Gaussian Linear Dynamical Systems, Bayesian Bandits, and Gaussian Processes

Razeen Wasif

April 17, 2025

1 Bayes' rule warm-up

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

where $p(\theta|y, X)$ is the posterior, $p(y, |\theta, X)$ is the likelihood, $p(\theta)$ is the prior, and $p(y|X)$ is the marginal likelihood. In Bayesian Linear Regression, we use a Gaussian likelihood and a Gaussian prior as follows,

$$p(\theta) = \mathcal{N}(\theta; \mu_0, \Sigma_0)$$
$$p(y|, \theta, X) = \mathcal{N}(y; X\theta, \sigma_y^2 \mathrm{I}),$$

where $\mathbf{X} \in \mathbb{R}^{N \times D}$, $y \in \mathbb{R}^N$, $\theta \in \mathbb{R}^D$, σ_y^2 is the observation noise, and μ_0 and Σ_0 are the prior mean and covariance, respectively. *Derive the posterior $p(\theta|y, X)$ and the marginal likelihood $p(y|X)$.*

Solution:
The Gaussian distribution over a D-dimensional vector X is defined by:

$$\mathcal{N}(x|\mu, \Sigma) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \exp\left\{ -\frac{1}{2}(x-\mu)^{\mathrm{T}} \Sigma^{-1} (x-\mu) \right\}$$

The posterior can be found given the following formula:

$$\textbf{posterior} \propto \textbf{likelihood} \times \textbf{prior}$$

where the likelihood and prior are:

$$p(y|, \theta, X) = \frac{1}{(2\pi)^{N/2}} \frac{1}{|\sigma_y^2|^{N/2}} \exp\left\{ -\frac{1}{2\sigma_y^2}(y - X\theta)^{\mathrm{T}} (\sigma_y^2 \mathrm{I})^{-1} (y - X\theta) \right\}$$

$$p(\theta) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma_0|^{1/2}} \exp\left\{ -\frac{1}{2}(\theta - \mu_0)^{\mathrm{T}} \Sigma_0^{-1} (\theta - \mu_0) \right\}$$

Before multiplying the two, we take the **log** of both terms so we can focus on terms involving θ:

$$p(\theta|y, X) = \left(-\frac{1}{2\sigma_y^2}(y - X\theta)^{\mathrm{T}} (y - X\theta) \right) \times \left(-\frac{1}{2}(\theta - \mu_0)^{\mathrm{T}} \Sigma_0^{-1} (\theta - \mu_0) \right)$$

$$= -\frac{1}{2\sigma_y^2} \left(y^{\mathrm{T}} y - y^{\mathrm{T}} X\theta - \theta^{\mathrm{T}} X^{\mathrm{T}} y + \theta^{\mathrm{T}} X^{\mathrm{T}} X\theta \right) \cdot - \frac{1}{2} \left(\theta^{\mathrm{T}} \Sigma_0^{-1} \theta - \theta^{\mathrm{T}} \Sigma_0^{-1} \mu_0 - \mu_0^{\mathrm{T}} \Sigma_0^{-1} \theta + \mu_0^{\mathrm{T}} \Sigma_0^{-1} \mu_0 \right)$$

ignore terms without θ and simplify:

$$= -\frac{1}{2\sigma_y^2} \left(y^{\mathrm{T}} X\theta - \theta^{\mathrm{T}} X^{\mathrm{T}} y + \theta^{\mathrm{T}} X^{\mathrm{T}} X\theta \right) \cdot - \frac{1}{2} \left(\theta^{\mathrm{T}} \Sigma_0^{-1} \theta - \theta^{\mathrm{T}} \Sigma_0^{-1} \mu_0 - \mu_0^{\mathrm{T}} \Sigma_0^{-1} \theta \right)$$

$$= \left(-\frac{1}{2\sigma_y^2}(\theta^{\mathrm{T}} X^{\mathrm{T}} X\theta) + \frac{1}{2\sigma_y^2}(y^{\mathrm{T}} X\theta) + \frac{1}{2\sigma_y^2}(\theta^{\mathrm{T}} X^{\mathrm{T}} y) \right) \cdot \left(-\frac{1}{2}(\theta^{\mathrm{T}} \Sigma_0^{-1} \theta) + \frac{1}{2}(\theta^{\mathrm{T}} \Sigma_0^{-1} \mu_0) + \frac{1}{2}(\mu_0^{\mathrm{T}} \Sigma_0^{-1} \theta) \right)$$

$y^T X\theta = \theta^T X^T y$ so simplify to $\theta^T X^T y$. Similarly $\theta^T \Sigma_0^{-1} \mu_0 = \mu_0^T \Sigma_0^{-1} \theta$ so simplify to $\theta^T \Sigma_0^{-1} \mu_0$

$$= \exp\bigg(-\frac{1}{2\sigma_y^2}(\theta^{\mathrm{T}} X^{\mathrm{T}} X\theta) + \frac{1}{\sigma_y^2}(\theta^{\mathrm{T}} X^{\mathrm{T}} y)\bigg) \cdot \exp\bigg(-\frac{1}{2}(\theta^{\mathrm{T}} \Sigma_0^{-1}\theta) + \frac{1}{2}(\theta^{\mathrm{T}} \Sigma_0^{-1}\mu_0)\bigg)$$

using the exponential property of $exp(a) \cdot exp(b) = exp(a+b)$ we can simplify the formula to:

$$\exp\bigg[-\frac{1}{2\sigma_y^2}(\theta^{\mathrm{T}} X^{\mathrm{T}} X\theta) + \frac{1}{\sigma_y^2}(\theta^{\mathrm{T}} X^{\mathrm{T}} y) - \frac{1}{2}(\theta^{\mathrm{T}} \Sigma_0^{-1}\theta) + \frac{1}{2}(\theta^{\mathrm{T}} \Sigma_0^{-1}\mu_0)\bigg]$$

