Bayesian Inference and Data Assimilation

Prof. Dr.-Ing. Sebastian Reich

Universität Potsdam

12 April 2021

5 Bayesian inference

In this section, we define Bayesian inference, explain what it is used for and introduce some mathematical tools for applying it.

We are required to make inferences whenever we need to make a decision in light of incomplete information. Sometimes the information is incomplete because of partial measurements.

Also, sometimes the information is incomplete because of inaccurate measurements.

Incomplete information results in uncertainties which make decision-making difficult. However, often we have to make a decision anyway, despite the presence of uncertainty.

When making a decision/inference we have to decide how much to trust the incomplete information and how much to trust prior assumptions (assumptions based on previous experience before taking the measurement).

5.1 Inverse problems from a probabilistic perspective

Definition (Inference model)

Consider a random parameter vector $X : \Omega \to \mathcal{X} = \mathbb{R}^{N_x}$, and an observation noise variable $\Xi : \Omega \to \mathcal{Y} = \mathbb{R}^{N_y}$, both over a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Assume that X and Ξ are independent random variables. Then the observed variable $Y : \Omega \to \mathcal{Y} \in \mathbb{R}^{N_y}$ is a random variable defined by

$$Y = h(X) + \Xi, \tag{1}$$

where $h : \mathcal{X} \to \mathcal{Y}$ is a continuous map, called the forward map.

Example (linear model)

Assume that a scalar variable $z \in \mathbb{R}$ varies in time according to

$$z(t) = b_1 t + b_0. \tag{2}$$

Our goal is to estimate the parameters b_0 and b_1 from observations of $y_{\mathrm{obs}}(\mathrm{t}_k)$ taken at discrete times t_k, $k = 1, \ldots, N_{\mathrm{obs}}$. The measurement process can be described mathematically as follows:

$$y_{\mathrm{obs}}(\mathrm{t}_k) = b_1 \mathrm{t}_k + b_0 + \xi^k, \quad k = 1, \ldots, N_{\mathrm{obs}}.$$

where the measurement errors ξ^k are viewed as independent realisations of $\mathrm{N}(0, \sigma^2)$. This can be rewritten in the form of (1) with

$$Y = (Y(\mathrm{t}_1), Y(\mathrm{t}_2), \ldots, Y(\mathrm{t}_{N_{\mathrm{obs}}}))^{\mathrm{T}} \in \mathbb{R}^{N_{\mathrm{obs}}},$$

$x = (b_0, b_1)^T \in \mathbb{R}^2$,

$$h(x) = (b_0 + b_1 \mathrm{t}_1,\ b_0 + b_1 \mathrm{t}_2 \cdots, b_0 + b_1 \mathrm{t}_{N_{\mathrm{obs}}})^{\mathrm{T}},$$

and $\xi = (\xi^1, \ldots, \xi^{\mathrm{N}_{\mathrm{obs}}})^T$ is a realisation of $\Xi \sim \mathrm{N}(0, \sigma^2 I)$.

Theorem (Distribution of observation variable Y)

Assume that X and Ξ are absolutely continuous with PDFs π_X and π_Ξ respectively, and Y is related to X and Ξ via (1). Then Y is also absolutely continuous with PDF

$$\pi_Y(y) = \int_{\mathbb{R}^{N_x}} \pi_\Xi(y - h(x))\pi_X(x)\mathrm{d}x. \tag{3}$$

If X is a deterministic variable, i.e., $X(\omega) = x_0$ almost surely for an appropriate $x_0 \in \mathcal{X}$, then the PDF simplifies to

$$\pi_Y(y) = \pi_\Xi(y - h(x_0)).$$

Proof.

