BMLIP-5SSD0

Bayesian Machine Learning and Information Processing

Solutions

Bert de Vries

Wouter Kouw (Flux 7.060)

Magnus T. Koudahl (Flux 7.060)

Ismail Senoz (Flux 7.060)

Probability Theory Review

- [1] (a) (#) Proof that the "elementary" sum rule $p(A) + p(\bar{A}) = 1$ follows from the (general) sum rule
$$p(A + B) = p(A) + p(B) - p(A, B)\,.$$

  (b) (###) Conversely, derive the general sum rule $p(A + B) = p(A) + p(B) - p(A, B)$ from the elementary sum rule $p(A) + p(\bar{A}) = 1$ and the product rule. Here, you may make use of the (Boolean logic) fact that $A + B = \overline{\bar{A}\bar{B}}$.

  $$\begin{aligned}
  p\left(A + B\right) &\underset{\text{bool}}{=} p\left(\overline{\bar{A}\bar{B}}\right) \\
  &\underset{\text{sum}}{=} 1 - p\left(\bar{A}\bar{B}\right) \\
  &\underset{\text{prod}}{=} 1 - p\left(\bar{A}|\bar{B}\right) p\left(\bar{B}\right) \\
  &\underset{\text{sum}}{=} 1 - \left(1 - p\left(A|\bar{B}\right)\right)\left(1 - p\left(B\right)\right) \\
  &= p(B) + \left(1 - p\left(B\right)\right) p\left(A|\bar{B}\right) \\
  &\underset{\text{prod}}{=} p(B) + \left(1 - p\left(B\right)\right) p\left(\bar{B}|A\right) \frac{p\left(A\right)}{p\left(\bar{B}\right)} \\
  &\underset{\text{sum}}{=} p(B) + p\left(\bar{B}|A\right) p\left(A\right) \\
  &\underset{\text{sum}}{=} p(B) + \left(1 - p\left(B|A\right)\right) p\left(A\right) \\
  &\underset{\text{sum}}{=} p\left(A\right) + p(B) - p\left(A, B\right)
  \end{aligned}$$

  Note that, aside from the first boolean rewrite, everything follows straight application of sum and product rules.

- [2] Box 1 contains 8 apples and 4 oranges. Box 2 contains 10 apples and 2 oranges. Boxes are chosen with equal probability.
  (a) (#) What is the probability of choosing an apple?
  (b) (##) If an apple is chosen, what is the probability that it came from box 1?

  The following probabilities are given in the problem statement,
  $$\begin{aligned}
  p(b_1) = p(b_2) &= 1/2 \\
  p(a|b_1) = 8/12, &\quad p(a|b_2) = 10/12 \\
  p(o|b_1) = 4/12, &\quad p(o|b_2) = 2/12
  \end{aligned}$$
  (a) $p(a) = \sum_i p(a, b_i) = \sum_i p(a|b_i)p(b_i) = \frac{8}{12} \cdot \frac{1}{2} + \frac{10}{12} \cdot \frac{1}{2} = \frac{3}{4}$
  (b) $p(b_1|a) = \frac{p(a,b_1)}{p(a)} = \frac{p(a|b_1)p(b_1)}{p(a)} = \frac{\frac{8}{12} \cdot \frac{1}{2}}{\frac{3}{4}} = \frac{4}{9}$

- [3] (###) The inhabitants of an island tell the truth one third of the time. They lie with probability $2/3$. On an occasion, after one of them made a statement, you ask another "was that statement true?" and he says "yes". What is the probability that the statement was indeed true?

We use variables S_1 and S_2 for statements 1 and 2 and shorthand "y", "n", "t" and "f" for "yes", "no", "true and "false", respectively. The problem statement provides us with the following probabilities,

$$\begin{aligned} p(S_1 = \mathrm{t}) &= 1/3 \\ p(S_1 = \mathrm{f}) &= 1 - p(S_1 = \mathrm{t}) = 2/3 \\ p(S_2 = \mathrm{y}|S_1 = \mathrm{t}) &= 1/3 \\ p(S_2 = \mathrm{y}|S_1 = \mathrm{f}) &= 1 - p(S_2 = \mathrm{y}|S_1 = \mathrm{t}) = 2/3 \end{aligned}$$

We are asked to compute $p(S_1 = \mathrm{t}|S_2 = \mathrm{y})$. Use Bayes rule,

$$\begin{aligned} p(S_1 = \mathrm{t}|S_2 = \mathrm{y}) &= \frac{p(S_1 = \mathrm{t}, S_2 = \mathrm{y})}{p(S_2 = \mathrm{y})} \\ &= \frac{p(S_2 = \mathrm{y}|S_1 = \mathrm{t})p(S_1 = \mathrm{t})}{p(S_2 = \mathrm{y}|S_1 = \mathrm{t})p(S_1 = \mathrm{t}) + p(S_2 = \mathrm{y}|S_1 = \mathrm{f})p(S_1 = \mathrm{f})} \\ &= \frac{\frac{1}{3} \cdot \frac{1}{3}}{\frac{1}{3} \cdot \frac{1}{3} + \frac{2}{3} \cdot \frac{2}{3}} = \frac{1}{5} \end{aligned}$$

- [4] (##) A bag contains one ball, known to be either white or black. A white ball is put in, the bag is shaken, and a ball is drawn out, which proves to be white. What is now the chance of drawing a white ball? (Note that the state of the bag, after the operations, is exactly identical to its state before.)

  There are two hypotheses: let $H = 0$ mean that the original ball in the bag was white and $H = 1$ that is was black. Assume the prior probabilities are equal. The data is that when a randomly selected ball was drawn from the bag, which contained a white one and the unknown one, it turned out to be white. The probability of this result according to each hypothesis is:

  $$P(D|H = 0) = 1, \quad P(D|H = 1) = 1/2$$

  So by Bayes theorem, the posterior probability of H is

  $$P(H = 0|D) = 2/3, \quad P(H = 1|D) = 1/3$$

- [5] A dark bag contains five red balls and seven green ones.
  (a) (#) What is the probability of drawing a red ball on the first draw?
  (b) (##) Balls are not returned to the bag after each draw. If you know that on the second draw the ball was a green one, what is now the probability of drawing a red ball on the first draw?

  (a) $p(S_1 = R) = \frac{N_R}{N_R + N_G} = \frac{5}{12}$
  (b) The outcome of the nth draw is referred to by variable S_n. Use Bayes rule to get

  $$\begin{aligned} p(S_1 = \mathrm{R}|S_2 = \mathrm{G}) &= \frac{p(S_2 = \mathrm{G}|S_1 = \mathrm{R})p(S_1 = \mathrm{R})}{p(S_2 = \mathrm{G}|S_1 = \mathrm{R})p(S_1 = \mathrm{R}) + p(S_2 = \mathrm{G}|S_1 = \mathrm{G})p(S_1 = \mathrm{G})} \\ &= \frac{\frac{7}{11} \cdot \frac{5}{12}}{\frac{7}{11} \cdot \frac{5}{12} + \frac{6}{11} \cdot \frac{7}{12}} = \frac{5}{11} \end{aligned}$$

- [6] (#) Is it more correct to speak about the likelihood of a model (or model parameters) than about the likelihood of an observed data set. And why?

  When a data generating distribution is considered as a function of the model parameters for given data, i.e. $L(\theta) \triangleq \log p(D|\theta)$, it is called a likelihood. It is more correct to speak about the likelihood of a model (or of the likelihood of the parameters).

- [7] (##) Is a speech signal a 'probabilistic' (random) or a deterministic signal?

That depends. The term 'probabilistic' refers to a state-of-knowledge (or beliefs) about something (in this case, about the values of a speech signal). The fundamental issue here is to realize that the signal itself is not probabilistic (nor deterministic), but rather that these attributes reflect a state-of-knowledge. If you had a perfect microphone and recorded a speech signal perfectly at its source, then you would know all the signal values perfectly. You could say that the signal is deterministic since there is no uncertainty. However, before you would record the signal, how would you describe your state-of-knowledge about the signal values that your are going to record? There is uncertainty, so you would need to describe that speech signal by a probability distribution over all possible values.

- [8] (##) Proof that, for any distribution of $\boldsymbol{x}$ and $\boldsymbol{y}$ and $\boldsymbol{z} = \boldsymbol{x} + \boldsymbol{y}$

$$\begin{aligned}\mathbb{E}[z] &= \mathbb{E}[x] + \mathbb{E}[y] \\ \mathbb{V}[z] &= \mathbb{V}[x] + \mathbb{V}[y] + 2\mathbb{V}[x, y]\end{aligned}$$

  where $\mathbb{E}[\cdot]$, $\mathbb{V}[\cdot]$ and $\mathbb{V}[\cdot, \cdot]$ refer to the expectation (mean), variance and covariance operators respectively. You may make use of the more general theorem that the mean and variance of any distribution $\boldsymbol{p(x)}$ is processed by a linear tranformation as

$$\begin{aligned}\mathbb{E}[Ax + b] &= A\mathbb{E}[x] + b \\ \mathbb{V}[Ax + b] &= A\,\mathbb{V}[x]\,A^T\end{aligned}$$

  Define $\boldsymbol{A} = [\boldsymbol{I}, \boldsymbol{I}]$, $\boldsymbol{w} = [\boldsymbol{x}; \boldsymbol{y}]$ (where the notation ";" stacks the columns of $\boldsymbol{x}$ and $\boldsymbol{y}$ and $\boldsymbol{I}$ is the identity matrix). Then $\boldsymbol{z} = \boldsymbol{Aw}$. Now apply the formula for the mean and variance of a RV after a linear transformation.

$$\begin{aligned}\mathbb{E}[z] &= \mathbb{E}[Aw] \\ &= \mathbb{E}[x + y] \\ &= \mathbb{E}[x] + \mathbb{E}[y] \\ \mathbb{V}[z] &= \mathbb{V}[Aw] \\ &= A\mathbb{V}[w]A^T \\ &= \begin{bmatrix} I & I \end{bmatrix} \begin{bmatrix} \mathbb{V}[x] & \mathbb{V}[x, y] \\ \mathbb{V}[x, y] & \mathbb{V}[y] \end{bmatrix} \begin{bmatrix} I \\ I \end{bmatrix} \\ &= \mathbb{V}[x] + \mathbb{V}[y] + 2\mathbb{V}[x, y]\end{aligned}$$

Bayesian Machine Learning

- [1] (#) (a) Explain shortly the relation between machine learning and Bayes rule.
  (b) How are Maximum a Posteriori (MAP) and Maximum Likelihood (ML) estimation related to Bayes rule and machine learning?