Thus we have:

$$\mathbf{log}(p(\theta|y,X)) \propto \mathbf{exp}\bigg[-\frac{1}{2}\theta^{\mathrm{T}}\bigg(\frac{X^T X}{\sigma_y^2} + \Sigma_0^{-1}\bigg)\theta + \theta^T\bigg(\frac{X^T y}{\sigma_y^2} + \Sigma_0^{-1}\mu_0\bigg)\bigg]$$

The standard form of the log of multivariate Gaussian is:

$$\mathbf{log}\mathcal{N}(\theta;\mu,\Sigma) = \mathbf{exp}\bigg(-\frac{1}{2}(\theta-\mu)^{\mathrm{T}}\Sigma^{-1}(\theta-\mu)\bigg)$$

Which expands to:

$$= -\frac{1}{2}\theta^{\mathrm{T}}\Sigma^{-1}\theta + \theta^{\mathrm{T}}\Sigma^{-1}\mu - \frac{1}{2}\mu^{\mathrm{T}}\Sigma^{-1}\mu$$

Comparing this with the expression we have, we can see that:

$$\Sigma^{-1} = \bigg(\frac{X^T X}{\sigma_y^2} + \Sigma_0^{-1}\bigg) \text{ and } \Sigma^{-1}\mu = \bigg(\frac{X^T y}{\sigma_y^2} + \Sigma_0^{-1}\mu_0\bigg).$$

Therefore:

$$\Sigma = \bigg(\frac{X^T X}{\sigma_y^2} + \Sigma_0^{-1}\bigg)^{-1} \text{ and } \mu = \Sigma\bigg(\frac{X^T y}{\sigma_y^2} + \Sigma_0^{-1}\mu_0\bigg)$$

This gives us the posterior distribution $p(\theta|y,X) = \mathcal{N}(\theta;\mu,\Sigma)$.

The marginal likelihood can be found through:

$$p(y|X) = \int p(y|\theta,X)p(\theta)\ d\theta$$

Start by expanding the likelihood:

$$p(y|\theta,X) = \mathcal{N}(y;X\theta,\sigma_y^2 I) \propto \exp\left(-\frac{1}{2\sigma_y^2}(y-X\theta)^T(y-X\theta)\right)$$

$$(y-X\theta)^T(y-X\theta) = y^T y - y^T X\theta - \theta^T X^T y + \theta^T X^T X\theta$$

Now expand the prior:

$$p(\theta) = \mathcal{N}(\theta;\mu_0,\Sigma_0) \propto \exp\left(-\frac{1}{2}(\theta-\mu_0)^T\Sigma_0^{-1}(\theta-\mu_0)\right)$$

$$(\theta-\mu_0)^T\Sigma_0^{-1}(\theta-\mu_0) = \theta^T\Sigma_0^{-1}\theta - \theta^T\Sigma_0^{-1}\mu_0 - \mu_0^T\Sigma_0^{-1}\theta + \mu_0^T\Sigma_0^{-1}\mu_0$$

Combine and organize the terms involving θ:

$$-\frac{1}{2\sigma_y^2}(y^T y - 2\theta^T X^T y + \theta^T X^T X\theta) - \frac{1}{2}(\theta^T\Sigma_0^{-1}\theta - 2\theta^T\Sigma_0^{-1}\mu_0 + \mu_0^T\Sigma_0^{-1}\mu_0) \tag{1}$$

$$= -\frac{1}{2\sigma^2}y^T y + \frac{1}{2\sigma^2}2\theta^T X^T y - \frac{1}{2\sigma^2}\theta^T X^T X\theta - \frac{1}{2}\theta^T\Sigma^{-1}\theta + \frac{1}{2}2\theta^T\Sigma^{-1}\mu - \frac{1}{2}\mu^T\Sigma^{-1}\mu \tag{2}$$

$$= -\frac{1}{2\sigma^2}y^T y + \frac{1}{\sigma^2}\theta^T X^T y - \frac{1}{2\sigma^2}\theta^T X^T X\theta - \frac{1}{2}\theta^T\Sigma^{-1}\theta + \theta^T\Sigma^{-1}\mu - \frac{1}{2}\mu^T\Sigma^{-1}\mu \tag{3}$$

$$= \frac{1}{\sigma^2}\theta^T X^T y - \frac{1}{2\sigma^2}\theta^T X^T X\theta - \frac{1}{2}\theta^T\Sigma^{-1}\theta + \theta^T\Sigma^{-1}\mu + C \tag{4}$$

$$= -\frac{1}{2}\theta^T\left(\frac{1}{\sigma_y^2}X^T X + \Sigma_0^{-1}\right)\theta + \theta^T\left(\frac{1}{\sigma_y^2}X^T y + \Sigma_0^{-1}\mu_0\right) + C \tag{5}$$

where C includes terms not involving θ.
To complete the square, define A and b as:

$$A = \frac{1}{\sigma_y^2} X^T X + \Sigma_0^{-1} \tag{6}$$

$$b = \frac{1}{\sigma_y^2} X^T y + \Sigma_0^{-1} \mu_0 \tag{7}$$

re-write the exponent as:

$$-\frac{1}{2}\theta^T A \theta + \theta^T b + C$$

Completing the square gives:

$$-\frac{1}{2}(\theta - A^{-1}b)^T A (\theta - A^{-1}b) + \frac{1}{2} b^T A^{-1} b + C$$