See Theorem 5.3 on page 133 of the textbook. □

Example (Gaussian distributions)

Consider the case in which X and Ξ are normal random variables with means $\bar{x}$ and zero, and covariance matrices P and R, respectively. The forward operator is given by $h(x) = Hx$, where H is an $N_x \times N_y$-dimensional matrix. Equation (3) then becomes

$$\pi_Y(y) \propto \int_{\mathbb{R}^{N_x}} \exp\left(-\frac{1}{2}(y - Hx)^T R^{-1}(y - Hx)\right) \times \\ \exp\left(-\frac{1}{2}(x - \bar{x})^T P^{-1}(x - \bar{x})\right) \mathrm{d}x,$$

where we have left out the constant of proportionality. Next we use the completing-the-square formula (see Example 5.4 on page 134) to obtain

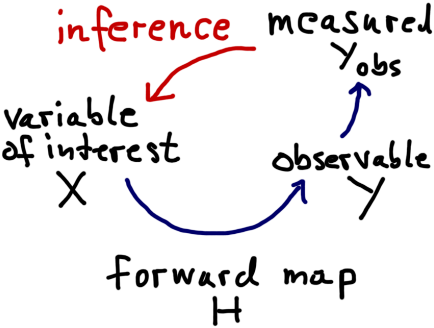
$$\pi_Y(y) \propto \exp\left(-\frac{1}{2}(y^T R^{-1} y - d^T C^{-1} d)\right),$$

where

$$C = P^{-1} + H^{\mathrm{T}} R^{-1} H, \quad d = H^{\mathrm{T}} R^{-1} y + P^{-1}\bar{x}.$$

Remark

We are actually more interested in the inverse inference problem in which we have obtained an observation which is modelled as a random observation variable $y \in \mathcal{Y}$, and we wish to infer the uncertainty in the parameters $x \in \mathcal{X}$. See the figure below for a schematic representation of the connection between the forward problem, an observation, and the inference problem.



Theorem (Bayes' Theorem)

Given a particular observation value $y_{\text{obs}} \in \mathbb{R}^{N_y}$, the conditional PDF $\pi_X(x|y_{\text{obs}})$ is given by Bayes' formula,

$$\pi_X(x|y_{\text{obs}}) = \frac{\pi_Y(y_{\text{obs}}|x)\pi_X(x)}{\pi_Y(y_{\text{obs}})}. \tag{4}$$

Proof.

See Theorem 5.5 on page 136 of the textbook. □

Remark (Bayes' Theorem)

Here π_X quantifies our uncertainty about the parameters X before observing y_{obs} (and hence we call it the prior PDF), whilst $\pi_X(x|y_{\text{obs}})$ quantifies our uncertainty after observing y_{obs} (and hence we call it the posterior PDF). The conditional PDF $\pi_Y(y|x)$ quantifies the likelihood of observing y given a particular value of x, and hence it is often called the likelihood function.

Example (Bayes for Gaussian distributions)

From the calculations in the previous example on Gaussian distributions we obtain

$$\begin{aligned}\pi_X(x|y) &\propto \exp\left(-\frac{1}{2}\left((y-Hx)^T R^{-1}(y-Hx) + (x-\bar{x})^T P^{-1}(x-\bar{x})\right)\right)\\ &\propto \exp\left(-\frac{1}{2}(x - C^{-1}d)^T C(x - C^{-1}d)\right),\end{aligned}$$

with C and d defined as before. This means that $\pi_X(x|y_{\rm obs}) = {\rm n}(x;\bar{x}^a, P^a)$ with covariance matrix $P^a = C^{-1} = (P^{-1} + H^T R^{-1} H)^{-1}$ and mean

$$\begin{aligned}\bar{x}^a &= C^{-1}d\\ &= \bar{x} - P^a H^T R^{-1}(H\bar{x} - y_{\rm obs}).\end{aligned}$$

Example (Likelihood for measurement model)

We introduced a (deterministic) measurement model in Section 1. In Section 2, we replaced it by the stochastic measurement model

$$Y = h(z_{\mathrm{ref}}(t)) + \Xi(t),$$

where $\Xi(t)$ is a Gaussian random variable with mean zero and variance R.