  > (a) Machine learning is inference over models (hypotheses, parameters, etc.) from a given data set. Bayes rule makes this statement precise. Let $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and $\boldsymbol{D}$ represent a model parameter vector and the given data set, respectively. Then, Bayes rule,
  >
  > $$p(\theta|D) = \frac{p(D|\theta)}{p(D)} p(\theta)$$
  >
  > relates the information that we have about $\boldsymbol{\theta}$ before we saw the data (i.e., the distribution $\boldsymbol{p(\theta)}$) to what we know after having seen the data, $\boldsymbol{p(\theta|D)}$.
  > (b) The Maximum a Posteriori (MAP) estimate picks a value $\hat{\boldsymbol{\theta}}$ for which the posterior distribution $\boldsymbol{p(\theta|D)}$ is maximal, i.e.,
  >
  > $$\hat{\theta}_{MAP} = \arg\max_{\theta} p(\theta|D)$$
  >
  > In a sense, MAP estimation approximates Bayesian learning, since we approximated $p(\theta|D)$ by $\delta(\theta - \hat{\theta}_{\mathrm{MAP}})$. Note that, by Bayes rule,
  >
  > $$\arg\max_{\theta} p(\theta|D) = \arg\max_{\theta} p(D|\theta)p(\theta)$$
  >
  > If we further assume that prior to seeing the data all values for $\boldsymbol{\theta}$ are equally likely (i.e., $p(\theta) = \mathbf{const.}$), then the MAP estimate reduces to the Maximum Likelihood estimate,
  >
  > $$\hat{\theta}_{ML} = \arg\max_{\theta} p(D|\theta)$$

- [2] (#) What are the four stages of the Bayesian design approach?

  > (1) Model specification, (2) parameter estimation, (3) model evaluation and (4) application of the model to tasks.

- [3] (##) The Bayes estimate is a summary of a posterior distribution by a delta distribution on its mean, i.e.,

  $$\hat{\theta}_{bayes} = \int \theta\, p\left(\theta|D\right)\, \mathrm{d}\theta$$

  Proof that the Bayes estimate minimizes the expected mean-squared error, i.e., proof that

  $$\hat{\theta}_{bayes} = \arg\min_{\hat{\theta}} \int_{\theta} (\hat{\theta} - \theta)^2 p\left(\theta|D\right)\, \mathrm{d}\theta$$

To minimize the expected mean-squared error we will look for $\hat{\theta}$ that makes the gradient of the integral with respect to $\hat{\theta}$ vanish.

$$\begin{aligned}
\nabla_{\hat{\theta}} \int_{\theta} (\hat{\theta} - \theta)^2 p(\theta|D) \, \mathrm{d}\theta &= 0 \\
\int_{\theta} \nabla_{\hat{\theta}} (\hat{\theta} - \theta)^2 p(\theta|D) \, \mathrm{d}\theta &= 0 \\
\int_{\theta} 2(\hat{\theta} - \theta) p(\theta|D) \, \mathrm{d}\theta &= 0 \\
\int_{\theta} \hat{\theta} p(\theta|D) \, \mathrm{d}\theta &= \int_{\theta} \theta p(\theta|D) \, \mathrm{d}\theta \\
\hat{\theta} \underbrace{\int_{\theta} p(\theta|D) \, \mathrm{d}\theta}_{1} &= \int_{\theta} \theta p(\theta|D) \, \mathrm{d}\theta \\
\hat{\theta} &= \int_{\theta} \theta p(\theta|D) \, \mathrm{d}\theta
\end{aligned}$$

- [4] (###) We make N IID observations $D = \{x_1 \dots x_N\}$ and assume the following model

  $$x_k = A + \epsilon_k$$

  where $\epsilon_k = \mathcal{N}(\epsilon_k|0, \sigma^2)$ with known $\sigma^2 = 1$. We are interested in deriving an estimator for A.
  (a) Make a reasonable assumption for a prior on A and derive a Bayesian (posterior) estimate.
  (b) (##) Derive the Maximum Likelihood estimate for A.
  (c) Derive the MAP estimates for A.
  (d) Now assume that we do not know the variance of the noise term? Describe the procedure for Bayesian estimation of both A and σ^2 (No need to fully work out to closed-form estimates).

  (a) Since there is no assumption on the values A can take it makes sense to assume a distribution that has support over the reals. A Gaussian prior is a good candidate. Let us assume $p(A) = \mathcal{N}(A|m_A, v_A)$. Since $p(D|A) = \prod_k \mathcal{N}(x_k|A, \sigma^2)$ is a Gaussian likelihood and $p(A)$ is a Gaussian prior, their multiplication is proportional to a Gaussian. We will work this out with the canonical parameterization of the Gaussian since it is easier to multiply Gaussians in that domain. This means the posterior $p(A|D)$ is

  $$\begin{aligned}
  p(A|D) &\propto p(A)p(D|A) \\
  &= \mathcal{N}(A|m_A, v_A) \prod_{k=1}^{N} \mathcal{N}(x_k|A, \sigma^2) \\
  &= \mathcal{N}(A|m_A, v_A) \prod_{k=1}^{N} \mathcal{N}(A|x_k, \sigma^2) \\
  &= \mathcal{N}_c\big(A \,\big|\, \frac{m_A}{v_A}, \frac{1}{v_A}\big) \prod_{k=1}^{N} \mathcal{N}_c\big(A \,\big|\, \frac{x_k}{\sigma^2}, \frac{1}{\sigma^2}\big) \\
  &\propto \mathcal{N}_c\big(A \,\big|\, \frac{m_A}{v_A} + \frac{1}{\sigma^2} \sum_k x_k, \frac{1}{v_A} + \frac{N}{\sigma^2}\big) \,,
  \end{aligned}$$

  where we have made use of the fact that precision-weighted means and precisions add when multiplying Gaussians. In principle this description of the posterior completes the answer.
  (b) The ML estimate can be found by

$$\nabla \log p(D|A) = 0$$
$$\nabla \sum_k \log \mathcal{N}(x_k|A,\sigma^2) = 0$$
$$\nabla \frac{-1}{2} \sum_k \frac{(x_k - A)^2}{\sigma^2} = 0$$
$$\sum_k (x_k - A) = 0$$
$$\hat{A}_{ML} = \frac{1}{N} \sum_{k=1}^{N} x_k$$

(c) The MAP is simply the location where the posterior has its maximum value, which for a Gaussian posterior is its mean value. We computed in (a) the precision-weighted mean, so we need to divide by precision (or multiply by variance) to get the location of the mean:

$$\hat{A}_{MAP} = \left(\frac{m_A}{v_A} + \frac{1}{\sigma^2} \sum_k x_k \right) \cdot \left(\frac{1}{v_A} + \frac{N}{\sigma^2} \right)^{-1}$$
$$= \frac{v_A \sum_k x_k + \sigma^2 m_A}{N v_A + \sigma^2}$$

(d) A Bayesian treatment requires putting a prior on the unknown variance. The variance is constrained to be positive hence the support of the prior distribution needs to be on the positive reals. (In a multivariate case positivity needs to be extended to symmetric positive definiteness.) Choosing a conjugate prior will simplify matters greatly. In this scenerio the inverse Gamma distribution is the conjugate prior for the unknown variance. In the literature this model is called a Normal-Gamma distribution. See https://www.seas.harvard.edu/courses/cs281/papers/murphy-2007.pdf for the analytical treatment.

- [5] (##) We consider the coin toss example from the notebook and use a conjugate prior for a Bernoulli likelihood function.
  (a) Derive the Maximum Likelihood estimate.
  (b) Derive the MAP estimate.
  (c) Do these two estimates ever coincide (if so under what circumstances)?

(a)

$$\begin{aligned}\nabla \log p(D|\mu) &= 0 \\ \nabla \left(n \log \mu + (N-n)\log(1-\mu)\right) &= 0 \\ \frac{n}{\mu} - \frac{N-n}{1-\mu} &= 0 \\ \hat{\mu}_{\mathrm{ML}} &= \frac{n}{N}\end{aligned}$$

(b) Assuming a beta prior $\mathcal{B}(\mu|\alpha,\beta)$, we can write the posterior as as

$$\begin{aligned}p(\mu|D) &\propto p(D|\mu)p(\mu) \\ &\propto \mu^n(1-\mu)^{N-n}\mu^{\alpha-1}(1-\mu)^{\beta-1} \\ &\propto \mathcal{B}(\mu|n+\alpha, N-n+\beta)\end{aligned}$$

The MAP estimate for a beta distribution $\mathcal{B}(a,b)$ is located at $\frac{a-1}{a+b-2}$, see wikipedia. Hence,

$$\begin{aligned}\hat{\mu}_{\mathrm{MAP}} &= \frac{(n+\alpha)-1}{(n+\alpha)+(N-n+\beta)-2} \\ &= \frac{n+\alpha-1}{N+\alpha+\beta-2}\end{aligned}$$

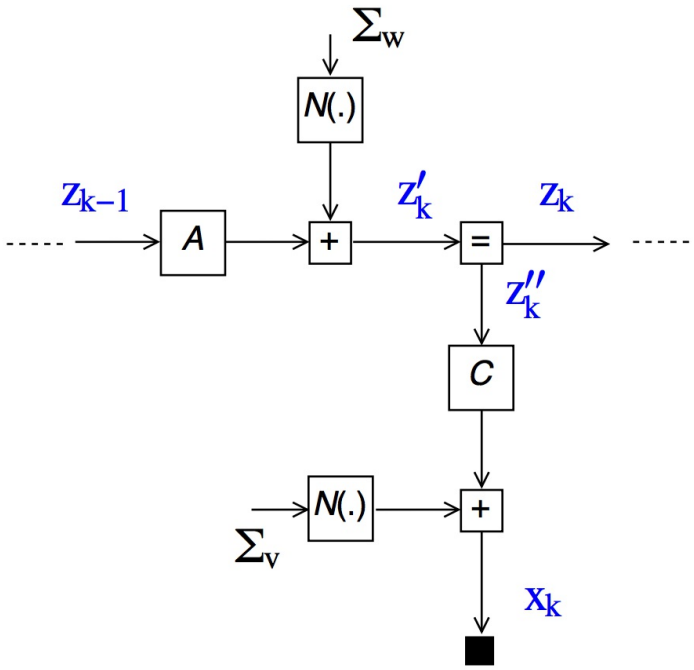
(c) As N gets larger, the MAP estimate approaches the ML estimate. In the limit the MAP solution converges to the ML solution.