The first term represents the posterior distribution $p(\theta|y, X) = \mathcal{N}(\theta; A^{-1}b, A^{-1})$.
When integrating with respect to θ, only the terms that don't involve θ remain, giving us:

$$p(y|X) \propto \exp\left(-\frac{1}{2\sigma_y^2} y^T y + \frac{1}{2} b^T A^{-1} b - \frac{1}{2}\mu_0^T \Sigma_0^{-1} \mu_0\right) \tag{8}$$

$$p(y|X) = \mathcal{N}(y; X\mu_0, \sigma_y^2 I + X\Sigma_0 X^T) \tag{9}$$

2 Gaussian linear dynamical systems

In many real-world engineering settings, the states of physical objects over time such as location or speed, often governed by complex and unknown dynamics, are only observed through noisy sensors. It is thus important to de-noise and estimate the true states. One way to do this is to write down a model of the true state, how it changes over time, and how it is linked to the observed sensor measurements, and subsequently perform "inference" to estimate the state given the observations. We will study a model class with linear dynamics and Gaussian noise as follows:

$$s_t = A s_{t-1} + w_t,$$

$$y_t = C s_t + v_t,$$

where $s_t \in \mathbb{R}^{D_s}$ is the state at time t, $w_t \in \mathbb{R}^{D_s}$ is the noise in the dynamics, $y_t \in \mathbb{R}^{D_y}$ is the observation at time t, v_t is the observation noise. We assume the noises are Gaussian-distributed, $\mathbf{w}_t \sim \mathcal{N}(0; \mathbf{Q})$ and $v_t \sim \mathcal{N}(0; \mathbf{R})$, and $\mathbf{A} \in \mathbb{R}^{D_s \times D_s}, \mathbf{C} \in \mathbb{R}^{D_y \times D_s}, \mathbf{Q} \in \mathbb{R}^{D_s \times D_s}, \mathbf{R} \in \mathbb{R}^{D_y \times D_y}$ are fixed over time. The initial state $\mathbf{s}_0$ is assumed to be drawn from $\mathcal{N}(\mu_0, \Sigma_0)$.
It is perhaps more pedagogical to see how to sequentially generate data from such a model using the following steps/densities:

$$\begin{aligned}
p(s_0) &= \mathcal{N}(s_0; \mu_0, \Sigma_0) \\
p(s_1|s_0) &= \mathcal{N}(s_1; \mathbf{A}\mathbf{s}_0, \mathbf{Q}) \\
p(y_1|s_1) &= \mathcal{N}(y_1; \mathbf{C}\mathbf{s}_0, \mathbf{R}) \\
&\vdots \\
p(s_t|s_{t-1}) &= \mathcal{N}(s_t; \mathbf{A}\mathbf{s}_{t-1}, \mathbf{Q}) \\
p(y_t|s_t) &= \mathcal{N}(y_t; \mathbf{C}\mathbf{s}_t, \mathbf{R})
\end{aligned}$$

We illustrate two trajectories from such a system below.
Assume that we know the dynamics and noise configurations, **A, C, Q, R,** μ_0, Σ_0 and have observed data up to $T, \mathbf{y}_{1:T}$, the typical task is to estimate the hidden state s as data are collected, e.g. estimate the location and speed of the robot in the example above.

2.1 Question 1.1: Find $p(s_1)$ and $p(s_1|y_1)$

Both s_0 and w are Gaussian with $s_0 \sim \mathcal{N}(\mu, \Sigma)$ and $w \sim \mathcal{N}(0; Q)$ and s_1 is a linear transformation of s_0 with some noise so the mean of s_1 can be found using the expectation of a Gaussian distribution:

$$\mathbb{E}[s_1] = \mathbb{E}[As_0 + w_1] = A\mathbb{E}[s_0] + \mathbb{E}[w_1] = A\mu + 0 = A\mu$$

and the covariance is:

$$\text{Cov}(s_1) = \text{Cov}(As_0 + w_1) = A\text{Cov}(s_0)A^{\text{T}} + \text{Cov}(w_1) = A\Sigma A^{\text{T}} + \mathbf{Q}$$

altogether, $p(s_1)$ is: $\mathcal{N}(s_1; A\mu,\ A\Sigma A^{\text{T}} + \mathbf{Q})$.

We know that $p(y_1|s_1)$ is given by $\mathcal{N}(y_1; \mathbf{C}s_1, \mathbf{R})$ and we have already found $p(s_1)$. Therefore $p(s_1|y_1)$ can be found using Gaussian conditioning where we have the following[7]:

$$\mu_n = A\mu_{n-1} + K_n(x_n - CA\mu_{n-1})$$
$$V_n = (I - K_nC)(AV_{n-1}A^{\text{T}} + \Gamma)$$
$$K_n = AV_{n-1}A^{\text{T}} + \Gamma C^{\text{T}}(CAV_{n-1}A^{\text{T}} + \Gamma C^{\text{T}} + \Sigma)^{-1}$$

where K is the Kalman gain matrix. Applying these to $p(s_1|y_1)$ we get:

$$\mu_1 = A\mu_0 + K_1(y_1 - CA\mu_0)$$
$$V_1 = (I - K_1C)(A\Sigma_0A^{\text{T}} + Q)$$
$$K_1 = A\Sigma_0A^{\text{T}} + QC^{\text{T}}(CA\Sigma_0A^{\text{T}} + QC^{\text{T}} + R)^{-1}$$

Therefore $p(s_1|y_1) = \mathcal{N}(s_1; A\mu_0 + K_1(y_1 - CA\mu_0), (I - K_1C)(A\Sigma_0A^{\text{T}} + Q))$ with $K = A\Sigma_0A^{\text{T}} + QC^{\text{T}}(CA\Sigma_0A^{\text{T}} + QC^{\text{T}} + R)^{-1}$

2.2 Question 1.2: Find $p(s_t|y_{1:t-1})$ and $p(s_t|y_{1:t})$

Given $p(s_{t-1}|y_{1:t-1}) = \mathcal{N}(s_{t-1}; \mu_{t-1}, \Sigma_{t-1})$, find $p(s_t|y_{1:t-1})$ and $p(s_t|y_{1:t})$ in terms of $\mu_{t-1}, \Sigma_{t-1}, \mathbf{A}, \mathbf{C}, \mathbf{Q}$ and $\mathbf{R}$.

