BML: exercise sheet

Rémi Bardenet, Julyan Arbel

Stars indicate the difficulty level, from 1 to 3. One star means that everyone should be able to do it without too much effort.

Contents

1 Bayesics

1.1 Conjugate priors 101: Gaussians ($\star$)

Let $y|\mu \sim \mathcal{N}(\mu, I_N)$ and $\mu \sim \mathcal{N}(0, aI_N)$, for some $a > 0$. Show that

$$\mu|y \sim \mathcal{N}(by, bI_N), \text{ where } b = a/(a+1). \tag{1}$$

> **Solution:** We apply Bayes' theorem and keep track of only the terms that will not end up in the normalization constant of the posterior. This gives
>
> $$\log p(\mu|y) \propto \log p(y|\mu) + \log p(\mu)$$
> $$\propto -\frac{\|y-\mu\|^2}{2} - \frac{\|\mu\|^2}{2a}$$
> $$\propto -\frac{1}{2}\|\mu\|^2\left(1 + \frac{1}{a}\right) + y^T\mu$$
> $$\propto -\frac{\|\mu - by\|^2}{2b}.$$

1.2 A conjugate prior on probability vectors ($\star$)

Let

$$\Delta_d = \{\theta \in [0,1]^d \text{ such that } \sum_{k=1}^{d} \theta_d = 1\}.$$

Let further $\alpha \in (\mathbb{R}_+)^d$. The Dirichlet pdf is defined by

$$\mathrm{Dir}(\theta|\alpha) = \frac{1}{B(\alpha)} \prod_{k=1}^{d} \theta_k^{\alpha_k - 1} 1_{\theta \in \Delta_d},$$

where $B(\alpha) = \prod_{k=1}^{d} \Gamma(\alpha_k)/\Gamma(\sum_{k=1}^{d} \alpha_k)$ is the so-called beta function.

Now put a prior $\mathrm{Dir}(\theta|\alpha)$ on θ, and consider drawing $y_{1:N}$ from the multinomial distribution with parameter $\theta \in \Delta_d$. Show that

$$p(\theta, y_{1:N}) = \frac{B(\alpha + c)}{B(\alpha)} \mathrm{Dir}(\theta|\alpha + c), \tag{2}$$

where $c = (\sum_{i=1}^{N} 1_{y_i = k})_{1 \leqslant k \leqslant d}$ is the vector of counts. Note that (2) implies that $\theta|y_{1:N} \sim \mathrm{Dir}(\theta|\alpha)$ and that the marginal likelihood $p(y_{1:n}) = B(\alpha + c)/B(\alpha)$.

Solution: Once you express the multinomial pdf, the Dirichlet distribution becomes the obvious conjugate prior. This time, we keep track of the normalizing constant, because the script requires it. This gives

$$p(\theta, y_{1:N}) = p(y_{1:N}|\theta)p(\theta)$$

$$= \prod_{i=1}^{N}\prod_{k=1}^{d}\theta_k^{1\{y_i=k\}} \times \frac{1}{B(\alpha)}\prod_{k=1}^{d}\theta_k^{\alpha_k-1}1_{\theta\in\Delta_d}$$

$$= \frac{1}{B(\alpha)}\prod_{k=1}^{d}\theta_k^{\alpha_k+c_k-1}1_{\theta\in\Delta_d}$$

$$= \frac{B(\alpha+c)}{B(\alpha)}\mathrm{Dir}(\theta|\alpha+c).$$

1.3 Empirical Bayes and the James-Stein effect ($\star\star$)

Let $\mu = (\mu_1, \ldots, \mu_N) \in \mathbb{R}^N$, and consider N i.i.d. real variables $y_i|\mu \sim \mathcal{N}(\mu_i, 1)$. We wish to infer μ.

1. What is the maximum likelihood estimator $\hat{\mu}_{\mathrm{MLE}}$?

2. Henceforth, we judge estimators by the square loss. The frequentist risk of an estimator $\hat{\mu}$ is
$$R(\hat{\mu}) = \mathbb{E}_{y|\mu}\|\mu - \hat{\mu}\|^2.$$
show that $R(\hat{\mu}_{\mathrm{MLE}}) = N$.

3. Suppose we have prior belief that μ lies near 0, and we choose to represent it by $\mu \sim \mathcal{N}(0, aI_N)$, $a > 0$. What is the Bayes estimator $\hat{\mu}_{\mathrm{Bayes}}$? What is its (frequentist) risk $R(\hat{\mu}_{\mathrm{Bayes}})$? What is its Bayes risk $\mathbb{E}_{\mu}R(\hat{\mu}_{\mathrm{Bayes}})$?