Factor Graphs

- [1] Consider the following state-space model:

$$z_k = Az_{k-1} + w_k$$
$$x_k = Cz_k + v_k$$

where $k = 1, 2, \ldots, n$ is the time step counter; z_k is an unobserved state sequence; x_k is an observed sequence; $w_k \sim \mathcal{N}(0, \Sigma_w)$ and $v_k \sim \mathcal{N}(0, \Sigma_v)$ are (unobserved) state and observation noise sequences respectively; $z_0 \sim \mathcal{N}(0, \Sigma_0)$ is the initial state and A, C, Σ_v,Σ_w and Σ_0 are known parameters. The Forney-style factor graph (FFG) for one time step is depicted here:



(a) Rewrite the state-space equations as a set of conditional probability distributions.

$$p(z_k|z_{k-1}, A, \Sigma_w) = \ldots$$
$$p(x_k|z_k, C, \Sigma_v) = \ldots$$
$$p(z_0|\Sigma_0) = \ldots$$

$$p(z_k|z_{k-1}, A, \Sigma_w) = \mathcal{N}(z_k|Az_{k-1}, \Sigma_w)$$
$$p(x_k|z_k, C, \Sigma_v) = \mathcal{N}(x_k|Cz_k, \Sigma_v)$$
$$p(z_0|\Sigma_0) = \mathcal{N}(z_0|0, \Sigma_0)$$

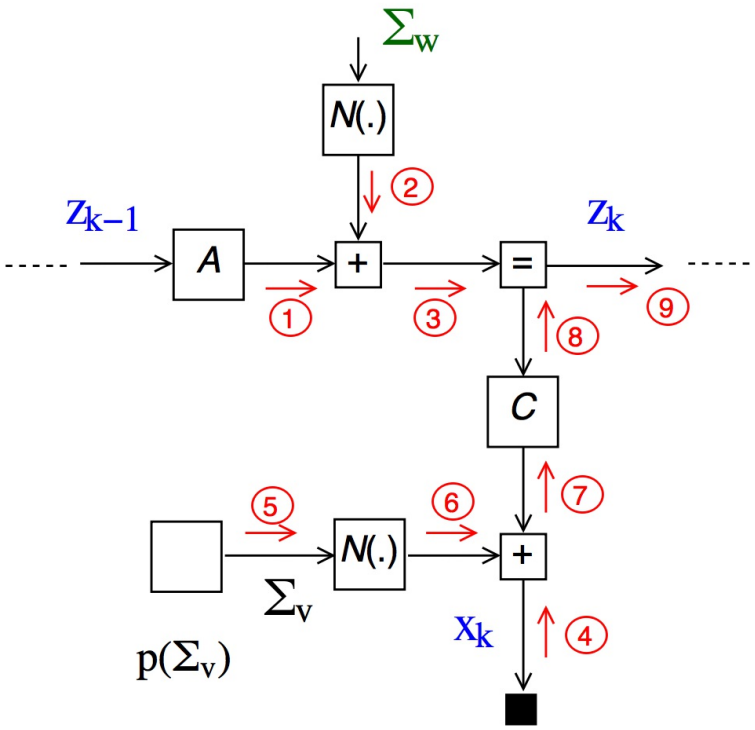
(b) Define $z^n \triangleq (z_0, z_1, \ldots, z_n)$, $x^n \triangleq (x_1, \ldots, x_n)$ and $\theta = \{A, C, \Sigma_w, \Sigma_v\}$. Now write out the generative model $p(x^n, z^n|\theta)$ as a product of factors.

$$\begin{aligned} p(x^n, z^n|\theta) &= p(z_0|\Sigma_0) \prod_{k=1}^{n} p(x_k|z_k, C, \Sigma_v)\, p(z_k|z_{k-1}, A, \Sigma_w) \\ &= \mathcal{N}(z_0|0, \Sigma_0) \prod_{k=1}^{n} \mathcal{N}(x_k|Cz_k, \Sigma_v)\, \mathcal{N}(z_k|Az_{k-1}, \Sigma_w) \end{aligned}$$

(c) We are interested in estimating z_k from a given estimate for z_{k-1} and the current observation x_k, i.e., we are interested in computing $p(z_k|z_{k-1}, x_k, \theta)$. Can $p(z_k|z_{k-1}, x_k, \theta)$ be expressed as a Gaussian distribution? Explain why or why not in one sentence.

Yes, since the generative model $p(x^n, z^n|\theta)$ is (one big) Gaussian.

(d) Copy the graph onto your exam paper and draw the message passing schedule for computing $p(z_k|z_{k-1}, x_k, \theta)$ by drawing arrows in the factor graph. Indicate the order of the messages by assigning numbers to the arrows.



Some permutations of this order are also possible. The most important thing here is that you recognize the tree with Z_k as a root of the tree and pass messages from the terminals (e.g., Z_{k-1}, X_k, etc.) towards the root.

(e) Now assume that our belief about parameter Σ_v is instead given by a distribution $p(\Sigma_v)$ (rather than a known value). Adapt the factor graph drawing of the previous answer to reflects our belief about Σ_v.

See drawing in previous answer.

Continuous Data and the Gaussian Distribution

- [1] (##) We are given an IID data set $\boldsymbol{D} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$, where $\boldsymbol{x}_n \in \mathbb{R}^M$. Let's assume that the data were drawn from a multivariate Gaussian (MVG),

$$p(\boldsymbol{x}_n|\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{x}_n|\,\boldsymbol{\mu}, \boldsymbol{\Sigma}) = |2\pi\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu})\right\}$$

  (a) Derive the log-likelihood of the parameters for these data.
  (b) Derive the maximum likelihood estimates for the mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$ by setting the derivative of the log-likelihood to zero.

(a) Let $\theta = \mu, \Sigma$. Then the log-likelihood can be worked out as

$$\begin{aligned}
\log p(D|\theta) &= \log \prod_n p(x_n|\theta) \\
&= \log \prod_n \mathcal{N}(x_n|\mu, \Sigma) \\
&= \log \prod_n (2\pi)^{-M/2} |\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(x_n - \mu)^T \Sigma^{-1}(x_n - \mu)\right) \\
&= \sum_n \left(\log(2\pi)^{-M/2} + \log|\Sigma|^{-1/2} - \frac{1}{2}(x_n - \mu)^T \Sigma^{-1}(x_n - \mu)\right) \\
&\propto \frac{N}{2}\log|\Sigma|^{-1} - \frac{1}{2}\sum_n (x - \mu)^T \Sigma^{-1}(x - \mu)
\end{aligned}$$

(b) First we take the derivative with respect to the mean.

$$\begin{aligned}
\nabla_\mu \log p(D|\theta) &\propto -\sum_n \nabla_\mu (x_n - \mu)^T \Sigma^{-1} (x_n - \mu) \\
&= -\sum_n \nabla_\mu \mathrm{Tr}\left[-2\mu^T \Sigma^{-1} x_n + \mu^T \Sigma^{-1} \mu\right] \\
&= -\sum_n \left(-2\Sigma^{-1} x_n + 2\Sigma^{-1}\mu\right) \\
&= \Sigma^{-1} \sum_n (x_n - \mu)
\end{aligned}$$

Setting the derivative to zeros leads to $\hat{\mu} = \frac{1}{N}\sum_n x_n$. The derivative with respect to covariance is a bit more involved. It's actually easier to compute this by taking the derivative to the precision:

$$\begin{aligned}
\nabla_{\Sigma^{-1}} \log p(D|\theta) &= \nabla_{\Sigma^{-1}} \left(\frac{N}{2}\log|\Sigma|^{-1} - \frac{1}{2}\sum_n (x_n - \mu)^T \Sigma^{-1}(x_n - \mu)\right) \\
&= \nabla_{\Sigma^{-1}} \left(\frac{N}{2}\log|\Sigma|^{-1} - \frac{1}{2}\sum_n \mathrm{Tr}\left[(x_n - \mu)(x_n - \mu)^T \Sigma^{-1}\right]\right) \\
&= \frac{N}{2}\Sigma - \frac{1}{2}\sum_n (x_n - \mu)(x_n - \mu)^T
\end{aligned}$$

Setting the derivative to zero leads to $\hat{\Sigma} = \frac{1}{N}\sum_n (x_n - \hat{\mu})(x_n - \hat{\mu})^T$.

- [2] (#) Shortly explain why the Gaussian distribution is often preferred as a prior distribution over other distributions with the same support? You can get this answer straight from the lession notebook. Aside from the computational advantages (operations on distributions tends to make them more Gaussian, and Gaussians tends to remain Gaussians in computational manipulations), the Gaussian distribution is also the maximum-entropy distribution among distributions that are defined over real numbers. This means that there is no distribution with the same variance that assumes less information about its argument.

- [3] (###) Proof that the Gaussian distribution is the maximum entropy distribution over the reals with specified mean and variance.

This is a challenging question (e.g., too diffucult for a written exam:) which requires calculus of variations to solve rigorously. We will show how to maximize the entropy functional which is $-\int q(x)\log q(x)\mathrm{d}x$ with the specified constraints. We have three constraints: (1) we require $q(x)$ to be normalized, (2) $\mathbb{E}[x]=m$ and (3) $\mathbb{E}[x^2]=m^2+\sigma^2$, where $m\in\mathbb{R}$ and $\sigma^2\in\mathbb{R}^+$ are arbitrary. Let us write entropy with the given constraints with undetermined multipliers as a Lagrangian

$$L[q]=-\int q(x)\log q(x)\mathrm{d}x+\lambda\left(\int xq(x)\mathrm{d}x-m\right)$$
$$+\gamma\left(\int x^2q(x)\mathrm{d}x-(\sigma^2+m^2)\right)+\psi\left(\int q(x)\mathrm{d}x-1\right).$$

We are searching for a distribution in a space of functions that minimizes the above Lagrangian. This is is a functional minimization problem that is defined over a function space as opposed to ordinary ($\mathbb{R}^N$). Even though the computational mechanics are somewhat different the idea is same with ordinary minimization problems. We look at the functional derivative that has a similar interpretation as a gradient(It can be thought of as the derivative of a functional with respect to a function). We want to solve

$$\frac{\delta L[q]}{\delta q}=0$$
$$-\log q(x)+\psi+\lambda x+\gamma x^2=0$$
$$q(x)=\exp(+\psi+\lambda x+\gamma x^2)$$

where $\frac{\delta L[q]}{\delta q}$ is the functional derivative. We can plug $q(x)$ back into the constraints and solve for the multipliers. Doing that we obtain $\lambda=\frac{m}{\sigma^2}$,$\gamma=-\frac{1}{2\sigma^2}$ and $\psi=-\frac{m^2}{2\sigma^2}-\log\sqrt{2\pi\sigma^2}$. This means the distribution that maximizes entropy, $q(x)$, is a Gaussian distribution.

- [4] (##) Proof that a linear transformation $z=Ax+b$ of a Gaussian variable $\mathcal{N}(x|\mu,\Sigma)$ is Gaussian distributed as

$$p(z)=\mathcal{N}\left(z\,|\,A\mu+b,A\Sigma A^T\right)$$

First, we show that a linear transformation of a Gaussian is a Gaussian. In general, the transformed distribution of $z=g(x)$ is given by

$$p_Z(z)=\frac{p_X(g^{-1}(z))}{\det[g(z)]}.$$

Since the transformation is linear, $\det[g]=\det[A]$, which is independent of z, and consequently $p_Z(z)$ has the same functional form as $p_X(x)$, i.e. $p_Z(z)$ is a also Gaussian. The mean and variance can easily be determined by the calculation that we used in question 8 of the Probability Theory exercises. This results in $p(z)=\mathcal{N}\left(z\,|\,A\mu+b,A\Sigma A^T\right)$.