The state transition equation is: $s_t = As_{t-1} + w_t,\ w_t \sim \mathcal{N}(0, Q)$ so the mean and covariance of s_t will be:

$$\mathbb{E}[s_t] = \mathbb{E}[As_{t-1} + w_t] = A\mathbb{E}[s_{t-1}] + \mathbb{E}[w_t] = A\mu_{t-1} + 0 = A\mu_{t-1}$$
$$\text{Cov}(s_t) = \text{Cov}(As_{t-1} + w_t) = A\text{Cov}(s_{t-1})A^{\text{T}} + \text{Cov}(w_t) = A\Sigma_{t-1}A^{\text{T}} + \mathbf{Q}$$

Therefore $p(s_t|y_{1:t-1}) = \mathcal{N}(s_t; A\mu_{t-1}, A\Sigma_{t-1}A^{\text{T}} + \mathbf{Q})$

The observation y at any given time t is given by $y_t = Cs_t + v_t,\ v_t \sim \mathcal{N}(0, \mathbf{R})$.
This will again lead to a Gaussian posterior distribution $p(s_t|y_{1:t})$ given by the following:

$$\mu = A\mu_{t-1} + K_t(y_t - CA\mu_{t-1})$$
$$\Sigma_t = (I - K_tC)A\Sigma_{t-1}A^{\text{T}} + \mathbf{Q}$$
$$K_t = A\Sigma_{t-1}A^{\text{T}} + \mathbf{Q}C^{\text{T}}(CA\Sigma_{t-1}A^{\text{T}} + \mathbf{Q}C^{\text{T}} + R)^{-1}$$

Therefore $p(s_t|y_{1:t}) = \mathcal{N}(s_t; A\mu_{t-1} + K_t(y_t - CA\mu_{t-1}), (I - K_tC)A\Sigma_{t-1}A^{\text{T}} + \mathbf{Q})$

2.3 Question 1.3: Implement the results in Question 1.2

2.4 Question 1.4: Find $p(y_t|y_{1:t-1})$

The mean and covariance of y_t are:

$$\mathbb{E}[y_t|y_{1:t-1}] = C\mathbb{E}[s_t|y_{1:t-1}] = CA\mu_{t-1}$$
$$Cov(y_t|y_{1:t-1}) = C\mathbf{Cov}(s_t|y_{1:t-1})C^{\text{T}} + R = CA\Sigma_{t-1}A^{\text{T}}C^{\text{T}} + R$$

Since y_t is Gaussian we have:

$$p(y_t|y_{1:t-1}) = \mathcal{N}(y_t; CA\mu_{t-1}, CA\Sigma_{t-1}A^{\text{T}}C^{\text{T}} + R)$$

2.5 Question 1.5 Find the log marginal likelihood $p(y_{1:T})$

Given the result in the last question, discuss how to efficiently compute the log marginal likelihood log $p(y_{1:T})$. In practice, this quantity can be used for model selection and hyper-parameter optimisation.
Solution: [7]
To compute the log marginal likelihood we can use the fact that the observations $y_{1:T}$ are generated sequentially, and the joint distribution can be factorized as:

$$p(y_{1:T}) = p(y_1) \prod_{t=2}^{\mathrm{T}} p(y_t|y_{1:t-1})$$

Each term $p(y_t|y_{1:t-1}$ is a Gaussian distribution and the log marginal likelihood is the sum of the log probabilities of all these Gaussian distributions. Start by initializing the state mean μ_o and covariance Σ_0. For each time step in t:

1. predict the state distribution $p(s_t|y_{1:t-1})$ using the Kalman filter prediction step
2. compute the observation distribution $p(y_t|y_{1:t-1})$ using the predicted state
3. update the state distribution $p(s_t|y_{1:t})$ using the Kalman filter update step
4. add the log probability of $p(y_t|y_{1:t-1})$ to the log marginal likelihood

Finally return the total log marginal likelihood.

2.6 Question 1.6 [Optional] How do we deal with a non linear dynamics between s_{t-1} and s_t and or a non linear mapping between s_t and y_t

The standard Kalman filter won't work since the predictions and updates are no longer Gaussian so we need to use approximation methods. One such way is using the Extended Kalman Filter (EKF) which linearizes about an estimate of the current mean and covariance. In EKF the state transition and observation models aren't linear but differentiable functions of the form:

$$x_k = f(x_{k-1}, u_{k-1}) + w_{k-1} \tag{10}$$
$$z_k = h(x_k) + v_k \tag{11}$$

The function f can be used to compute the predicted state from the previous estimate and the function h can be used to compute the predicted measurement from the predicted state. But f and h cannot be applied directly to the covariance, instead the Jacobian (a matrix of partial derivatives) is computed. At each time step the Jacobian is evaluated with current predicted states. These matrices can be used in the Kalman filter equations. This process essentially linearizes the non-linear function around the current estimate[3].

