathbf{a}^*(\mathbf{x}) = \operatorname*{argmax}_{\mathbf{h}} p(\mathbf{h}|\mathbf{x}) \tag{33}$$

$$= \operatorname*{argmax}_{\mathbf{h}} \log p(\mathbf{h}|\mathbf{x}) \tag{34}$$

  which is called the maximum a-posteriori (MAP) estimate.

- The MAP estimate is thus the optimal action under the posterior expected 0-1 loss, i.e. it minimises $\mathbb{E}_{p(\mathbf{h}|\mathbf{x})}[\ell(\mathbf{h}, \mathbf{a})]$ for $\ell(\mathbf{h}, \mathbf{a}) = \mathbb{1}(\mathbf{h} \neq \mathbf{a})$.

0-1 loss and MAP estimation

- Using Bayes' rule, $p(\mathbf{h}|\mathbf{x}) = p(\mathbf{x}|\mathbf{h})p(\mathbf{h})/p(\mathbf{x})$. Since $p(\mathbf{x})$ does not depend on $\mathbf{h}$, we obtain

$$\mathbf{a}^*(\mathbf{x}) = \operatorname*{argmax}_{\mathbf{h}} \log p(\mathbf{x}|\mathbf{h}) + \log p(\mathbf{h}) \tag{35}$$

- Let $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ be n observed samples, our data $\mathcal{D}$, and assume that they are conditionally independent given $\mathbf{h}$, so that $p(\mathbf{x}|\mathbf{h}) = \prod_i p(\mathbf{x}_i|\mathbf{h})$. Then

$$\mathbf{a}^*(\mathbf{x}) = \operatorname*{argmax}_{\mathbf{h}} \sum_{i=1}^{n} \log p(\mathbf{x}_i|\mathbf{h}) + \log p(\mathbf{h}) \tag{36}$$

- The first term is the log-likelihood and the second the log-prior. This holds for any unobserved $\mathbf{h}$, e.g. model parameters $\boldsymbol{\theta}$.

0-1 loss for parameter estimation

- If $\mathbf{h}$ are the parameters $\boldsymbol{\theta}$ of a model, the MAP estimate is

$$\hat{\boldsymbol{\theta}}_{MAP} = \operatorname*{argmax}_{\boldsymbol{\theta}} \sum_{i=1}^{n} \log p(\mathbf{x}_i|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) \tag{37}$$

- The log-prior acts as a regulariser, e.g. a Gaussian prior $p(\boldsymbol{\theta}) \propto \exp(-\|\boldsymbol{\theta}\|^2/(2\sigma_0^2))$ gives the L2 (ridge) penalty.
- The log-likelihood has n terms but the log-prior only one: the influence of the prior fades as n grows.
- If the prior is uniform, $\log p(\boldsymbol{\theta})$ does not depend on $\boldsymbol{\theta}$ and we obtain the maximum likelihood estimate (MLE)

$$\hat{\boldsymbol{\theta}}_{MLE} = \operatorname*{argmax}_{\boldsymbol{\theta}} \sum_{i=1}^{n} \log p(\mathbf{x}_i|\boldsymbol{\theta}) \tag{38}$$

- The MLE is thus the optimal action under the posterior expected 0-1 loss with a flat prior.

Example: estimating the mean of a Gaussian

- Data $x_1, \ldots, x_n$ with $p(x_i|\theta) = \mathcal{N}(x_i; \theta, \sigma^2)$, unknown mean θ, and prior $p(\theta) = \mathcal{N}(\theta; 0, \sigma_0^2)$, where

$$\mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- Log-likelihood and log-prior (up to constants)

$$-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \theta)^2 \qquad \text{and} \qquad -\frac{\theta^2}{2\sigma_0^2}$$

- Maximum likelihood estimate (sample mean) and MAP estimate

$$\hat{\theta}_{MLE} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \hat{\theta}_{MAP} = \frac{n}{n + \sigma^2/\sigma_0^2}\,\bar{x} \tag{39}$$

- The prior shrinks the estimate towards its mean 0; the smaller σ_0^2 relative to σ^2, the stronger the shrinkage.

Log-loss for predictive density estimation

$$\mathbf{a}^*(f) = \operatorname{argmin}_q \mathbb{E}_{f(\mathbf{h})}[-\log q(\mathbf{h})]$$

- Let $\mathbf{h}$ be a data point that is being revealed to us. Before we see $\mathbf{h}$, we are asked to report a distribution $q(\mathbf{h})$ over it, i.e. we predict $\mathbf{h}$ probabilistically.
- Let $f(\mathbf{h})$ be the true data distribution. We know that the optimal q is f itself, but f is unknown, so the risk cannot be computed.
- Assume that we have observed data $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, with $\mathbf{x}_i \sim f$ iid. The unobserved $\mathbf{h}$ is the next data point, i.e. $\mathbf{h} = \mathbf{x}_{n+1}$.
- Two ways to proceed:
  - Bayesian: replace f by our belief about $\mathbf{h}$ given the data $\mathcal{D}$.
  - Frequentist: approximate the expectation over f using the data $\mathcal{D}$.