- [5] (#) Given independent variables $x\sim\mathcal{N}(\mu_x,\sigma_x^2)$ and $y\sim\mathcal{N}(\mu_y,\sigma_y^2)$, what is the PDF for $z=A\cdot(x-y)+b$?

z is also Gaussian with

$$p_z(z)=\mathcal{N}(z\,|\,A(\mu_x-\mu_y)+b,\;A(\sigma_x^2+\sigma_y^2)A^T)$$

- [6] (###) Compute

$$\int_{-\infty}^{\infty}\exp(-x^2)\mathrm{d}x\,.$$

For a Gaussian with zero mean and varance equal to 1 we have

$$\int \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}x^2) = 1$$

Substitution of $y = \sqrt{2}x$ with $\mathrm{d}x = \sqrt{2}\mathrm{d}y$ will simply lead you to $\int_{-\infty}^{\infty} \exp(-x^2)\mathrm{d}x = \sqrt{\pi}$. If you don't want to use the result of the Gaussian integral, you can still do this integral, see youtube clip.

Discrete Data and the Multinomial Distribution

- [1] (##) We consider IID data $D = \{x_1, x_2, \ldots, x_N\}$ obtained from tossing a K-sided die. We use a binary selection variable

$$x_{nk} \equiv \begin{cases} 1 & \text{if } x_n \text{ lands on } k\text{th face} \\ 0 & \text{otherwise} \end{cases}$$

with probabilities $p(x_{nk} = 1) = \theta_k$.
(a) Write down the probability for the nth observation $p(x_n|\theta)$ and derive the log-likelihood $\log p(D|\theta)$.
(b) Derive the maximum likelihood estimate for θ.

> See lecture notes (on class homepage).
> (a) $p(x_n|\theta) = \prod_k \theta_k^{x_{nk}}$ subject to $\sum_k \theta_k = 1$.
>
> $$\ell(\theta) = \sum_k m_k \log \theta_k$$
>
> where $m_k = \sum_n x_{nk}$.
> (b) $\hat{\theta} = \frac{m_k}{N}$, the sample proportion.

- [2] (#) In the notebook, Laplace's generalized rule of succession (the probability that we throw the kth face at the next toss) was derived as

$$p(x_{\bullet,k} = 1|D) = \frac{m_k + \alpha_k}{N + \sum_k \alpha_k}$$

Provide an interpretation of the variables $m_k, N, \alpha_k, \sum_k \alpha_k$.

> m_k is the total number of occurances that we threw k eyes, α_k is the prior pseudo counts representing the number of observations in the kth that we assume to have seen already. $\sum_k m_k = N$ is the total number of rolls and $\sum_k \alpha_k$ is the total number of prior pseudo rolls.

- [3] (##) Show that Laplace's generalized rule of succession can be worked out to a prediction that is composed of a prior prediction and data-based correction term.

$$\begin{aligned} p(x_{\bullet,k} = 1|D) &= \frac{m_k + \alpha_k}{N + \sum_k \alpha_k} \\ &= \frac{N}{N + \sum_k \alpha_k} \frac{m_k}{N} + \frac{\sum_k \alpha_k}{N + \sum_k \alpha_k} \frac{\alpha_k}{\sum_k \alpha_k} \\ &= \underbrace{\frac{\alpha_k}{\sum_k \alpha_k}}_{\text{prior prediction}} + \underbrace{\frac{N}{N + \sum_k \alpha_k} \cdot \underbrace{\left(\frac{m_k}{N} - \frac{\alpha_k}{\sum_k \alpha_k} \right)}_{\text{prediction error}}}_{\text{data-based correction}} \end{aligned}$$

- [4] (#) Verify that
(a) the categorial distribution is a special case of the multinomial for $N = 1$.
(b) the Bernoulli is a special case of the categorial distribution for $K = 2$.
(c) the binomial is a special case of the multinomial for $K = 2$.

(a) The probability mass function of a multinomial distribution is $p(D_m|\mu) = \frac{N!}{m_1!m_2!...m_K!} \prod_k \mu_k^{m_k}$ over the data frequencies $D_m = \{m_1, \ldots, m_K\}$ with the constraint that $\sum_k \mu_k = 1$ and $\sum_k m_k = N$. Setting $N = 1$ we see that $p(D_m|\mu) \propto \prod_k \mu_k^{m_k}$ with $\sum_k m_k = 1$, making the sample space one-hot coded given by the categorical distribution.
(b) When $K = 2$, the constraint for the categorical distribution takes the form $m_1 = 1 - m_2$ leading to $p(D_m|\mu) \propto \mu_1^{m_1} (1 - \mu_1)^{1-m_1}$ which is associated with the Bernoulli distribution.
(c) Plugging $K = 2$ into the multinomial distribution leads to $p(D_m|\mu) = \frac{N!}{m_1!m_2!}\mu_1^{m_1} (\mu_2^{m_2})$ with the constraints $m_1 + m_2 = N$ and $\mu_1 + \mu_2 = 1$. Then plugging the constraints back in we obtain $p(D_m|\mu) = \frac{N!}{m_1!(N-m1)!}\mu_1^{m_1} (1 - \mu_1)^{N-m_1}$ as the binomial distribution.

- [5] (###) Determine the mean, variance and mode of a Beta distribution.

The Beta distribution is given by $\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}$. Define $\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \triangleq \mathcal{B}(\alpha,\beta)$, which is the normalization constant. Notice that this definition makes $\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\mathrm{d}x = \mathcal{B}(\alpha,\beta)$. Together with $\Gamma(x+1) = x\Gamma(x)$ we can use these identities to obtain the requested statistics:

$$\begin{aligned}
\mathbb{E}[x] &= \frac{1}{\mathcal{B}(\alpha,\beta)}\int_0^1 x x^{\alpha-1}(1-x)^{\beta-1}\mathrm{d}x \\
&= \frac{1}{\mathcal{B}(\alpha,\beta)}\int_0^1 x^{\alpha}(1-x)^{\beta-1}\mathrm{d}x \\
&= \frac{\mathcal{B}(\alpha+1,\beta)}{\mathcal{B}(\alpha,\beta)} \\
&= \frac{\Gamma(\alpha+1)\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\alpha+\beta+1)} \\
&= \frac{\alpha\Gamma(\alpha)\Gamma(\alpha+\beta)}{(\alpha+\beta)\Gamma(\alpha)\Gamma(\alpha+\beta)} \\
&= \frac{\alpha}{\alpha+\beta} \\
\mathbb{V}[x] &= \mathbb{E}[x^2] - \mathbb{E}[x]^2 \\
&= \frac{1}{\mathcal{B}(\alpha,\beta)}\int_0^1 x^2 x^{\alpha-1}(1-x)^{\beta-1}\mathrm{d}x - \frac{\alpha^2}{(\alpha+\beta)^2} \\
&= \frac{\mathcal{B}(\alpha+2,\beta)}{\mathcal{B}(\alpha,\beta)} - \frac{\alpha^2}{(\alpha+\beta)^2} \\
&= \frac{\alpha}{\alpha+\beta}\left(\frac{\alpha+1}{\alpha+\beta+1} - \frac{\alpha}{\alpha+\beta}\right) \\
&= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}
\end{aligned}$$

If $\alpha = \beta$, then the Beta distribution is identical to a uniform distribution, which doesn't have a unique mode. If one of the parameters is < 1, then the mode is at one of the edges. When both parameters are > 1, then the mode is well-defined and is within the interior of the distribution. Assuming the parameters are > 1 we can evaluate the mode as

$$\begin{aligned}
\nabla_x x^{\alpha-1}(1-x)^{\beta-1} &= 0 \\
\frac{\alpha-1}{\beta-1} &= \frac{x}{1-x} \\
\alpha - 1 &= x(\alpha+\beta-2) \\
\Rightarrow x_{mode} &= \frac{\alpha-1}{\alpha+\beta-2}.
\end{aligned}$$

- [6] (###) Consider a data set of binary variables $D = \{x_1, x_2, \ldots, x_N\}$ with a Bernoulli distribution $\mathrm{Ber}(x_k|\mu)$ as data generating distribution and a Beta prior for μ. Assume that you make n observations with $x = 1$ and $N - n$ observations with $x = 0$. Now consider a new draw $x_\bullet$. We are interested in computing $p(x_\bullet|D)$. Show that the mean value for $p(x_\bullet|D)$ lies in between the prior mean and Maximum Likelihood estimate.

In the lectures we have seen that $p(x_\bullet = 1|D) = \frac{a+n}{a+b+N}$, where a and b are parameters of the Beta prior. The ML estimate is $\frac{n}{N}$ and the prior mean is $\frac{a}{a+b}$. To show that the prediction lies in between ML and prior estimate, we will try to write the prediction as a convex combination of the latter two. That is we want to solve for λ

$$(1-\lambda)\frac{n}{N} + \lambda\frac{a}{a+b} = \frac{a+n}{a+b+N}$$
$$\lambda = \frac{1}{1+\frac{N}{a+b}}$$

Since a, b and N are positive, it follows that $0 < \lambda < 1$. This means the prediction is a convex combination of prior and ML estimates and thus lies in between the two.

- [7] Consider a data set $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$ with 1-of-K notation for the discrete classes, i.e., \begin{equation*} y_{nk} = \begin{cases} 1 & \text{if } y_n \text{ in } k\text{th class} \\
```
  0 & \text{otherwise}
 \end{cases}
```
\end{equation} together with class-conditional distribution $p(x_n|y_{nk} = 1, \theta) = \mathcal{N}(x_n|\mu_k, \Sigma)$ and multinomial prior $p(y_{nk} = 1) = \pi_k$.
(a) Proof that the joint log-likelihood is given by $$\begin{equation} \log p(D|\theta) = \sum_{n,k} y_{nk} \log \mathcal{N}(x_n|\mu_k,\Sigma) + \sum_{n,k} y_{nk} \log \pi_k \end{equation*}$$

$$\begin{aligned}\log p(D|\theta) &= \sum_n \log \prod_k p(x_n, y_{nk}|\theta)^{y_{nk}} \\ &= \sum_{n,k} y_{nk} \log p(x_n, y_{nk}|\theta) \\ &= \sum_{n,k} y_{nk} \log \mathcal{N}(x_n|\mu_k, \Sigma) + \sum_{n,k} y_{nk} \log \pi_k\end{aligned}$$

(b) Show now that the MLE of the class-conditional mean is given by

$$\hat{\mu}_k = \frac{\sum_n y_{nk} x_n}{\sum_n y_{nk}}$$

Regression

- [1] (#) (a) Write down the generative model for Bayesian linear ordinary regression (i.e., write the likelihood and prior).
  (b) State the inference task for the weight parameter in the model.
  (c) Why do we call this problem linear?