3 Bayesian Bandits

let θ_a and θ_b , both $\in [0, 1]$, be the unknown efficacy of the drugs respectively. The outcome for each patient allocated to each drug pool can be modelled as a Bernoulli trial as follows,

$$p(y_a|\theta_a) = \theta_a^{y_a}(1-\theta_a)^{1-y_a} \tag{12}$$
$$p(y_b|\theta_b) = \theta_b^{y_b}(1-\theta_b)^{1-y_b}, \tag{13}$$

where $y_a = 1$ meas treated, $y_a = 0$ means not treated, and similarly for y_b. Assume that we have allocated Y_a and Y_b patients to group A and B, respectively, and of which, the treatment for S_a and S_b patients were successful. We place a Beta prior over θ_a and θ_b, $p(\theta_a) = \mathrm{Beta}(\theta_a; 1, 1)$ and $p(\theta_b) = \mathrm{Beta}(\theta_b; 1, 1)$. Note the pdf of $\mathrm{Beta}(\theta; \alpha, \beta)$ is $p(\theta) = B(\alpha, \beta)^{-1}\theta^{\alpha-1}(1-\theta)^{\beta-1}$, where B is the beta function.

3.1 Question 2.1 Find $p(\theta_a|Y_a)$ and $p(\theta_b|Y_b)$

The likelihood of observing S_n successes out of Y_n patients is given by the binomial distribution.

$$p(S_a|\theta_a, Y_a) = \binom{Y_a}{S_a}\theta_a^{S_a}(1-\theta_a)^{Y_a-S_a} \tag{14}$$

$$p(S_b|\theta_b, Y_b) = \binom{Y_b}{S_b}\theta_b^{S_b}(1-\theta_b)^{Y_b-S_b} \tag{15}$$

The priors are given by:

$$p(\theta_a) = \text{Beta}(\theta; 1, 1) = \text{B}(1,1)^{-1}\theta^{1-1}(1-\theta)^{1-1} = 1 \tag{16}$$

$$p(\theta_b) = \text{Beta}(\theta; 1, 1) = \text{B}(1,1)^{-1}\theta^{1-1}(1-\theta)^{1-1} = 1 \tag{17}$$

Using the likelihood and prior we can now find the posterior:

$$p(\theta_n|Y_n, S_n) \propto p(S_n|\theta_n, Y_n) \cdot p(\theta_n) \tag{18}$$

$$= \theta_n^{S_n}(1-\theta_n)^{Y_n-S_n} \times 1 \tag{19}$$

$$\tag{20}$$

To match the form of the Beta distribution, equate the exponents where:

$$\alpha = S_n + 1, \quad \beta = Y_n - S_n + 1$$

$\therefore$ the posterior distributions of $p(\theta_a|Y_a)$ and $p(\theta_b|Y_b)$ are:

$$p(\theta_a|Y_a) = \text{Beta}(\theta_a; S_a + 1, Y_a - S_a + 1) \tag{21}$$

$$p(\theta_b|Y_b) = \text{Beta}(\theta_b; S_b + 1, Y_b - S_b + 1) \tag{22}$$

3.2 Question 2.2

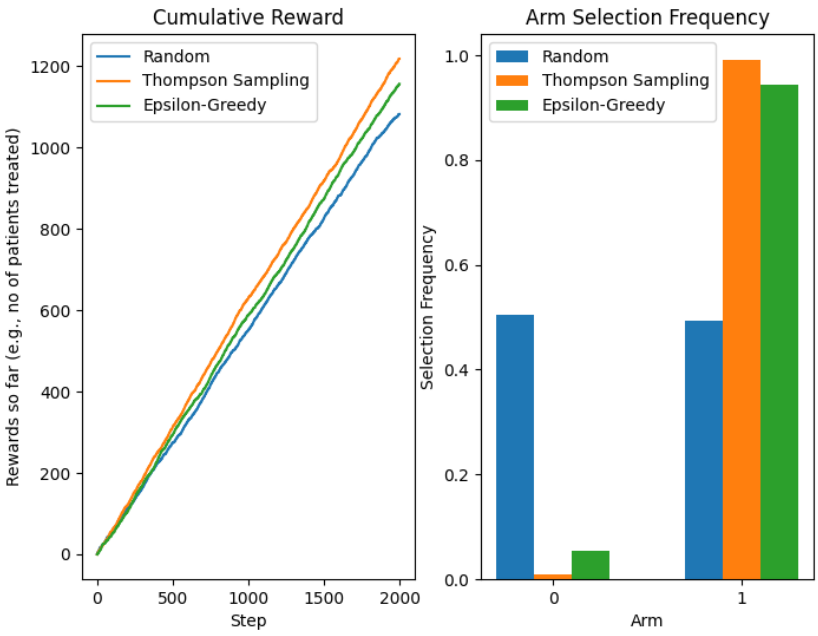

Figure 1: Caption

The figure shows that the Thompson Sampling and Epsilon-Greedy accumulated rewards faster than Random Selection. They also showed a higher selection frequency for arm 2 (with the higher probability) since random selection would choose both arms with roughly equal frequency.