Following our previous discussion, with the state variable z now replacing the parameter variable x, this model gives rise to the conditional PDF (likelihood)

$$\pi_Y(y|z) = \frac{1}{\sqrt{2\pi R}} e^{-\frac{(y-h(z))^2}{2R}}.$$

Example (MLE and data assimilation)

A more complex case was considered later in Section 1:

$$
\begin{aligned}
y_{\mathrm{obs}}(\mathfrak{t}_1) &= H\psi(z^0) + \xi^1, \\
y_{\mathrm{obs}}(\mathfrak{t}_2) &= H\psi^2(z^0) + \xi^2, \\
&\vdots \\
y_{\mathrm{obs}}(\mathfrak{t}_{N_a}) &= H\psi^{N_a}(z^0) + \xi^{N_a},
\end{aligned}
$$

where the variables $\{\xi^k\}$ are realisations of independent and identically distributed Gaussian random variables with mean zero and variance R. We obtain the likelihood function

$$
\pi_{Y_{\mathfrak{t}_{1:N_a}}}(y^{\mathrm{obs}}_{\mathfrak{t}_{1:N_a}}|z) = \frac{1}{(2\pi R)^{N_a/2}} \Pi_{k=1}^{N_a} \exp\left(-\frac{(y_{\mathrm{obs}}(\mathfrak{t}_k) - H\psi^k(z))^2}{2R}\right), \qquad (5)
$$

where $y^{\mathrm{obs}}_{\mathfrak{t}_{1:N_a}} \in \mathbb{R}^{N_a}$ represents the values of all measurements $y_{\mathrm{obs}}(\mathfrak{t}_k) \in \mathbb{R}$ from $\mathfrak{t}_1$ to $\mathfrak{t}_{N_a}$. The nonlinear method of least squares can now be viewed as maximum likelihood estimator (MLE). Bayesian estimates will replace MLEs as a tool for performing data assimilation in Section 6.

Definition (Bayesian estimate)

Given a posterior PDF $\pi_X(x|y_{\rm obs})$ we define a Bayesian estimate $\hat{x} \in \mathcal{X}$ by

$$\hat{x} = \arg \min_{x' \in \mathcal{X}} \int \ell(x', x) \pi_X(x|y_{\rm obs}) {\rm d}x,$$

where $\ell(x', x)$ is an appropriate loss function. Popular choices include the maximum a posteriori (MAP) estimate with $\hat{x}$ corresponding to the modal value (global maximum) of $\pi_X(x|y_{\rm obs})$. The MAP estimate formally corresponds to the loss function

$$\ell_\varepsilon(x', x) = \left\{ \begin{array}{ll} 1 & \text{if } \|x' - x\| > \varepsilon, \\ 0 & \text{otherwise,} \end{array} \right.$$

in the limit $\varepsilon \to 0$. The posterior median estimate corresponds to $\ell(x', x) = \|x' - x\|$ while the posterior mean estimate,

$$\hat{x} = \int_{\mathcal{X}} x \pi_X(x|y_{\rm obs}) {\rm d}x,$$

results from $\ell(x', x) = \|x' - x\|^2$.

Remark (maximum likelihood estimate (MLE))

The MLE is formally obtained as a special case of the MAP estimate with the prior PDF (formally) set equal to a constant. This step is justified provided that

$$\int_{\mathbb{R}^{N_x}} \pi_Y(y|x)\mathrm{d}x < \infty.$$

This situation is referred to as a non-informative prior.

Remark (Estimators for Gaussian distributions)

Note that the MAP estimate, the posterior mean and the posterior median coincide for Gaussian random variables.