  (a)

  $$\text{likelihood: } p(y_n|x_n, w) = \mathcal{N}(y_n|w^T\phi(x_n), \beta^{-1})$$
  $$\text{prior: } p(w|\alpha) = \mathcal{N}(w|0, \alpha^{-1}I)$$

  (b) The inference task is to compute

  $$p(w|D) = \frac{p(D|w)p(w)}{p(D)}$$

  (c) The model is linear with respect to $\boldsymbol{w}$, which is the reason we call it linear.

- [2] (##) Consider a linear regression problem

  $$\begin{aligned} p(y \,|\, \mathbf{X}, w, \beta) &= \mathcal{N}(y \,|\, \mathbf{X}w, \beta^{-1}\mathbf{I}) \\ &= \prod_n \mathcal{N}(y_n \,|\, w^T x_n, \beta^{-1}) \end{aligned}$$

  with $\boldsymbol{y}, \boldsymbol{X}$ and $\boldsymbol{w}$ as defined in the notebook.
  (a) Work out the maximum likelihood solution for linear regression by solving

  $$\nabla_w \log p(y|X, w) = 0\,.$$

  (b) Work out the MAP solution. How does it relate to the ML solution?

  (a) The gradient of the log-likelihood is

  $$\nabla_w \log p(y|X, w) = X^T(y - Xw)$$

  Setting the derivation to zero leads to

  $$w_{ML} = (X^TX)^{-1}X^Ty$$

  (b) We now add a prior $w \sim \mathcal{N}(0, \alpha^{-1})$, and a similar derivation leads to

  $$\nabla_w \log p(y, w|X) = -\beta X^T(y - Xw) + \alpha w$$

  Setting the derivation to zero leads to

  $$w_{MAP} = (X^TX + \frac{\alpha}{\beta}I)^{-1}X^Ty$$

  The MAP solution weighs both the prior and likelihood. If $\frac{\alpha}{\beta}$ is close to zero (if the prior is uninformative), then the ML solution and MAP solutions are close to each other.

- [3] (###) Show that the variance of the predictive distribution for linear regression decreases as more data becomes available.

Variance of the predictive distribution is given by

$$\begin{aligned}\sigma^2_{N+1}(x) &= 1/\beta + \phi(x)^T S_{N+1}\phi(x)\\ S_{N+1} &= (S_N^{-1} + \beta\phi_{N+1}\phi_{N+1}^T)^{-1}\\ &= S_N - \frac{\beta S_N\phi_{N+1}\phi_{N+1}^T S_N}{1+\beta\phi_{N+1}^T S_N\phi_{N+1}}\end{aligned}$$

where in the last equality, we applied Woodbury's matrix identity, which is also listed in Sam Roweis' matrix notes, eq. 10. Using the recursive relation for S_N we can write the variance for the next observation as

$$\sigma^2_{N+1}(x) = \sigma^2_N(x) - \frac{\beta\phi(x)^T S_N\phi_{N+1}\phi_{N+1}^T S_N\phi(x)}{1+\beta\phi_{N+1}^T S_N\phi_{N+1}}.$$

Because S_N is positive definite, the numerator and the denominator of the second term wil be non-negative, hence $\sigma^2_N(x) \geqslant \sigma^2_{N+1}(x)$. This shows that the predictive variance decrease as more data becomes available.

- [4] (#) Assume a given data set $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$ with $x \in \mathbb{R}^M$ and $y \in \mathbb{R}$. We propose a model given by the following data generating distribution and weight prior functions:

$$p(y_n|x_n, w)\cdot p(w)\,.$$

(a) Write down Bayes rule for generating the posterior $p(w|D)$ from a prior and likelihood.
(b) Work out how to compute a distribution for the predicted value $y_\bullet$, given a new input $x_\bullet$.

(a)

$$p(w|D) = \frac{p(w)\prod_{n=1}^N p(y_n|x_n, w)}{\int p(w)\prod_{n=1}^N p(y_n|x_n, w)\mathrm{d}w}$$

(b)

$$p(y_\bullet|x_\bullet, D) = \int p(y_\bullet|x_\bullet, w)p(w|D)\mathrm{d}w$$

- [5] (#) In the class we use the following prior for the weights:

$$p(w|\alpha) = \mathcal{N}\left(w|0, \alpha^{-1}I\right)$$

(a) Give some considerations for choosing a Gaussian prior for the weights.
(b) We could have chosen a prior with full (not diagonal) covariance matrix $p(w|\alpha) = \mathcal{N}\left(w|0, \Sigma\right)$. Would that be better? Give your thoughts on that issue.
(c) Generally we choose α as a small positive number. Give your thoughts on that choice as opposed to choosing a large positive value. How about choosing a negative value for α?

(a) These considerations can be both computational (eg, Gaussian prior times Gaussian likelihood leads to a Gaussian posterior) or based on available information (eg, among all distributions with the same variance, the Gaussian distribution has the largest entropy. Roughly this means that the Gaussian makes the least amount of assumptions across all distributions with the same variance).
(b) If you have no prior information about co-variances, why make that assumption? If you do have some prior information, eg based on the physical process, then by all means feel free to add those constraints to the prior. Note that the posterior variance is given by $S_N = \left(\alpha\mathbf{I} + \beta\mathbf{X}^T\mathbf{X}\right)^{-1}$. Importantly, the term $\alpha\mathbf{I}$ for small α makes sure that the matrix is invertible, even for zero observations.
(c) As you can see from the posterior variance (see answer to (b)), for smaller values of α, the data term $\mathbf{X}^T\mathbf{X}$ gets to play a role after fewer observations. Hence, if you have little prior information, it's better to choose a small value for α.

- [6] Consider an IID data set $D = \{(x_1, y_1), \ldots, (x_N, y_N)\}$. We will model this data set by a model

$$y_n = \theta^T f(x_n) + e_n\,,$$

where $f(x_n)$ is an M-dimensional feature vector of input x_n; y_n is a scalar output and $e_n \sim \mathcal{N}(0, \sigma^2)$.

(a) Rewrite the model in matrix form by lumping input features in a matrix $F = [f(x_1), \ldots, f(x_N)]^T$, outputs and noise in the vectors $y = [y_1, \ldots, y_N]^T$ and $e = [e_1, \ldots, e_N]^T$, respectively.

$$y = F\theta + e$$

(b) Now derive an expression for the log-likelihood $\log p(y|\, F, \theta, \sigma^2)$.

$$\begin{aligned}\log p(D|\theta, \sigma^2) &= \log \mathcal{N}(y|\, F\theta, \sigma^2) \\ &\propto -\frac{1}{2\sigma^2}(y - F\theta)^T (y - F\theta)\end{aligned}$$

(c) Proof that the maximum likelihood estimate for the parameters is given by

$$\hat{\theta}_{\text{ml}} = (F^T F)^{-1} F^T y$$

.

Taking the derivative to θ

$$\nabla_\theta \log p(D|\theta) = \frac{1}{\sigma^2} F^T (y - F\theta)$$

Set derivative to zero for maximum likelihood estimate

$$\hat{\theta}_{\text{ml}} = (F^T F)^{-1} F^T y$$

(d) What is the predicted output value $y_\bullet$, given an observation $x_\bullet$ and the maximum likelihood parameters $\hat{\theta}_{\text{ml}}$. Work this expression out in terms of F, y and $f(x_\bullet)$.

Prediction of new data point: $\hat{y}_\bullet = \hat{\theta}^T f(x_\bullet) = \left((F^T F)^{-1} F^T y\right)^T f(x_\bullet)$

(e) Suppose that, before the data set D was observed, we had reason to assume a prior distribution $p(\theta) = \mathcal{N}(0, \sigma_0^2)$. Derive the Maximum a posteriori (MAP) estimate $\hat{\theta}_{\text{map}}$.(hint: work this out in the $\log$ domain.)

$$\begin{aligned}\log p(\theta|D) &\propto \log p(D|\theta) p(\theta) \\ &\propto -\frac{1}{2\sigma^2}(y - F\theta)^T (y - F\theta) + \frac{1}{2\sigma_0^2}\theta^T\theta\end{aligned}$$

Derivative $\nabla_\theta \log p(\theta|D) = -(1/\sigma^2) F^T (y - F\theta) + (1/\sigma_0^2)\theta$

Set derivative to zero for MAP estimate leads to

$$\hat{\theta}_{\text{map}} = \left(F^T F + \frac{\sigma^2}{\sigma_0^2} I\right)^{-1} F^T y$$

Generative Classification

- [1] You have a machine that measures property $\boldsymbol{x}$, the "orangeness" of liquids. You wish to discriminate between $C_1 = \text{`Fanta'}$ and $C_2 = \text{`Orangina'}$. It is known that

$$p(x|C_1) = \begin{cases} 10 & 1.0 \leq x \leq 1.1 \\ 0 & \text{otherwise} \end{cases}$$
$$p(x|C_2) = \begin{cases} 200(x-1) & 1.0 \leq x \leq 1.1 \\ 0 & \text{otherwise} \end{cases}$$

The prior probabilities $p(C_1) = 0.6$ and $p(C_2) = 0.4$ are also known from experience.
(a) (##) A "Bayes Classifier" is given by

$$\text{Decision} = \begin{cases} C_1 & \text{if } p(C_1|x) > p(C_2|x) \\ C_2 & \text{otherwise} \end{cases}$$

Derive the optimal Bayes classifier.
(b) (###) The probability of making the wrong decision, given $\boldsymbol{x}$, is

$$p(\text{error}|x) = \begin{cases} p(C_1|x) & \text{if we decide } C_2 \\ p(C_2|x) & \text{if we decide } C_1 \end{cases}$$

Compute the total error probability $p(\text{error})$ for the Bayes classifier in this example.

> (a) We choose C_1 if $p(C_1|x)/p(C_2|x) > 1$. This condition can be worked out as
> $$\frac{p(C_1|x)}{p(C_2|x)} = \frac{p(x|C_1)p(C_1)}{p(x|C_2)p(C_2)} = \frac{10 \times 0.6}{200(x-1) \times 0.4} > 1$$
> which evaluates to choosing
> $$\begin{array}{ll} C_1 & \text{if } 1.0 \leq x < 1.075 \\ C_2 & \text{if } 1.075 \leq x \leq 1.1 \end{array}$$
> The probability that $\boldsymbol{x}$ falls outside the interval $[1.0, 1.1]$ is zero.
> (b) The total probability of error $p(\text{error}) = \int_x p(\text{error}|x)p(x)\mathrm{d}x$. We can work this out as

$$\begin{aligned} p(\text{error}) &= \int_x p(\text{error}|x)p(x)\mathrm{d}x \\ &= \int_{1.0}^{1.075} p(C_2|x)p(x)\mathrm{d}x + \int_{1.075}^{1.1} p(C_1|x)p(x)\mathrm{d}x \\ &= \int_{1.0}^{1.075} p(x|C_2)p(C_2)\mathrm{d}x + \int_{1.075}^{1.1} p(x|C_1)p(C_1)\mathrm{d}x \\ &= \int_{1.0}^{1.075} 0.4 \cdot 200(x-1)\mathrm{d}x + \int_{1.075}^{1.1} 0.6 \cdot 10\mathrm{d}x \\ &= 80 \cdot [x^2/2 - x]_{1.0}^{1.075} + 6 \cdot [x]_{1.075}^{1.1} \\ &= 0.225 + 0.15 \\ &= 0.375 \end{aligned}$$

- [2] (#) (see Bishop exercise 4.8): Using (4.57) and (4.58) (from Bishop's book), derive the result (4.65) for the posterior class probability in the two-class generative model with Gaussian densities, and verify the results (4.66) and (4.67) for the parameters $\boldsymbol{w}$ and $\boldsymbol{w0}$.