3.3 Question 2.3 Find $p(\theta_a|\mathcal{D}_a)$ and $p(\theta_b|\mathcal{D}_b)$

We will now consider a two-arm setting with noisy positive outcomes (e.g., basket size in an online order). The outcome for each arm can be modelled using a log-Normal distribution with an unknown positive mean θ,

$$p(r_a|\theta_a) = \text{logNormal}(r_a; \log\theta_a - \sigma_y^2/2, \sigma_y^2) = \frac{1}{r_a\sigma_y\sqrt{2\pi}}\exp\left(-\frac{(\log\, r_a - \log\, \theta_a + \sigma_y^2/2)^2}{2\sigma_y^2}\right) \tag{23}$$

$$p(r_b|\theta_b) = \text{logNormal}(r_b; \log\theta_b - \sigma_y^2/2, \sigma_y^2) = \frac{1}{r_b\sigma_y\sqrt{2\pi}}\exp\left(-\frac{(\log\, r_b - \log\, \theta_b + \sigma_y^2/2)^2}{2\sigma_y^2}\right) \tag{24}$$

We place a log-Normal prior over θ,

$$p(\theta_a) = \text{logNormal}(\theta_a; \mu_o, \sigma_o^2) = \frac{1}{\theta_a\sigma_o\sqrt{2\pi}}\exp\left(-\frac{(\log\, \theta_a - \mu_o)^2}{2\sigma_o^2}\right) \tag{25}$$

$$p(\theta_b) = \text{logNormal}(\theta_b; \mu_o, \sigma_o^2) = \frac{1}{\theta_b\sigma_o\sqrt{2\pi}}\exp\left(-\frac{(\log\, \theta_b - \mu_o)^2}{2\sigma_o^2}\right) \tag{26}$$

Assume we have collected $\mathcal{D}_a = \{r_{a,i}\}_{i=1}^{N_a}$ and $\mathcal{D}_b = \{r_{b,i}\}_{i=1}^{N_b}$ for the two arms.
Solution:
The likelihood of observing the data $\mathcal{D}_a = \{r_{a,i}\}_{i=1}^{N_a}$ for arm a is:

$$p(\mathcal{D}_a|\theta_a) = \prod_{i=1}^{N} p(r_{a,i}|\theta_a)$$

Now we substitute the log-normal likelihood:

$$p(\mathcal{D}_a|\theta_a) = \prod_{i=1}^{N} \frac{1}{r_{a,i}\sigma_y\sqrt{2\pi}}\mathbf{exp}\left(-\frac{(\log\, r_{a,i} - \log\, \theta_a + \sigma_y^2/2)^2}{2\sigma_y^2}\right)$$

The prior for θ_a is:

$$p(\theta_a) = \frac{1}{\theta_a\sigma_o\sqrt{2\pi}}\exp\left(-\frac{(\log\, \theta_a - \mu_o)^2}{2\sigma_o^2}\right)$$

Now we just multiply the two two get the posterior distribution:

$$p(\theta_a|\mathcal{D}_a) \propto \prod_{i=1}^{N} \frac{1}{r_{a,i}\sigma_y\sqrt{2\pi}}\mathbf{exp}\left(-\frac{(\log\, r_{a,i} - \log\, \theta_a + \sigma_y^2/2)^2}{2\sigma_y^2}\right) \cdot \frac{1}{\theta_a\sigma_o\sqrt{2\pi}}\exp\left(-\frac{(\log\, \theta_a - \mu_o)^2}{2\sigma_o^2}\right) \tag{27}$$

$$\propto \frac{1}{\theta_a}\mathbf{exp}\left(-\frac{1}{2}\left(\sum_{i=1}^{N_a}\frac{(\log\, r_{a,i} - \log\, \theta_a + \sigma_y^2/2)^2}{\sigma_y^2} + \frac{(\log\, \theta_a - \mu_o)^2}{\sigma_o^2}\right)\right) \tag{28}$$

Let $S_a = \sum_{i=1}^{N_a} \log\, r_{a,i}$. The exponent becomes:

$$-\frac{1}{2}\left(\frac{N_a(\log\theta_a)^2 - 2\log\theta_a(S_a - N_a\sigma_y^2/2)}{\sigma_y^2} + \frac{(\log\theta_a)^2 - 2\log\theta_a\mu_o}{\sigma_o^2}\right) \tag{29}$$

$$-\frac{1}{2}\left(\left(\frac{N_a}{\sigma_y^2} + \frac{1}{\sigma_o^2}\right)(\log\theta_a)^2 - 2\log\theta_a\left(\frac{S_a - N_a\sigma_y^2/2}{\sigma_y^2} + \frac{\mu_o}{\sigma_o^2}\right)\right) \tag{30}$$

This is the quadratic form of a log-normal distribution. The posterior parameters are:

$$\mu_a' = \frac{\frac{S_a - N_a\sigma_y^2/2}{\sigma_y^2} + \frac{\mu_o}{\sigma_o^2}}{\frac{N_a}{\sigma_y^2} + \frac{1}{\sigma_o^2}}$$

$$\sigma_a'^2 = \frac{1}{\frac{N_a}{\sigma_y^2} + \frac{1}{\sigma_o^2}}$$

$\therefore$ the posterior distribution of θ_a is:

$$p(\theta_a|\mathcal{D}_a) = \text{logNormal}(\theta_a; \mu_a', \sigma_a'^2)$$

Similarly for θ_b:

$$p(\theta_b|\mathcal{D}_b) = \text{logNormal}(\theta_b; \mu_b', \sigma_b'^2)$$

4 Gaussian process with negative binomial likelihood

Assume that we have observed data $\mathcal{D} = \{(x_n, y_n)\}_{n=1}^N$, where $x_n \in \mathbb{R}^D$ and $y_n \in \{0, 1, 2, \dots\}$. We can model such data using a Gaussian process (GP) model with a negative binomial likelihood. That is, we have a GP prior over the underlying function, $f \sim \mathcal{GP}(0, k(\cdot, \cdot))$. Denoting , $f = [f(x_1), \dots, f(x_n)]$. By definition of GPs, we have $f \sim \mathcal{N}(0, \mathbf{K})$ where $\mathbf{K}$ is the covariance evaluated at the training points. We assume $\mathbf{K}$ to be invertible. We consider the following negative binomial likelihood:

$$p(y_n|f(x_n)) = \binom{y_n + r - 1}{y_n} \pi_n^{y_n}(1 - \pi_n)^r, \qquad i = 1. \dots, n$$

where $\pi_n = (1 + \exp(-f(x_n)))^{-1}$ and $r > 0$ is a fixed dispersion parameter.