4. Since we actually have no idea what a should be, we propose to estimate it from data.[1] Show that the marginal of y is

$$\int p(y, \mu)\mathrm{d}\mu = \mathcal{N}(y|0, (a+1)I_N).$$

In particular, what is the law of $S = \|y\|^2$? Deduce from it that $(N-2)/S$ is an unbiased estimator of $1/(a+1)$, and consider the empirical Bayes estimator
$$\hat{\mu}_{\mathrm{EB}} = \left(1 - \frac{N-2}{S}\right)y.$$

---

[1]This procedure of using data to tune the prior is called *empirical Bayes* (EB). The expected utility principle allows it, but statisticians who like to interpret their prior as encoding their belief before the data is collected are uncomfortable with EB. At the other extreme, Bayesians who insist on using estimators with good frequentist properties are happy using the data or the likelihood to design their prior.

Note that this is just $\hat{\mu}_{\text{Bayes}}$, but with $1/(a+1)$ replaced by an unbiased estimator. What is the Bayes risk of $\hat{\mu}_{\text{EB}}$?

5. *Note: This particular item is $(\star\star\star)$ because it is longer to solve, but all individual arguments are elementary; do this only if you have solved all the preceding exercises, though.* Show that for $N \geqslant 3$, for every $\mu \in \mathbb{R}^N$,

$$R(\hat{\mu}_{\text{EB}}) < R(\hat{\mu}_{\text{MLE}}). \tag{3}$$

Frequentists say that $\hat{\mu}_{\text{EB}}$ dominates μ_{MLE}, in the sense that whatever the value of μ, the risk of $\hat{\mu}_{\text{EB}}$ is the smallest of the two. This happens even when μ is far from zero, in which case one might have thought that our $\mathcal{N}(0, aI_N)$ prior would have been a poor choice. Finally, if you are a strict Waldian, you should thus prefer $\hat{\mu}_{\text{EB}}$ to $\hat{\mu}_{\text{MLE}}$. Many applied frequentists still use $\hat{\mu}_{\text{MLE}}$, however; see (Efron, 2012, Section 1.3) for a tentative answer.

Equation 3 is called the James-Stein effect, and is a standard example of why following Bayesian guidelines can end up giving good frequentist estimators. Shrinkage, like $\hat{\mu}_{\text{EB}}$ shrinks $\hat{\mu}_{\text{MLE}}$ towards zero, is now commonplace in large-dimensional regression. For more on frequentist guarantees for Bayesian estimators and shrinkage, see (Parmigiani and Inoue, 2009, Sections 7, 8, 9).

Solution: The solution is basically Efron, 2012, Section 1.2, and we give some details below. The book is also highly recommended, especially if you are into large-scale hypothesis tests. At least, read the prologue for statistical culture.

1. By definition,

$$\hat{\mu}_{\text{MLE}} \in \arg\max_{\mu} \mathcal{N}(y|\mu, I_N) = \arg\min_{\mu} \|y - \mu\|^2 = y.$$

2. Since $y|\mu \sim \mathcal{N}(\mu, I_N)$, the risk of $\hat{\mu}_{\text{MLE}}$ is

$$R(\hat{\mu}_{\text{MLE}}) = \mathbb{E}_{y|\mu} \|y - \mu\|^2 = \sum_{i=1}^{N} \mathbb{E}_{y_i|\mu_i} (y_i - \mu_i)^2 = N.$$

3. Because the loss is the squared loss, the Bayes estimator is the posterior mean. By Exercise 1.1, this is $\hat{\mu}_{\text{Bayes}} = \frac{a}{a+1} y$. Its frequentist

risk is

$$R(\hat{\mu}_{\text{Bayes}}) = \mathbb{E}_{y|\mu} \left\| \mu - \frac{a}{a+1}y \right\|^2$$

$$= \|\mu\|^2 - \frac{2a}{a+1}\mu^T \mathbb{E}_{y|\mu}y + \left(\frac{a}{a+1}\right)^2 \mathbb{E}_{y|\mu}\|y\|^2$$

$$= \frac{1-a}{a+1}\|\mu\|^2 + \left(\frac{a}{a+1}\right)^2 (N + \|\mu\|^2).$$

$$= \frac{1}{(a+1)^2}\|\mu\|^2 + \left(\frac{a}{a+1}\right)^2 N.$$

Denoting by $b = a/(a+1)$, it comes

$$R(\hat{\mu}_{\text{Bayes}}) = (1-b)^2\|\mu\|^2 + b^2 N. \tag{4}$$

Finally, upon noting that $a = \frac{b}{1-b}$, the Bayes risk is

$$\mathbb{E}_\mu R(\hat{\mu}_{\text{Bayes}}) = (1-b)^2 aN + b^2 N = ((1-b)b + b^2)N = bN.$$

Note that, as expected, this is smaller than the constant Bayes risk $\mathbb{E}_\mu R(\hat{\mu}_{\text{MLE}}) = N$ of the MLE.