Substitute 4.64 into 4.58 to get

$$
\begin{aligned}
a &= \log\left(\frac{\frac{1}{(2\pi)^{D/2}}\cdot\frac{1}{|\Sigma|^{1/2}}\cdot\exp\left(-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\right)\cdot p(C_1)}{\frac{1}{(2\pi)^{D/2}}\cdot\frac{1}{|\Sigma|^{1/2}}\cdot\exp\left(-\frac{1}{2}(x-\mu_2)^T\Sigma^{-1}(x-\mu_2)\right)\cdot p(C_2)}\right) \\
&= \log\left(\exp\left(-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)+\frac{1}{2}(x-\mu_2)^T\Sigma^{-1}(x-\mu_2)\right)\right)+\log\frac{p(C_1)}{p(C_2)} \\
&= \ldots \\
&= (\mu_1-\mu_2)^T\Sigma^{-1}x-0.5\left(\mu_1^T\Sigma^{-1}\mu_1-\mu_2^T\Sigma^{-1}\mu_2\right)+\log\frac{p(C_1)}{p(C_2)}
\end{aligned}
$$

Substituting this into the right-most form of (4.57) we obtain (4.65), with w and $w0$ given by (4.66) and (4.67), respectively.

- [3] (###) (see Bishop exercise 4.9).

The Log-likelihood is given by

$$\log p(\{\phi_n,t,n\}|\{\pi_k\})=\sum_n\sum_k t_{nk}\left(\log p(\phi_n|C_k)+\log\pi_k\right)\ .$$

Using the method of Lagrange multipliers (Bishop app.E), we augment the log-likelihood with the constraint and obtain the Lagrangian

$$\log p(\{\phi_n,t_{nk}\}|\{\pi_k\})+\lambda\left(\sum_k\pi_k-1\right)\ .$$

In order to maximize, we set the derivative with respect to π_k equal to zero and obtain

$$
\begin{aligned}
\sum_n\frac{t_{nk}}{\pi_k}+\lambda&=0 \\
-\pi_k\lambda&=\sum_n t_{nk}=N_k \\
-\lambda\sum_k\pi_k&=\sum_k\sum_n t_{nk} \\
\lambda&=-N
\end{aligned}
$$

- [4] (##) (see Bishop exercise 4.10).

We can write the log-likelihood as

$$\log p(\{\phi_n,t_n\}|\{\pi_k\})\propto -0.5\sum_n\sum_k t_{nk}\left(\log|\Sigma|+(\phi_n-\mu_k)^T\Sigma^{-1}(\phi-\mu)\right)$$

The derivatives of the likelihood with respect to mean and shared covariance are respectively

$$
\begin{aligned}
\nabla_{\mu_k}\log p(\{\phi_n,t_n\}|\{\pi_k\})&=\sum_n\sum_k t_{nk}\Sigma^{-1}\left(\phi_n-\mu_k\right)=0 \\
\sum_n t_{nk}\left(\phi_n-\mu_k\right))&=0 \\
\mu_k&=\frac{1}{N_k}\sum_n t_{nk}\phi_n \\
\nabla_\Sigma\log p(\{\phi_n,t_n\}|\{\pi_k\})&=\sum_n\sum_k t_{nk}\left(\Sigma-(\phi_n-\mu_k)(\phi_n-\mu_k)^T\right)=0 \\
\sum_n\sum_k t_{nk}\Sigma&=\sum_n\sum_k t_{nk}(\phi_n-\mu_k)(\phi_n-\mu_k)^T \\
\Sigma&=\frac{1}{N}\sum_k\sum_n t_{nk}(\phi_n-\mu_k)(\phi_n-\mu_k)^T
\end{aligned}
$$

Discriminative Classification

- [1] Given a data set $D = \{(x_1, y_1), \ldots, (x_N, y_N)\}$, where $x_n \in \mathbb{R}^M$ and $y_n \in \{0, 1\}$. The probabilistic classification method known as logistic regression attempts to model these data as

$$p(y_n = 1|x_n) = \sigma(\theta^T x_n + b)$$

where $\sigma(x) = 1/(1 + e^{-x})$ is the logistic function. Let's introduce shorthand notation $\mu_n = \sigma(\theta^T x_n + b)$. So, for every input x_n, we have a model output μ_n and an actual data output y_n.
(a) Express $p(y_n|x_n)$ as a Bernoulli distribution in terms of μ_n and y_n.
(b) If furthermore is given that the data set is IID, show that the log-likelihood is given by

$$L(\theta) \triangleq \log p(D|\theta) = \sum_n \{y_n \log \mu_n + (1 - y_n) \log(1 - \mu_n)\}$$

(c) Prove that the derivative of the logistic function is given by

$$\sigma'(\xi) = \sigma(\xi) \cdot (1 - \sigma(\xi))$$

(d) Show that the derivative of the log-likelihood is

$$\nabla_\theta L(\theta) = \sum_{n=1}^{N} \left(y_n - \sigma(\theta^T x_n + b)\right) x_n$$

(e) Design a gradient-ascent algorithm for maximizing $L(\theta)$ with respect to θ.

(a) $p(y_n|x_n) = p(y_n = 1|x_n)^{y_n} p(y_n = 0|x_n)^{1-y_n} = \mu_n^{y_n}(1-\mu_n)^{1-y_n}$
(b) The log-likelihood is given by

$$\begin{aligned} L(\theta) = \log p(D|\theta) &= \sum_n \log p(y_n|x_n, \theta) \\ &= \sum_n \{y_n \log \mu_n + (1-y_n)\log(1-\mu_n)\} \end{aligned}$$

(c)

$$\begin{aligned} \frac{d}{dx}\left(\frac{1}{1+e^{-x}}\right) &= \frac{(1+e^{-x})\cdot 0 - (-e^{-x}\cdot 1)}{(1+e^{-x})^2} \\ &= \frac{e^{-x}}{(1+e^{-x})^2} = \frac{1}{1+e^{-x}} \cdot \frac{e^{-x}}{1+e^{-x}} \\ &= \sigma(x)\,(1-\sigma(x)) \end{aligned}$$

(d)

$$\begin{aligned} \nabla_\theta L(\theta) &= \sum_n \frac{\partial L}{\partial \mu_n} \cdot \frac{\partial \mu_n}{\partial(\theta^T x_n + b)} \cdot \frac{\partial(\theta^T x_n + b)}{\partial \theta} \\ &= \sum_n \left(\frac{y_n}{\mu_n} - \frac{1-y_n}{1-\mu_n}\right) \cdot \mu_n(1-\mu_n) \cdot x_n \\ &= \sum_n \frac{y_n - \mu_n}{\mu_n(1-\mu_n)} \cdot \mu_n(1-\mu_n) \cdot x_n \\ &= \sum_n (y_n - \mu_n) \cdot x_n \end{aligned}$$

(e)

$$\theta^{(t+1)} = \theta^{(t)} + \rho \sum_n (y_n - \mu_n^{(t)}) x_n$$

- [2] Describe shortly the similarities and differences between the discriminative and generative approach to classification.

Both aim to build an algorithm for $p(y|x)$ where y is a discrete class label and x is a vector of real (or possibly discretely valued) variables. In the discriminative approach, we propose a model $p(y|x,\theta)$ and use a training data set $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$ to infer good values for the parameters. For instance, in a maximum likelihood setting, we choose the parameters $\hat{\theta}$ that maximize $p(D|\theta)$. The classification algorithm is now given by

$$p(y|x) = p(y|x, \hat{\theta})\,.$$

In the generative approach, we also aim to design an algorithm $p(y|x)$ through a parametric model that is now given by $p(y, x|\theta) = p(x|y,\theta)p(y|\theta)$. Again, we use the data set to train the parameters, eg, $\hat{\theta} = \arg\max_\theta p(D|\theta)$, and the classification algorithm is now given by Bayes rule:

$$p(y|x) \propto p(x|y, \hat{\theta}) \cdot p(y|\hat{\theta})$$

- [3] (Bishop ex.4.7) (#) Show that the logistic sigmoid function $\sigma(a) = \frac{1}{1+\exp(-a)}$ satisfies the property $\sigma(-a) = 1 - \sigma(a)$ and that its inverse is given by $\sigma^{-1}(y) = \log\{y/(1-y)\}$.

$$1-\sigma(a) = 1 - \frac{1}{1+\exp(-a)} = \frac{1+\exp(-a)-1}{1+\exp(-a)}$$
$$= \frac{\exp(-a)}{1+\exp(-a)} = \frac{1}{\exp(a)+1} = \sigma(-a)$$

Regarding the inverse,

$$y = \sigma(a) = \frac{1}{1+\exp(-a)}$$
$$\Rightarrow \frac{1}{y} - 1 = \exp(-a)$$
$$\Rightarrow \log\left(\frac{1-y}{y}\right) = -a$$
$$\Rightarrow \log\left(\frac{y}{1-y}\right) = a = \sigma^{-1}(y)$$

- [4] (Bishop ex.4.16) (###) Consider a binary classification problem in which each observation x_n is known to belong to one of two classes, corresponding to $y_n = 0$ and $y_n = 1$. Suppose that the procedure for collecting training data is imperfect, so that training points are sometimes mislabelled. For every data point x_n, instead of having a value y_n for the class label, we have instead a value π_n representing the probability that $y_n = 1$. Given a probabilistic model $p(y_n = 1|x_n, \theta)$, write down the log-likelihood function appropriate to such a data set.