Bayesian approach

- Our belief about the next data point $\mathbf{h}$ given the data $\mathcal{D}$ is the posterior predictive distribution $p(\mathbf{h}|\mathcal{D})$.
- With $f(\mathbf{h}) = p(\mathbf{h}|\mathcal{D})$, the optimal action under the log-loss is

$$\mathbf{a}^*(\mathcal{D}) = p(\mathbf{h}|\mathcal{D}) \tag{40}$$

- No optimisation is needed: the work lies in computing $p(\mathbf{h}|\mathcal{D})$, i.e. in inference.

Frequentist approach

- We approximate the expectation over f with a sample average over $\mathcal{D}$. This gives the so-called empirical risk

$$\hat{R}_n(q) = \frac{1}{n} \sum_{i=1}^{n} -\log q(\mathbf{x}_i) \tag{41}$$

- By the law of large numbers (see later), $\hat{R}_n(q) \to R(q; f) = \mathbb{E}_{f(\mathbf{h})} [-\log q(\mathbf{h})]$ as $n \to \infty$.
- Here, f is a fixed but unknown truth, and the data are only used to approximate the risk.

MLE as empirical risk minimisation

- Minimising $\hat{R}_n(q)$ over all distributions is degenerate: the optimal q concentrates all its probability mass on the observed data points. This is overfitting in its purest form.
- We thus restrict q to be a member of a parametric family $\mathcal{Q} = \{q(.\,; \boldsymbol{\theta})\}$. The problem then becomes

$$\hat{\boldsymbol{\theta}}_n = \operatorname*{argmin}_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^{n} -\log q(\mathbf{x}_i; \boldsymbol{\theta}) = \operatorname*{argmax}_{\boldsymbol{\theta}} \sum_{i=1}^{n} \log q(\mathbf{x}_i; \boldsymbol{\theta})$$

This is the maximum likelihood estimate. Note that $\hat{q}_n = q(.\,; \hat{\boldsymbol{\theta}}_n)$ depends on the samples through $\hat{\boldsymbol{\theta}}_n$.

Interpretation: MLE as KL minimisation

- Recall that risk $\mathbb{E}_{f(\mathbf{h})}\left[-\log q(\mathbf{h})\right] = \mathrm{KL}(f||q) + \text{entropy of } f$, where the entropy does not depend on q.
- Minimising the risk over $\mathcal{Q}$ thus gives

$$q^* = \operatorname*{argmin}_{q \in \mathcal{Q}} \mathrm{KL}(f||q), \tag{42}$$

  i.e. the member of $\mathcal{Q}$ that is closest to f in the KL sense.
- The maximum likelihood estimate $\hat{q}_n$ minimises the empirical risk $\hat{R}_n$, which approaches the risk as $n \to \infty$, and $\hat{q}_n$ approaches q^*.
- If $f \in \mathcal{Q}$ (well-specified model), then $q^* = f$: $\hat{q}_n$ recovers the data distribution as $n \to \infty$.
- If $f \notin \mathcal{Q}$ (misspecified model), $\hat{q}_n$ approaches the best approximation of f within $\mathcal{Q}$, in the KL sense.

Two routes to maximum likelihood estimation

	Via 0-1 loss	Via log-loss
approach	Bayesian (MAP)	frequentist (empirical risk)
unknown $\mathbf{h}$	parameters $\boldsymbol{\theta}$	next data point
$f(\mathbf{h})$	posterior $p(\boldsymbol{\theta}\|\mathcal{D})$	true data distribution (approximated by samples)
requires	model $p(\mathbf{x}\|\boldsymbol{\theta})$, flat prior	model $q(\mathbf{x};\boldsymbol{\theta})$
interpretation	most probable $\boldsymbol{\theta}$ (posterior mode)	for large n: closest $q \in \mathcal{Q}$ to f in KL divergence

Both routes lead to the same estimator (with $q(.\,;\boldsymbol{\theta}) = p(.|\boldsymbol{\theta})$), but justify it differently.

Beyond point estimates: posterior over the parameters

- The MAP estimate and the MLE are point estimates of $\boldsymbol{\theta}$. How about a distribution over $\boldsymbol{\theta}$?
- Same game as before, but now the unknown is the parameter $\boldsymbol{\theta}$ of a model $p(\mathbf{x}|\boldsymbol{\theta})$: before $\boldsymbol{\theta}$ is revealed, you report a distribution $q(\boldsymbol{\theta})$ over it and incur the log-loss $-\log q(\boldsymbol{\theta})$.
- Averaging over our belief about $\boldsymbol{\theta}$ given the data, $f(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|\mathcal{D})$, the optimal action is

$$\mathbf{a}^*(\mathcal{D}) = \operatorname*{argmin}_q \mathbb{E}_{p(\boldsymbol{\theta}|\mathcal{D})}\left[-\log q(\boldsymbol{\theta})\right] = p(\boldsymbol{\theta}|\mathcal{D}) \tag{43}$$

- Minimising the posterior expected loss thus gives the posterior itself under the log-loss, and its mode, the MAP estimate, under the 0-1 loss.

Program recap

1. Brief introduction to decision theory
   - Loss and decision principles
   - Connection to causality
   - Mild cognitive impairment example

2. Its use in statistics and machine learning
   - Common loss functions
   - Learning as a decision under the 0-1 loss
   - Learning as a decision under the log-loss