4. This is the same computation as for Exercise 1.1, but this time we keep the terms in y. More precisely,

$$\log p(\mu, y) = \log p(y|\mu) + \log p(\mu)$$

$$\propto -\frac{\|y-\mu\|^2}{2} - \frac{\|\mu\|^2}{2a}$$

$$\propto -\frac{1}{2}\|\mu\|^2 \left(1 + \frac{1}{a}\right) + y^T\mu - \frac{1}{2}\|y\|^2$$

$$\propto -\frac{1}{2} \begin{pmatrix} \mu & y \end{pmatrix} \begin{pmatrix} (1+\frac{1}{a})I_N & I_N \\ I_N & I_N \end{pmatrix} \begin{pmatrix} \mu \\ y \end{pmatrix}.$$

We recognize a Gaussian in (μ, y), with mean zero and covariance matrix

$$\begin{pmatrix} (1+\frac{1}{a})I_N & I_N \\ I_N & I_N \end{pmatrix}^{-1} = a \begin{pmatrix} I_N & -I_N \\ -I_N & (1+\frac{1}{a})I_N \end{pmatrix}.$$

The marginal of y is thus a Gaussian with mean zero and variance

$$a\left(1 + \frac{1}{a}\right) I_N = (1+a)I_N.$$

5

In particular, $S = \|y\|^2 = \sum_{i=1}^N y_i^2$ has the same distribution as $(1+a)X$, where $X \sim \chi_N^2$. Letting $N > 2$, $(N-2)/S$ thus has the same law as $(N-2)/(1+a)$ times an inverse chi-squared variable with N degrees of freedom. The latter has mean $1/(N-2)$, so that $(N-2)/S$ is an unbiased estimator of $1/(1+a)$. Finally, the Bayes risk of $\hat\mu_{\mathrm{EB}}$ is

$$\mathbb{E}_\mu R(\hat\mu_{\mathrm{EB}}) = \mathbb{E}_\mu \mathbb{E}_{y|\mu}\left(1 - \frac{N-2}{S}y\right) = \mathbb{E}_\mu(\mu())$$

1.4   Classification with asymmetric loss ($\star$)

Consider the classification problem, but with loss

$$L(a_g, s) = \alpha 1_{y \neq g(x; x_{1:n}, y_{1:n})} 1_{y=0} + \beta 1_{y \neq g(x; x_{1:n}, y_{1:n})} 1_{y=1},$$

for some $\alpha, \beta > 0$. Show that the Bayes decision rule is

$$g^\star(x; x_{1:n}, y_{1:n}) = 1_{p(y|x, x_{1:n}, y_{1:n}) \geqslant \frac{\alpha}{\alpha+\beta}}.$$

In particular, if $\alpha \ll \beta$, one will often decide for predicting 1, because the cost for misclassifying a 0 is low.

Solution: For brevity, we drop the dependence of g in the training set and write $g(x)$ for $g(x; x_{1:n}, y_{1:n})$. Following the posterior expected loss rationale, we pick action

$$a^\star = a_{g^\star} \in \arg\min \int L(a_g, s) p(s_u | s_o) \mathrm{d}s_u$$

$$= \arg\min \int \left[\alpha 1_{y \neq g(x)} 1_{y=0} + \beta 1_{y \neq g(x)} 1_{y=1}\right] p(y | x_{1:N}, y_{1:N}, x) \mathrm{d}y$$

$$= \arg\min \alpha 1_{0 \neq g(x)} p(y = 0 | x_{1:N}, y_{1:N}, x)$$
$$+ \beta 1_{1 \neq g(x)} p(y = 1 | x_{1:N}, y_{1:N}, x).$$

This is equivalent to setting $g^\star(x) = 1$ if and only if

$$\alpha p(y = 0 | x_{1:N}, y_{1:N}, x) \leqslant \beta p(y = 1 | x_{1:N}, y_{1:N}, x).$$

Letting $q = p(y = 1 | x_{1:N}, y_{1:N}, x)$, this becomes

$$\alpha(1 - q) \leqslant \beta q,$$

or, equivalently,

$$q \geqslant \alpha/(\alpha + \beta).$$

1.5 Linear regression with a Gaussian prior ($\star$)

Consider $y_i|x_i, \theta \sim \mathcal{N}(x_i^T\theta, \sigma^2)$ i.i.d., $i = 1, \ldots, N$. Take a Gaussian prior $\theta \sim \mathcal{N}(0, \sigma_0^2)$. Show that the posterior $\theta|x_{1:N}, y_{1:N}$ is Gaussian, with mean the ridge regression estimator.