  If the values of the $\{y_n\}$ were known then each data point for which $y_n = 1$ would contribute $\log p(y_n = 1|x_n, \theta)$ to the log likelihood, and each point for which $y_n = 0$ would contribute $\log p(y_n = 0|x_n, \theta) = \log(1 - p(y_n = 1|x_n, \theta))$ to the log likelihood. A data point whose probability of having $y_n = 1$ is given by π_n will therefore contribute

  $$\pi_n \log p(y_n = 1|x_n, \theta) + (1 - \pi_n) \log p(y_n = 0|x_n, \theta)$$

  and so the overall log-likelihood given the data set is

  $$\sum_{n=1}^{N} \pi_n \log p(y_n = 1|x_n, \theta) + (1 - \pi_n) \log p(y_n = 0|x_n, \theta)$$

- [5] (###) Let X be a real valued random variable with probability density

  $$p_X(x) = \frac{e^{-x^2/2}}{\sqrt{2\pi}}, \quad \text{for all } x.$$

  Also Y is a real valued random variable with conditional density

  $$p_{Y|X}(y|x) = \frac{e^{-(y-x)^2/2}}{\sqrt{2\pi}}, \quad \text{for all } x \text{ and } y.$$

  (a) Give an (integral) expression for $p_Y(y)$. Do not try to evaluate the integral.
  (b) Approximate $p_Y(y)$ using the Laplace approximation. Give the detailed derivation, not just the answer. Hint: You may use the following results. Let

  $$g(x) = \frac{e^{-x^2/2}}{\sqrt{2\pi}}$$

  and

  $$h(x) = \frac{e^{-(y-x)^2/2}}{\sqrt{2\pi}}$$

  for some real value y. Then:

$$\frac{\partial}{\partial x}g(x) = -xg(x)$$
$$\frac{\partial^2}{\partial x^2}g(x) = (x^2-1)g(x)$$
$$\frac{\partial}{\partial x}h(x) = (y-x)h(x)$$
$$\frac{\partial^2}{\partial x^2}h(x) = ((y-x)^2-1)h(x)$$

(a)

$$p_Y(y) = \int_{-\infty}^{\infty} p_X(x)p_{Y|X}(y|x)\,dx = \int_{-\infty}^{\infty} \frac{e^{-\frac{1}{2}(x^2+(y-x)^2)}}{2\pi}\,dx$$

(b) Using the hint we determine the first derivative of

$$f(x) = g(x)h(x),$$
$$\frac{\partial}{\partial x}f(x) = \frac{\partial}{\partial x}g(x)\cdot h(x) = -xg(x)h(x) + g(x)(y-x)h(x) = (y-2x)f(x)$$

Setting this to zero gives

$$y - 2x = 0; \quad \text{so} \quad x = \frac{1}{2}y.$$
$$\frac{\partial}{\partial x}\ln f(x) = \frac{\frac{\partial}{\partial x}f(x)}{f(x)} = (y-2x)$$
$$\frac{\partial^2}{\partial x^2}\ln f(x) = \frac{\partial}{\partial x}(y-2x) = -2.$$

So, we find $A = 2$, see lecture notes, and thus

$$\begin{aligned} p_Y(y) &= \int_{-\infty}^{\infty} f(x)\,dx \approx f(\frac{y}{2})\sqrt{\frac{2\pi}{A}} \\ &= g(\frac{y}{2})h(\frac{y}{2})\sqrt{\frac{2\pi}{A}} \\ &= \frac{1}{\sqrt{2\pi\cdot 2}}e^{-y^2/4}. \end{aligned}$$

So Y is a Gaussian with mean $m = 0$ and variance $\sigma^2 = 2$.

Latent Variable Models and Variational Bayes

- [1] (##) For a Gaussian mixture model, given by generative equations

$$p(x,z) = \prod_{k=1}^{K} \underbrace{(\pi_k \cdot \mathcal{N}(x|\mu_k, \Sigma_k))}_{p(x,z_k=1)}^{z_k}$$

proof that the marginal distribution for observations $\boldsymbol{x_n}$ evaluates to

$$p(x) = \sum_{j=1}^{K} \pi_k \cdot \mathcal{N}(x|\mu_j, \Sigma_j)$$

> $$\begin{aligned} p(x) &= \sum_z p(x,z) \\ &= \sum_z \prod_{k=1}^{K} (\pi_k \cdot \mathcal{N}(x|\mu_k, \Sigma_k))^{z_k} \end{aligned}$$
>
> Exploiting the one-hot coding scheme for $\boldsymbol{z}$, we can re-write the RHS as
>
> $$\sum_{j=1}^{K} \prod_{k=1}^{K} (\pi_k \cdot \mathcal{N}(x|\mu_k, \Sigma_k))^{I_{kj}} = \sum_{j=1}^{K} \pi_j \cdot \mathcal{N}(x|\mu_j, \Sigma_j)$$
>
> where $I_{kj} = 1$ if $k = j$ and 0 otherwise.

- [2] (#) Given the free energy functional $F[q] = \sum_z q(z) \log \frac{q(z)}{p(x,z)}$, proof the EE, DE and AC decompositions.

> The Energy-Entropy decomposition follows simply from $\log \frac{a}{b} = \log(a) - \log(b)$. The Divergence-Evidence decomposition follows from $p(x,z) = p(z|x)p(x)$ and the Complexity-Accuracy decomposition follows from substituting $p(x,z) = p(x|z)p(z)$. Altogether leading to
>
> $$\begin{aligned} \mathrm{F}[q] &= \underbrace{-\sum_z q(z) \log p(x,z)}_{\text{energy}} \underbrace{-\sum_z q(z) \log \frac{1}{q(z)}}_{\text{entropy}} && \text{(EE)} \\ &= \underbrace{\sum_z q(z) \log \frac{q(z)}{p(z|x)}}_{\text{KL divergence} \geq 0} \underbrace{-\log p(x)}_{\text{log-evidence}} && \text{(DE)} \\ &= \underbrace{\sum_z q(z) \log \frac{q(z)}{p(z)}}_{\text{complexity}} \underbrace{-\sum_z q(z) \log p(x|z)}_{\text{accuracy}} && \text{(CA)} \end{aligned}$$

- [3] (#) The Free energy functional $\mathrm{F}[q] = -\sum_z q(z) \log p(x,z) - \sum_z q(z) \log \frac{1}{q(z)}$ decomposes into "Energy minus Entropy". So apparently the entropy of the posterior $q(z)$ is maximized. This entropy maximization may seem puzzling at first because inference should intuitively lead to more informed posteriors, i.e., posterior distributions whose entropy is smaller than the entropy of the prior. Explain why entropy maximization is still a reasonable objective.

> Note that Free Energy minimization is a balancing act: FE minimization implies entropy maximization and at the same time energy minimization. Minimizing the energy term leads to aligning $q(z)$ with $\log p(x, z)$, ie, it tries to move the bulk of the function $q(z)$ to areas in z-space where $p(x, z)$ is large ($p(x, z)$ is here just a function of z, since x is observed). However, aside from aligning with $p(x, z)$, we want $q(z)$ to be as uninformative as possible. Everything that can be inferred should be represented in $p(x, z)$ (which is prior times likelihood). We don't want to learn anything that is not in either the prior or the likelihood. The entropy term balances the energy term by favoring distributions that are as uninformative as possible.

- [4] (#) Explain the following update rule for the mean of the Gaussian cluster-conditional data distribution (from the example about mean-field updating of a Gaussian mixture model):

$$m_k = \frac{1}{\beta_k}(\beta_0 m_0 + N_k \bar{x}_k) \tag{B-10.61}$$

> We see here an example of "precision-weighted means add" when two sources of information are fused, just like precision-weighted means add when two Gaussians are multiplied, eg a prior and likelihood. In this case, the prior is m_0 and the likelihood estimate is $\bar{x}$. β_0 can be interpreted as the number of pseudo-observations in the prior.

- [5] (##) Consider a model $p(x, z|\theta)$, where $D = \{x_1, x_2, \ldots, x_N\}$ is observed, z are unobserved variables and θ are parameters. The EM algorithm estimates the parameters by iterating over the following two equations (i is the iteration index):

$$\begin{aligned} q^{(i)}(z) &= p(z|D, \theta^{(i-1)}) \\ \theta^{(i)} &= \arg\max_\theta \sum_z q^{(i)}(z) \cdot \log p(D, z|\theta) \end{aligned}$$

Proof that this algorithm minimizes the Free Energy functional

$$F[q, \theta] = \sum_z q(z) \log \frac{q(z)}{p(D, z|\theta)}$$

> Let's start with a prior estimate $\theta^{(i-1)}$ and we want to minimize the free energy functional wrt q. This leads to
>
> $$\begin{aligned} q^{(i)}(z) &= \arg\min_q F[q, \theta^{(i-1)}] \\ &= \arg\min_q \sum_z q(z) \log \frac{q(z)}{p(D, z|\theta^{(i-1)})} \\ &= \arg\min_q \sum_z q(z) \log \frac{q(z)}{p(z|D, \theta^{(i-1)}) \cdot p(D|\theta^{(i-1)})} \\ &= p(z|D, \theta^{(i-1)}) \end{aligned}$$
>
> Next, we use $q^{(i)}(z) = p(z|D, \theta^{(i-1)})$ and minimize the free energy w.r.t. θ, leading to
>
> $$\begin{aligned} \theta^{(i)} &= \arg\min_\theta F[q^{(i)}(z), \theta] \\ &= \arg\min_\theta \sum_z p(z|D, \theta^{(i-1)}) \log \frac{p(z|D, \theta^{(i-1)})}{p(D, z|\theta)} \\ &= \arg\max_\theta \sum_z \underbrace{p(z|D, \theta^{(i-1)})}_{q^{(i)}(z)} \log p(D, z|\theta) \end{aligned}$$

- [6] (###) Consult the internet on what overfitting and underfitting is and then explain how FE minimization finds a balance between these two (unwanted) extremes.

> Overfitting relates to learning a posterior that "listens" too much to the data (and not enough to the prior). Underfitting does the opposite. The CA decompostion
>
> $$\underbrace{\sum_z q(z) \log \frac{q(z)}{p(z)}}_{\text{complexity}} - \underbrace{\sum_z q(z) \log p(x|z)}_{\text{accuracy}} \tag{CA}$$
>
> exposes this dilemma nicely. The complrxity term tries to keep the posterior $q(z)$ near the prior $p(z)$ whereas the accuracy term tries to align the posterior $q(z)$ with the likelihood $p(x|z)$. Thus, minimizing Free Energy keeps an eye on avoiding both under- and over-fitting.

- [7] (##) Consider a model $p(x, z|\theta) = p(x|z, \theta)p(z|\theta)$ where x and z relate to observed and unobserved variables, respectively. Also available is an observed data set $D = \{x_1, x_2, \ldots, x_N\}$. One iteration of the EM-algorithm for estimating the parameters θ is described by (m is the iteration counter)

$$\hat{\theta}^{(m+1)} := \arg\max_\theta \left(\sum_z p(z|x = D, \hat{\theta}^{(m)}) \log p(x = D, z|\theta) \right) .$$

  (a) Apparently, in order to execute EM, we need to work out an expression for the 'responsibility' $p(z|x = D, \hat{\theta}^{(m)})$. Use Bayes rule to show how we can compute the responsibility that allows us to execute an EM step.

  > Use Bayes rule:
  >
  > $$p(z|x = D, \hat{\theta}^{(m)}) = \frac{p(x = D|z, \hat{\theta}^{(m)})\, p(z|\hat{\theta}^{(m)})}{\int p(x = D|z, \hat{\theta}^{(m)})\, p(z|\hat{\theta}^{(m)})\, \mathrm{d}z}$$
  >
  > Note that the RHS is an expression in z since D and $\hat{\theta}$ are given. If you want to evaluate the RHS, you need to make a specific choice for your model
  >
  > $$p(x, z|\theta) = \underbrace{p(x|z, \theta)}_{\text{likelihood}}\, \underbrace{p(z|\theta)}_{\text{prior}}$$

  (b) Why do we need multiple iterations in the EM algorithm?