4.1 Question 3.1 Find the log-posterior and gradient

Compute $\log p(\mathbf{f}|\mathcal{D})$ up to an additive constant and derive its gradient with respect to $\mathbf{f}$
Solution:

$$p(\mathbf{f}|\mathcal{D}) \propto p(\mathbf{f}) \cdot p(\mathcal{D}|\mathbf{f})$$

Taking the log would leave us with:

$$\log p(\mathbf{f}|\mathcal{D}) = \log p(\mathbf{f}) + \log p(\mathcal{D}|\mathbf{f})$$

we're given the following expressions for the likelihood and prior respectively:

$$\log p(\mathcal{D}|\mathbf{f}) = \sum_{n=1}^{N} \log \binom{y_n + r - 1}{y_n} + y_n \log \pi_n + r \log (1 - \pi_n)$$

$$\log p(\mathbf{f}) = -\frac{1}{2}\mathbf{f}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{f} - \frac{1}{2}\log|\mathbf{K}| - \frac{N}{2}\log 2\pi$$

Now we just add them up to get the log posterior:

$$p(\mathbf{f}|\mathcal{D}) = \sum_{n=1}^{N}(y_n \log \pi_n + r \log (1 - \pi_n)) + -\frac{1}{2}\mathbf{f}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{f} + \textbf{constant} \tag{31}$$

where *constant* contains all the terms that don't depend of $\mathbf{f}$

To calculate the gradient of the log-posterior, we can find the gradients of the likelihood and prior

individually and then add them together as such:

$$\frac{\partial \log\, p(\mathcal{D}|f)}{\partial f(x_n)} = \frac{\partial}{\partial f(x_n)} y_n \log \pi_n + \frac{\partial}{\partial f(x_n)} r \log(1-\pi_n) \tag{32}$$

$$= \frac{\partial}{\partial f(x_n)} y_n \log(1+\exp(-f(x_n)))^{-1}) + \frac{\partial}{\partial f(x_n)} r \log(1-(1+\exp(-f(x_n)))^{-1}) \tag{33}$$

$$= \frac{\partial}{\partial f(x_n)} y_n \log(\frac{1}{1+\exp(-f(x_n))}) + \frac{\partial}{\partial f(x_n)} r \log(1-\frac{1}{1+\exp(-f(x_n))}) \tag{34}$$

$$= \frac{\partial}{\partial f(x_n)} \log\left(\frac{1}{1+\exp(-f(x_n))}\right) + \frac{\partial}{\partial f(x_n)} \log\left(\frac{\exp(-f(x_n))}{1+\exp(-f(x_n))}\right) \tag{35}$$

$$= -\frac{\partial}{\partial f(x_n)} \log\left(1+\exp(-f(x_n))\right) + \frac{\partial}{\partial f(x_n)}\left(-f(x_n) - \log(1+\exp(-f(x_n)))\right) \tag{36}$$

$$= \frac{1}{1+\exp(f(x_n))} - 1 + \frac{\exp(-f(x_n))}{1+\exp(-f(x_n)} \tag{37}$$

$$= y_n(1-\pi_n) + r(-1+(1-\pi_n)) \Rightarrow y_n(1-\pi_n) - r\pi_n \tag{38}$$

For all observations the gradient would be:

$$\nabla_{\mathbf{f}}\, \mathrm{p}(\mathcal{D}|\mathbf{f}) = \begin{bmatrix} y_1(1-\pi_n) - r\pi_n \\ \vdots \\ y_N(1-\pi_N) - r\pi_N \end{bmatrix} = \mathbf{y} \odot (1-\pi) - r\pi$$

The gradient of the log-prior is:

$$\nabla \log\, p(\mathrm{f}) = \nabla(-\frac{1}{2}\mathbf{f}^{\mathrm{T}} K^{-1}\mathbf{f}) = -\frac{1}{2} 2\mathbf{K}^{-1}\mathbf{f} = \mathbf{K}^{-1}\mathbf{f}$$

Therefore the gradient of the log posterior is:

$$\nabla \mathbf{log}\; p(\mathbf{f}|\mathcal{D}) = -\mathbf{K}^{-1}\mathbf{f} + \mathbf{y} \odot (1-\pi) - r\pi$$

4.2 Question 3.2 Find the Hessian

Compute the Hessian matrix of the log-posterior and prove that it is negative definite
Solution:
To compute the Hessian of the log-posterior we can simply find the second derivatives of the log-prior and the log-likelihood and combine them together. We've computed the derivatives of the log-prior and log-likelihood to be:

$$\nabla_{\mathbf{f}}\, \log\, p(\mathrm{f}) = \mathbf{K}^{-1}\mathbf{f} \tag{39}$$

$$\nabla_{\mathbf{f}}\, \log\, p(\mathcal{D}|\mathrm{f}) = y_n(1-\pi_n) - r\pi_n \tag{40}$$