Solution: We write Bayes' theorem and keep track only of the terms that won't end up in the normalization constant. This gives

$$\log p(\theta|y_{1:N}, x_{1:N}) \propto \log p(y_{1:N}|x_{1:N}, \theta) + \log p(\theta)$$

$$\propto -\sum_{i=1}^{N} \frac{(y_i - x_i^T\theta)^2}{2\sigma^2} + \frac{1}{2\sigma_0^2}\|\theta\|^2$$

$$= -\frac{1}{2\sigma^2}\|y - X\theta\|^2 + \frac{1}{2\sigma_0^2}\|\theta\|^2$$

$$\propto -\frac{1}{2\sigma^2}\left[\theta^T\left(X^TX + \frac{\sigma^2}{\sigma_0^2}I_d\right)\theta - 2y^TX\theta\right]$$

$$= -\frac{1}{2}\left[\theta^T\Lambda\theta - \frac{2}{\sigma^2}y^TX\theta\right],$$

where $\Lambda := \frac{1}{\sigma^2}X^TX + \frac{1}{\sigma_0^2}I_d$ is symmetric and positive definite. This leads to

$$\log p(\theta|y_{1:N}, x_{1:N}) \propto -\frac{1}{2}\left(\theta - \frac{1}{\sigma^2}\Lambda^{-1}X^Ty\right)^T\Lambda\left(\theta - \frac{1}{\sigma^2}\Lambda^{-1}X^Ty\right),$$

so that $\theta|y_{1:N}, x_{1:N}$ is indeed Gaussian, with mean the ridge regression estimator

$$\frac{1}{\sigma^2}\Lambda^{-1}X^Ty = \left(X^TX + \frac{\sigma^2}{\sigma_0^2}I_d\right)^{-1}X^Ty$$

and variance Λ^{-1}. Note how the ratio σ/σ_0 is playing the role of the regularization parameter in ridge regression.

1.6 For more exercises on Bayesian derivations

- Exercises 5.1 to 5.4 of (Murphy, 2012).

- Go through Sections 4.4 to 4.6 of (Murphy, 2012) with pen and paper. Linear Gaussian models appear all the time.

- Exercises 2.6, 2.9, 2.10, 2.13, 2.14, and 2.15 of (Marin and Robert, 2007). Solutions are here.
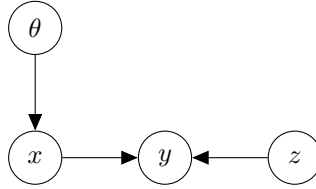
Figure 1: A DAG

2 MCMC

2.1 DAGs and dependence ($\star$)

Consider the DAG from Figure 1.

1. Write the corresponding factorization of $p(x, y, z, \theta)$.

2. Deduce from the factorization that $x \perp z$.

3. Deduce from the factorization that $x \perp z | \theta$.

4. Give an example of joint distribution that factorizes over the DAG, and such that $x \not\perp z | \theta, y$.

In particular, note how Item 3 is a case of *being independent from your non-descendents given your parents*, while Item 4 illustrates how conditioning on common children can induce dependence between parents. In more complicated DAGs, the so-called *Bayes ball* algorithm determines whether two sets of nodes are independent given a third one; see Murphy, 2012, Section 10.5.

Solution:

1. By definition, we write the product of the conditionals of each node given its parents, that is,

$$p(x, y, z, \theta) = p(y|z, x)p(x|\theta)p(\theta)p(z). \tag{5}$$

2. By (5),

$$p(x, z) = \int p(x, y, z, \theta)\mathrm{d}y\mathrm{d}\theta = p(z) \int p(x|\theta)p(\theta)\mathrm{d}\theta.$$

In particular,

$$p(x) = \int p(x, z)\mathrm{d}z = \int p(x|\theta)p(\theta)\mathrm{d}\theta,$$

so that $p(x, z) = p(x)p(z)$.

---

3. We use Bayes' theorem and (5),

$$p(x, z|\theta) = \int p(x, y, z|\theta)\mathrm{d}y$$

$$= \int \frac{p(x, y, z, \theta)}{p(\theta)}\mathrm{d}y$$

$$= \int p(y|z, x)p(x|\theta)p(z)\mathrm{d}y$$

$$= p(x|\theta)p(z).$$

In particular,

$$p(z|\theta) = \int p(x, z|\theta)\mathrm{d}x = p(z),$$

so that $p(x, z|\theta) = p(x|\theta)p(z|\theta)$.