  > We must have a parameter estimate in order to compute the responsibilities, and vice versa, we need responsibilities to update the parameter estimate. Thus, in the EM algorithm, we iterate between updating responsibilities (beliefs about z) and parameter estimates (beliefs about θ).

  (c) Why can't we just use simple maximum log-likelihood to estimate parameters, as described by

$$\hat{\theta} := \arg\max_\theta \log p(x = D, z|\theta) \, ?$$

  > Because z is not observed.

- [8] In a particular model with hidden variables, the log-likelihood can be worked out to the following expression:

$$L(\theta) = \sum_n \log \left(\sum_k \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k) \right)$$

  Do you prefer a gradient descent or EM algorithm to estimate maximum likelihood values for the parameters? Explain your answer. (No need to work out the equations.)

Since this expression does not degenerate into simple MVGs, the EM approach is in practice preferred.

Intelligent Agents and Active Inference

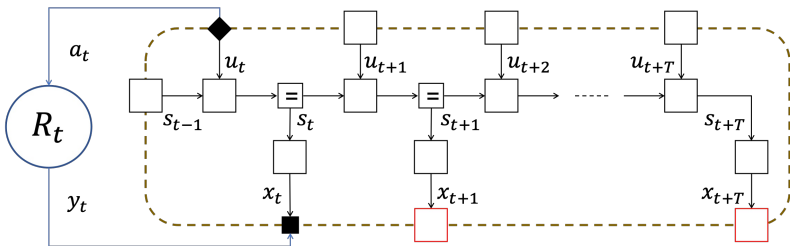
- [1] (##) I asked you to watch a video segment (https://www.vibby.com/watch?vib=71iPtUJxd) where Karl Friston talks about two main approaches to goal-directed acting by agents: (1) choosing actions that maximize (the expectation of) a value function $V(s)$ of the state (s) of the environment; or (2) choosing actions that minimize a functional ($F[q(s)]$) of beliefs ($q(s)$) over environmental states (s). Discuss the advantage of the latter appraoch.

  > We'll discuss two advantages here. Either one would suffice for full credit and there are likely multiple alternative answers that would be adequate as well. (1) One advantage is that the value function V needs to be uniquely chosen for each problem. Brains cannot afford to come up with a new value function for each problem as thousands of new problems are encountered each day. In contrast, $F[q(s)]$ holds the free-energy functional (a given cost functional) for posterior beliefs that technically are defined by a generative model p and Bayes rule. In other words, in the latter approach, there is one value (cost) function for all problems. (2) A second advantage of $F[q(s)]$ is that inference for actions can take into account the uncertainty about our state-of-knowledge of the environment. This may lead to actions that are information seeking rather than goal-driven if our belief are very uncertain. For instance, if I want to cross a street, my first actions will be to seek information (look for cars and how fast they go), and only after enough information has been collected, the goal-driven action (decision) cross-vs-stay will be executed. When minimization of $F[q(s)]$ drives actions, both information-seeking and goal-driven actions can be accomodated in the same framework. This is very difficult when the value function is a direct function of the state of the world ($V(s)$, because there is no accomodation to represent our uncertainties about the state of the world.

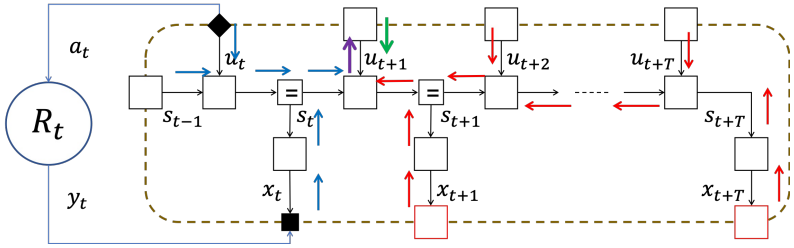
- [2] (#) The good regulator theorem states that a "successful and efficient" controller of the world must contain a model of the world. But it's hard to imagine how just learning a model of the world leads to goal-directed behavior, like learning how to read or drive a car. Which other ingredient do we need to get learning agents to behave as goal-directed agents?

  > In the Free Energy Principle framework, goals (targets) are encoded in a generative model of the environment as prior distributions on future observations. Actions are inferred through free energy minimization in this extended model. As a result, the inferred actions aim to generate future observations that are maximally consistent with the goal priors. This kind of behavior can be interpreted as goal-directed behavior.

- [3] (##) The figure below reflects the state of a factor graph realization of an active inference agent after having pushed action a_t onto the environment and having received observation x_t. In this graph, the variables $x_\bullet$, $u_\bullet$ and $s_\bullet$ correspond to observations, and unobserved control and internal states respectively. Copy the figure onto your sheet and draw a message passing schedule to infer a posterior belief (i.e. after observing x_t) over the next control state u_{t+1}.



Imagine picking up the tree at the u_{t+1} edge (call this edge the root of the tree). Then pass messages from the leaves of the tree towards the root, see Figure below. Note that the posterior belief over next control (action) u_{t+1} incorporates information from the recent past (blue messages), from prior information about what worked in the past (green arrow), and from expectations about future observations (red messages).



- [4] (##) The Free Energy Principle (FEP) is a theory about biological self-organization, in particular about how brains develop through interactions with their environment. Which of the following statements is not consistent with FEP (and explain your answer):
  (a) We act to fullfil our predictions about future sensory inputs.
  (b) Perception is inference about the environmental causes of our sensations.
  (c) Our actions aim to reduce the complexity of our model of the environment.

  Statement (c) is not consistent with the FEP formulation of biological self-organization. The Complexity-Accuracy decomposition of the Free Energy reveals that the "data" (observations) is exclusively part of the accuracy term (not in the complexity term). Observations are controlled by actions and hence actions aim to minimize accuracy rather than model complexity.

Dynamic Models

- [1] (##) Given the Markov property

$$p(x_n|x_{n-1}, x_{n-2}, \ldots, x_1) = p(x_n|x_{n-1}) \tag{A1}$$

proof that, for any n,

$$p(x_n, x_{n-1}, \ldots, x_{k+1}, x_{k-1}, \ldots, x_1|x_k) = p(x_n, x_{n-1}, \ldots, x_{k+1}|x_k) \cdot p(x_{k-1}, x_{k-2}, \ldots, x_1|x_k)\,. \tag{A2}$$

In other words, proof that, if the Markov property A1 holds, then, given the "present" (x_k), the "future" $(x_n, x_{n-1}, \ldots, x_{k+1})$ is independent of the "past" $(x_{k-1}, x_{k-2}, \ldots, x_1)$.

> First, we rewrite A2 as
>
> $$\begin{aligned} p(x_n, x_{n-1}, \ldots, x_{k+1}, x_{k-1}, \ldots, x_1|x_k) &= \frac{p(x_n, x_{n-1}, \ldots, x_1)}{p(x_k)} \\ &= \frac{p(x_n, x_{n-1}, \ldots, x_{k+1}|x_k, \ldots, x_1) \cdot p(x_k, x_{k-1}, \ldots, x_1)}{p(x_k)} \\ &= p(x_n, x_{n-1}, \ldots, x_{k+1}|x_k, \ldots, x_1) \cdot p(x_{k-1}, \ldots, x_1|x_k) \end{aligned} \tag{A3}$$
>
> The first term in A3 can be simplified if A1 holds to
>
> $$\begin{aligned} p(x_n, &x_{n-1}, \ldots, x_{k+1}|x_k, x_{k-1}, \ldots, x_1) \\ &= p(x_n|x_{n-1}, x_{n-2}, \ldots, x_1) \cdot p(x_{n-1}|x_{n-2}, x_{n-3}, \ldots, x_1) \cdots \\ &\quad \cdots p(x_{k+1}|x_k, x_{k-2}, \ldots, x_1) \\ &= p(x_n|x_{n-1}, x_{n-2}, \ldots, x_k) \cdot p(x_{n-1}|x_{n-2}, x_{n-3}, \ldots, x_k) \cdots \\ &\quad \cdots p(x_{k+1}|x_k) \\ &= p(x_n, x_{n-1}, \ldots, x_{k+1}|x_k) \end{aligned} \tag{A4}$$
>
> Substitution of A4 into A3 leads to A2. QED.

- [2] (#)
  (a) What's the difference between a hidden Markov model and a linear Dynamical system?

  > HMM has binary-valued (on-off) states, where the LDS has continuously valued states.

  (b) For the same number of state variables, which of these two models has a larger memory capacity, and why?

  > The latter holds more capacity because, eg, a 16-bit representation of a continuously-valued variable holds 2^{16} different states.

- [3] (#) (a) What is the 1st-order Markov assumption?
  (b) Derive the joint probability distribution $p(x_{1:T}, z_{0:T})$ (where x_t and z_t are observed and latent variables respectively) for the state-space model with transition and observation models $p(z_t|z_{t-1})$ and $p(x_t|z_t)$.
  (c) What is a Hidden Markov Model (HMM)?
  (d) What is a Linear Dynamical System (LDS)?
  (e) What is a Kalman Filter?
  (f) How does the Kalman Filter relate to the LDS?
  (g) Explain the popularity of Kalman filtering and HMMs?
  (h) How relates a HMM to a GMM?

(a) An auto-regressive model is first-order Markov if

$$p(x_t|x_{t-1}, x_{t-2}, \ldots, x_1) = p(x_t|x_{t-1}) \,.$$

(b)

$$p(x_{1:T}, z_{0:T}) = p(z_0) \prod_{t=1}^{T} p(z_t|z_{t-1}) \prod_{t=1}^{T} p(x_t|z_t)$$

(c) A HMM is a state-space model (as described in (b)) where the latent variable z_t is discretely valued. Iow, the HMM has hidden clusters.
(d) An LDS is a state-space model (also described by the eq in (b)), but now the latent variable z_t is continuously valued.
(e) A Kalman filter is a recursive solution to the inference problem $p(z_t|x_t, x_{t-1}, \ldots, x_1)$, based on a state estimate at the previous time step $p(z_{t-1}|x_{t-1}, x_{t-2}, \ldots, x_1)$ and a new observation x_t. Basically, it's a recursive filter that updates the optimal Bayesian estimate of the current state z_t based on all past observations $x_t, x_{t-1}, \ldots, x_1$.
(f) The LDS describes a (generative) model. The Kalman filter does not describe a model, but rather describes an inference task on the LDS model.
(g) The LDS and HMM models are both quite general and flexible generative probabilistic models for time series. There exists very efficient algorithms for executing the latent state inference tasks (Kalman filter for LDS and there is a similar algorithm for the HMM). That makes these models flexible and practical. Hence the popularity of these models.
(h) An HMM can be interpreted as a Gaussian-Mixture-model-over-time.