The second derivatives are:

$$\nabla^2_{\mathbf{f}}\, \log\, p(\mathrm{f}) = \nabla^2_{\mathbf{f}}\, \mathbf{K}^{-1}\mathbf{f} = -\mathbf{K}^{-1}$$

and:

$$\frac{\partial^2}{\partial f(x_n)^2} \log\, p(\mathcal{D}|\mathbf{f}) = \frac{\partial}{\partial f(x_n)}\; y_n(1-\pi_n) - r\pi_n = y_n\; \pi_n(1-\pi_n) - r\pi_n(1-\pi_n) = -(\pi_n(1-\pi_n)(y_n+r))$$

here we used the derivative of sigmoid: $\sigma(x)\cdot(1-\sigma(x))$ to get $\pi_n(1-\pi_n)$.
Now we combine the two terms to get the Hessian of the log-posterior:

$$\nabla^2_{\mathbf{f}}\, \log\, p(\mathbf{f}|\mathcal{D}) = -\mathbf{K}^{-1} - \mathbf{diag}((\pi_n(1-\pi_n)(y_n+r)))$$

$\mathbf{K}$ is positive definite[6] so $\mathbf{K}^{-1}$ is also positive definite. Therefore $-\mathbf{K}^{-1}$ is negative definite.
The log-likelihood for the negative binomial model is separable across observations where each term only depends of $f(x_n)$, making it a diagonal[5].
All the entries are positive as well since $\pi_n \in (0,1) \therefore \pi_n(1-\pi_n) > 0$. We also have $y_n > 0$ and $r > 0$ so $y_n + r > 0$. This leads to $(\pi_n(1-\pi_n)(y_n+r))$ being positive definite, so $-(\pi_n(1-\pi_n)(y_n+r))$ is negative definite. Thus the Hessian of the log-posterior is also negative definite as the sum of two negative definite matrices is negative definite.

4.3 Question 3.3 Laplace approximation

Implement a procedure to obtain the Laplace approximation based on the above results
Solution: [4] [1] [2]
To implement a Laplace approximation for a Gaussian Process model with a negative binomial likelihood, we'll have to approximate the posterior as a Gaussian centered at the **MAP estimate $\mathbf{f}_{\mathbf{MAP}}$**, with a covariance matrix given by the negative inverse Hessian of the log-posterior.

$$\mathbf{f}_{\text{MAP}} = \text{argmax} \log p(\mathbf{f}|\mathcal{D}).$$

Since the Hessian is negative definite of the log-posterior, we can use Newton-Raphson for optimization. At iteration t, using the Newton-Raphson update rule, we have:

$$\mathbf{f}^{(t+1)} = \mathbf{f}^{(t)} - \left(\nabla^2_{\mathbf{f}} \log p(\mathbf{f}|\mathcal{D})\right)^{-1} \nabla_{\mathbf{f}} \log p(\mathbf{f}|\mathcal{D}),$$

where the gradient and hessian are:

$$\nabla \mathbf{log}\ p(\mathbf{f}|\mathcal{D}) = -\mathbf{K}^{-1}\mathbf{f} + \mathbf{y} \odot (1-\pi) - r\pi \tag{41}$$

$$\nabla^2_{\mathbf{f}} \log p(\mathbf{f}|\mathcal{D}) = -\mathbf{K}^{-1} - \mathbf{W}, \quad \mathbf{W} = \mathbf{diag}((\pi_n(1-\pi_n)(y_n + r))). \tag{42}$$

This update rule moves f in the direction of steepest ascent, adjusting for curvature using the Hessian. Once $\mathbf{f}_{MAP}$ is found, we can approximate the posterior as:

$$p(\mathbf{f}|\mathcal{D}) \approx \mathcal{N}(\mathbf{f}|\mathbf{f}_{\text{MAP}}, \Sigma),$$

where $\Sigma = -\left(\nabla^2_{\text{f}} \log p(\mathbf{f}|\mathcal{D})\right)^{-1}$.
Invert the negative Hessian to get the covariance matrix Σ:

$$\Sigma = -\mathbf{H}^{-1} = (\mathbf{K}^{-1} + \mathbf{W})^{-1}.$$

python implementation:

```
import numpy as np
    def MAP(K, y, r, max_iter='iter', tol='tol'):
        N = len(y)
        f = np.zeros(N)
        for _ in range(max_iter):
            pi = 1 / (1 + np.exp(-f))
            grad = -K @ f + y * (1 - pi) - r * pi
            W = pi * (1 - pi) * (y + r)
            H = -np.linalg.inv(K) - np.diag(W)
            delta = np.lingalg.solve(H, grad) # Newton step
            f -= delta # update f
            if np.linalg.norm(delta) < tol:
                break
        return f

    def laplace_approx(K, y, r):
        f_map = MAP(K, y, r) # find the MAP
        pi_map = 1 / (1 + np.exp(-f_map))
        W_map = np.diag(pi_map * (1 - pi_map) * (y + r))
        H = -np.linalg.inv(K) - W_map
        Sigma = np.linalg.inv(-H)
        return f_map, Sigma
```

References

[1] Carl Edward Rasmussen Malte Kuss. Assessing approximations for gaussian process classification. *https://proceedings.neurips.cc/paper_files/paper/2005/file/3c333aadfc3ee8ecb8d77ee31197d96a − Paper.pdf*.

[2] Jarno Vanhatalo Marcelo Hartmann. Laplace approximation and natural gradient for gaussian process regression with heteroscedastic student-t model. *https://doi.org/10.1007/s11222-018-9836-0*.

[3] Wikipedia. Extended kalman filter. *Wikipedia*.

[4] Wikipedia. Laplace's approximation. *Wikipedia*.

[5] Wikipedia. Negative binomial distribution. *Wikipedia*.

[6] Wikipedia. Positive-definite kernel. *Wikipedia*.

[7] Bishop, Christopher M. Pattern Recognition and Machine Learning. New York :Springer, 2006.