2.2 Self-normalized importance sampling ($\star\star$)

Show a central limit theorem for the self-normalized importance sampling estimator. Hint: use the delta method.

2.3 The random scan Gibbs sampler always accepts ($\star$)

Consider the MH kernel with proposal

$$q(\theta'|\theta) = \frac{1}{d}\sum_{k=1}^{d}\pi(\theta_k|\theta_{\backslash k}), \quad \theta_{\backslash k} := (\theta_1, \ldots, \theta_{k-1}, \theta_{k+1}, \ldots, \theta_d).$$

Show that the MH acceptance probability $\alpha(\theta, \theta')$ is 1. When implementing a Gibbs sampler, it is thus enough to repeatedly draw from a conditional chosen uniformly at random.

2.4 Systematic scan Gibbs sampler ($\star\star$)

Show that the systematic scan Gibbs kernel, while not satisfying detailed balance, leaves π invariant.

2.5 Gibbs ($\star$) and collapsed ($\star\star$) Gibbs for LDA

Rederive all conditionals in the LDA and collapsed LDA model. *Hint: use (2); Check (Murphy, 2012, Section 27.3.4) for the solution.*

2.6 The invariant distribution of the HMC kernel

Go through Sections 5.3 to 5.5 of **BoSa18** with pen and paper.

3 Variational inference

3.1 VB 101: fitting a univariate Gaussian ($\star$)

Consider a univariate Gaussian model $y|\mu, \lambda \sim \mathcal{N}(\mu, \lambda^{-1})$, where $\lambda = 1/\sigma^2$ is called the precision parameter.

1. Take as prior

$$p(\mu, \lambda) = \mathcal{N}(\mu|\mu_0, (\kappa_0\lambda)^{-1})\text{Gamma}(\lambda|\alpha_0, \beta_0).$$

   What is the posterior? *Hint: the prior is conjugate.*

2. Derive the updates for mean field VB in this model, i.e., with approximation

$$q(\mu, \lambda) = q(\mu)q(\lambda).$$

3. Since we know the actual posterior, what can you say of the mean field solution in that case? Could you extend VB to nonconjugate priors?

The solution is in (Murphy, 2012, Section 21.5.1).

3.2 A useful lemma for variational LDA ($\star$)

Let $\Psi(\cdot) := \Gamma'(\cdot)/\Gamma(\cdot)$ be the digamma function. Show that

$$\mathbb{E}_{\text{Dir}(\theta|\alpha)} \log \theta_i = \Psi(\alpha_i) - \Psi(\|\alpha\|_1).$$

We used that lemma when deriving the coordinatewise updates for VB with mean field approximation.

3.3 VB for LDA with counts ($\star\star$)

Derive the coordinatewise updates for VB on the count version of LDA. The variational approximation should read

$$q(\pi_i, c_i, B) = \text{Dir}(\pi_i|\tilde{\pi}_i) \prod_v \text{Multinomial}(c_{iv\cdot}|n_{iv}, \tilde{c}_{iv\cdot}) \prod_k \text{Dir}(b_{\cdot k}|\tilde{b}_{\cdot k}).$$

Hint: See Murphy, 2012, Section 27.3.6.

4 Bayesian nonparametrics

4.1 Combinatorial properties of K_n for Dirichlet process

Let K_n be the number of clusters observed when drawing n observations from a Dirichlet process with concentration parameter $\alpha \in \mathbb{R}_+$.

1. Show that

$$\mathbb{E}[K_n] = \sum_{i=0}^{n-1} \frac{\alpha}{\alpha + i} \quad \text{and} \quad \text{Var}(K_n) = \sum_{i=0}^{n-1} \frac{\alpha i}{(\alpha + i)^2}.$$

2. Show the following large n asymptotics for the expectation and variance of K_n:

$$\mathbb{E}[K_n] \sim \alpha \log n \quad \text{and} \quad \text{Var}(K_n) \sim \alpha \log n.$$

Solution:

1. The expressions are obtained by writing K_n as a sum of independent Bernoulli random variables of parameter $\frac{\alpha}{\alpha+i}$, for $i = 0, \ldots, n-1$, due to the DP predictive distribution.

2. This is obtained by factorizing by α and by Riemann sums of the intergal of $x \mapsto 1/x$ over interval $[1, n]$.

4.2 Combinatorial properties of K_n for Pitman–Yor process $(\star)$

Let K_n be the number of clusters observed when drawing n observations from a Pitman–Yor process with discount parameter $\sigma \in (0, 1)$ and concentration parameter $\alpha \in \mathbb{R}_+$.

1. Show that

$$\mathbb{E}[K_{n+1}] = \frac{\alpha}{n + \alpha} + \frac{\sigma + \alpha + n}{n + \alpha} \mathbb{E}[K_n].$$

Hint: use the PY predictive distribution and a conditional expectation to get this iterative formula from n to $n + 1$.

2. Deduce that
$$\mathbb{E}[K_n] = \frac{\alpha}{\sigma} \left(\frac{(\alpha + \sigma)_n}{(\alpha)_n} - 1 \right),$$

where $(x)_n = x(x + 1) \ldots (x + n - 1)$.

3. Show the following large n asymptotics for the expectation of K_n:

$$\mathbb{E}[K_n] \sim \frac{\Gamma(\alpha + 1)}{\sigma \Gamma(\alpha + \sigma)} n^\sigma.$$

4. Show that the following recursive formula holds for the variance of K_n:

$$\text{Var}(K_{n+1}) = \text{Var}(K_n) \frac{n + \alpha + 2\sigma}{n + \alpha} + \frac{(\alpha + \sigma)_n}{(\alpha + 1)_n} \left(1 - \frac{(\alpha + \sigma)_n}{(\alpha + 1)_n} \right).$$

Hint: use the law of total variance.

5. Derive again the simpler expression of expectation and variance of K_n in the Dirichlet process case.

6. Pitman, 2002 (Exercise 3.3.4) obtains the following limiting behaviour of moments of K_n of any positive order $p > 0$:

$$\mathbb{E}[K_n^p] \sim n^{\sigma p} \frac{\Gamma(\alpha/\sigma + p + 1)\Gamma(\alpha + 1)}{\Gamma(\alpha/\sigma + 1)\Gamma(\alpha + p\sigma + 1)}.$$

Use this to characterise the asymptotic behaviour of $\mathrm{Var}(K_n)$.

Solution:

1. By the Pólya urn representation of PY (or, equivalently, its predictive distribution), we have

$$\mathbb{P}(K_{n+1} = K + 1 \mid K_1, \ldots, K_n = K) = \frac{K\sigma + \alpha}{n + \alpha},$$

$$\mathbb{P}(K_{n+1} = K \mid K_1, \ldots, K_n = K) = \frac{n - K\sigma}{n + \alpha}$$

hence

$$\mathbb{E}[K_{n+1}] = \mathbb{E}[\mathbb{E}[K_{n+1} \mid K_1, \ldots, K_n = K]] = \frac{\alpha}{n + \alpha} + \frac{\sigma + \alpha + n}{n + \alpha}\mathbb{E}[K]. \tag{6}$$

2. This is obtained by induction and by noting that the proposed formula is verified for $n = 1$, leading to $\mathbb{E}[K_1] = 1$.

3. By properties of the gamma function:

$$\mathbb{E}[K_n] \sim \frac{\alpha}{\sigma} \frac{(\alpha + \sigma)_n}{(\alpha)_n} = \frac{\alpha}{\sigma} \frac{\Gamma(\alpha + \sigma + n)\Gamma(\alpha)}{\Gamma(\alpha + n)\Gamma(\alpha + \sigma)}$$

$$\sim \frac{\alpha\Gamma(\alpha)}{\sigma\Gamma(\alpha + \sigma)}n^\sigma = \frac{\Gamma(\alpha + 1)}{\sigma\Gamma(\alpha + \sigma)}n^\sigma.$$

4. Using the law of total variance we have,

$$\mathrm{Var}(K_{n+1}) = \mathbb{E}[\mathrm{Var}(K_{n+1} \mid K_1, \ldots, K_n = K)] \tag{7}$$
$$+ \mathrm{Var}(\mathbb{E}[K_{n+1} \mid K_1, \ldots, K_n = K]).$$

Using the fact that the variance of a two-point distribution with outcomes a and b and probabilities p and $1 - p$ is $p(1 - p)(a - b)^2$, we

can write the first term of the right-hand side of (7) as

$$\mathbb{E}[\mathrm{Var}(K_{n+1} \mid K_1, \ldots, K_n = K)] = \mathbb{E}\left[\frac{(K\sigma + \alpha)(n - K\sigma)}{(n + \alpha)^2}\right]$$

$$= \frac{\mathbb{E}[K_n]\sigma(n - \alpha) + n\alpha - \sigma^2(\mathbb{E}[K_n]^2 + \mathrm{Var}(K_n))}{(n + \alpha)^2}, \quad (8)$$

while (6) provides for second term of the right-hand side of (7)

$$\mathrm{Var}(\mathbb{E}[K_{n+1} \mid K_1, \ldots, K_n = K]) = \left(\frac{\sigma + n + \alpha}{n + \alpha}\right)^2 \mathrm{Var}(K_n). \quad (9)$$

Then combining (8) and (9), the following recursive formula holds

$$\mathrm{Var}(K_{n+1}) = \mathrm{Var}(K_n)\frac{n + \alpha + 2\sigma}{n + \alpha} + \frac{\sigma\mathbb{E}[K_n](n - \alpha - \sigma\mathbb{E}[K_n]) + n\alpha}{(n + \alpha)^2}$$

$$= \mathrm{Var}(K_n)\frac{n + \alpha + 2\sigma}{n + \alpha} + \frac{(\sigma\mathbb{E}[K_n] + \alpha)(n - \sigma\mathbb{E}[K_n])}{(n + \alpha)^2},$$

$$(10)$$

which simplifies to the desired expression by using

$$\mathbb{E}[K_n] = \frac{\alpha}{\sigma}\left(\frac{(\alpha + \sigma)_n}{(\alpha)_n} - 1\right).$$

5. This simply comes from setting $\sigma = 0$ in the above formulas.

4.3 For more exercises on Bayesian nonparametrics $(\star\star)$

- Exercises 4.4, 4.8, 4.9, 4.10, 4.12, 4.13, 4.18, 4.24, 4.25, 4.26, 4.32, 4.39 of Ghosal and Van der Vaart, 2017.

5 Foundations

5.1 A simple application of the likelihood principle $(\star)$

Consider experiments E_1: tossing a coin 10 times, vs. E_2: tossing the same coin until obtaining 4 heads. Say we ran E_1 and E_2, and we obtained two samples of the same size $n = 10$.

1. Write down the binomial and negative binomial likelihoods corresponding to E_1 and E_2, respectively.

2. Build two credible intervals for the bias θ of the coin, one for each experiment. Are the two intervals the same?

3. Build two (frequentist) confidence intervals for the bias θ of the coin, one for each experiment. Are the two intervals the same?

4. Which answer bothers you the most?

5.2 The von Neumann-Morgenstern axioms ($\star\star$)

In this problem, we will derive the utility principle from a set of axioms on a preference among actions, like Savage, 1954, but in the simplified setting of the seminal contribution of Von Neumann and Morgenstern, 1944. This will allow us to see how an expected utility can come almost out of the blue. This exercise is drawn from (Parmigiani and Inoue, 2009, Section 3.4.2).

Consider a finite set of rewards $\mathcal{Z} = \{z_1, \ldots, z_n\}$. Actions $\mathcal{A}$ are all functions from the set $\mathcal{S}$ of states to $\mathcal{Z}$. Let π be a probability over $(\mathcal{S}, \Sigma_{\mathcal{S}})$, with $\Sigma_{\mathcal{S}}$ a $\sigma-$algbra, to be interpreted as a "well-understood chance mechanism"; think being able to draw from the uniform distribution. Consider $\succ$ a binary relation on $\mathcal{A}$. Assume that

$$\succ \text{ is complete and transitive.} \tag{vNM1}$$

By complete, we mean that for any pair of actions $a, b \in \mathcal{A}$, then either $a \prec b$ holds, or $b \prec a$ holds, or neither $a \succ b$ nor $b \succ a$. In the latter case, we write $a \sim b$, and think of the decision maker as being indifferent to picking a or b. For simplicity, we also define $a \prec b$ as $b \succ a$, and $a \preceq b$ as the disjunction of $a \prec b$ and $a \sim b$. By transitivity, we mean that

$$a \preceq b \preceq c \Rightarrow a \preceq c.$$

We further assume that

$$\forall a, b, c \in \mathcal{A}, \alpha \in (0, 1], \quad a \succ b \Rightarrow (1-\alpha)c + \alpha a \succ (1-\alpha)c + \alpha b, \tag{vNM2}$$

and that

$$\forall a, b, c \in \mathcal{A} \text{ such that } a \succ b \succ c, \exists \alpha, \beta \in (0, 1) \text{ such that}$$
$$\alpha a + (1-\alpha)c \succ b \succ \beta a + (1-\beta)c. \tag{vNM3}$$

1. How do you interpret each of the three assumptions (**??**), (**??**), and (**??**)?

2. Let $a \succ b$. Show that $\beta a + (1-\beta)b \succ \alpha a + (1-\alpha)b$ iff $0 \leqslant \alpha < \beta \leqslant 1$.

3. Show that $a \succeq b \succeq c$ and $a \succ c$ together imply the existence of a unique $\alpha^\star$ such that
$$b \sim \alpha a + (1-\alpha)c.$$

4. Let $a \sim b$ and $\alpha \in [0, 1]$. Prove that for any $c \in \mathcal{A}$,
$$\alpha a + (1-\alpha)c \sim \alpha b + (1-\alpha)c.$$

5. For $z \in \mathcal{Z}$, define $\chi_z : s \mapsto z$ to be the corresponding constant action. Prove by induction on n that there exists $\underline{z}, \overline{z} \in \mathcal{Z}$ such that for all $a \in \mathcal{A}$, $\chi_{\underline{z}} \preceq a \preceq \chi_{\overline{z}}$. In the sequel, we further assume that the decision maker is not indifferent to all actions, so that $\chi_{\underline{z}} \prec \chi_{\overline{z}}$.

6. Let $a \in \mathcal{A}$. Prove the existence of a unique $\varphi(a) \in [0, 1]$ such that

$$a \sim \varphi(a)\chi_{\overline{z}} + (1 - \varphi(a))\chi_{\underline{z}}.$$

7. Show that $a \prec b \Leftrightarrow \varphi(a) < \varphi(b)$. It thus remains to prove that φ takes the form of an expected utility.

8. For $a, b \in \mathcal{A}$, $\alpha \in [0, 1]$, prove that $\varphi(\alpha a + (1-\alpha)b) = \alpha\varphi(a) + (1-\alpha)\varphi(b)$.

9. For $z \in \mathcal{Z}$, let $u(z) = \varphi(\chi_z)$. Prove by induction on n that

$$\varphi(a) = \mathbb{E}_\pi u(a(s)).$$

> **Solution:** See (Parmigiani and Inoue, 2009, Section 3.4.2).

5.3 The Blackwell-McQueen urn scheme and exchangeability ($\star\star$)

1. Show that the colors $X_1, \dots$ drawn in the BMC urn scheme are exchangeable.

2. Prove that the corresponding measure on $\mathcal{P}(\mathcal{X})$ given by de Finetti's theorem is a Dirichlet process.

5.4 McAllester's PAC bound ($\star\star\star$)

Prove McAllester's PAC bound. Hint: Check out Chapter 31 of (Shalev-Shwartz and Ben-David, 2014).

6 Bayesian deep learning

6.1 Laplace approximation for Bayesian neural networks ($\star$)

1. Exercises 5.38, 5.39, 5.40, 5.41 of Bishop, 2006.
   Note: Solutions to those exercises will not be provided; "Auditeurs libres" can work on them for their project.

6.2   Gaussian process limit of wide-regime Bayesian neural networks (⋆⋆)

1. Derive the Gaussian process limit of wide-regime Bayesian neural networks with isotropic (iid) Gaussian priors on the weights, in the case of 1-hidden-layer neural networks (Neal's result). In particular, show that variances of the Gaussian priors need to be scaled in $1/H$, with H the hidden layer width.

2. Derive the Gaussian process limit for the more general case of non shallow neural networks (refer to the proof by Matthews et al., 2018).

References

[1] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[2] B. Efron. *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*. Vol. 1. Cambridge University Press, 2012.

[3] Subhashis Ghosal and Aad Van der Vaart. *Fundamentals of nonparametric Bayesian inference*. Vol. 44. Cambridge University Press, 2017.

[4] J.-M. Marin and C.P. Robert. *Bayesian Core: A Practical Approach to Computational Bayesian Statistics*. New York: Springer-Verlag, 2007.

[5] A. Matthews et al. "Gaussian process behaviour in wide deep neural networks". In: *International Conference on Learning Representations*. Vol. 1804.11271. 2018.

[6] K. Murphy. *Machine learning: a probabilistic perspective*. MIT Press, 2012.

[7] G. Parmigiani and L. Inoue. *Decision theory: principles and approaches*. Vol. 812. John Wiley & Sons, 2009.

[8] J. Pitman. *Combinatorial stochastic processes*. Vol. 1875. Lecture Notes in Mathematics. Lectures from the 32nd Summer School on Probability Theory held in Saint-Flour, July 7–24, 2002. Berlin: Springer-Verlag, 2002, pp. x+256.

[9] L. J. Savage. *The Foundations of Statistics*. John Wiley & Sons, 1954.

[10] S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[11] J. Von Neumann and O. Morgenstern. *Theory of games and economic behavior*. John Wiley & Sons, 1944.