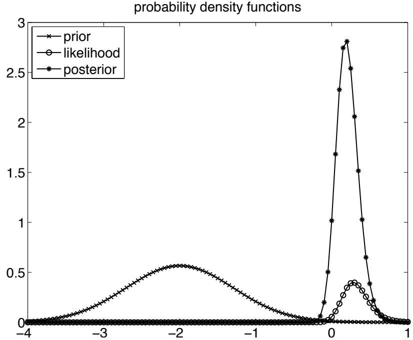
Example (Nonlinear forward map)

We consider a nonlinear forward map given by

$$y = h(x) = \frac{7}{12}x^3 - \frac{7}{2}x^2 + 8x.$$

The observed value is $y_{\rm obs} = 2$ and the measurement error is $\mathrm{N}(0,1)$. The prior is assumed to be normal with mean $\bar{x} = -2$ and variance $\sigma^2 = 1/2$.

We obtain numerically $\bar{x} \approx 0.2095$, a median of approximately 0.2006 and a MAP value of approximately 0.1837.



Example (Kalman)

Consider a scalar observation with $\Xi \sim N(0, R)$:

$$\pi_Y(y|x) = \pi_\Xi(y - h(x)) = \frac{1}{\sqrt{2\pi R}} e^{-\frac{1}{2R}(h(x)-y)^2}.$$

We also assume that $X \sim \mathrm{N}(\bar{x}, P)$ and that $h(x) = Hx$. Then, the posterior distribution is also Gaussian with mean

$$\begin{aligned} \bar{x}^a &= \bar{x} - P^a H^T R^{-1}(H\bar{x} - y_{\mathrm{obs}}), \\ &= \bar{x} - PH^T(HPH^T + R)^{-1}(H\bar{x} - y_{\mathrm{obs}}), \end{aligned} \tag{6}$$

and covariance matrix

$$\begin{aligned} P^a &= (P^{-1} + H^T R^{-1} H)^{-1}, \\ &= P - PH^T(HPH^T + R)^{-1} HP. \end{aligned} \tag{7}$$

These are the famous Kalman update formulas.

Remark (Tikhonov regularisation)

We note that $\bar{x}^a$ solves the minimisation problem

$$\bar{x}^a = \arg \min_{x \in \mathbb{R}^{N_x}} \left\{ \frac{1}{2}(x - \bar{x})^T P^{-1}(x - \bar{x}) + \frac{1}{2R}(Hx - y_{\rm obs})^2 \right\},$$

which can be viewed as a Tikhonov regularisation of the ill-posed inverse problem

$$y_{\rm obs} = Hx, \quad x \in \mathbb{R}^{N_x},$$

for $N_x > 1$.

A standard Tikhonov regularisation would use $\bar{x} = 0$ and $P^{-1} = \delta I$ with the regularisation parameter $\delta > 0$ appropriately chosen. In the Bayesian approach to inverse problems, the regularisation term is instead determined by the prior PDF π_X.

Example (Gaussian mixtures)

We extend the previous example to a Gaussian mixture prior on X:

$$\pi_X(x) = \sum_{j=1}^{J} \frac{\alpha_j}{(2\pi)^{N_x/2}|P_j|^{1/2}} \exp\left(-\frac{1}{2}(x-\bar{x}_j)^T P_j^{-1}(x-\bar{x}_j)\right) = \sum_{j=1}^{J} \alpha_j \,\mathrm{n}(x;\bar{x}_j,P_j),$$

where $\alpha_j > 0$, $j = 1,\ldots,J$, denote the mixture weights which sum to one. The posterior distribution is again a Gaussian mixture

$$\pi_X(x|y_{\rm obs}) = \sum_{j=1}^{J} \alpha_j^a \,\mathrm{n}(x;\bar{x}_j^a,P_j^a),$$

with

$$\bar{x}_j^a = \bar{x}_j - P_j H^T (HP_jH^T + R)^{-1}(H\bar{x}_j - y_{\rm obs}),$$

$$P_j^a = P_j - P_jH^T(HP_jH^T+R)^{-1}HP_j,$$

$$\alpha_j^a \propto \frac{\alpha_j}{\sqrt{2\pi(HP_jH^T+R)}} \exp\left(-\frac{(H\bar{x}_j - y_{\rm obs})^2}{2(HP_jH^T+R)}\right).$$

Remark (Radon-Nikodym derivative)

We mention that Bayes' formula must be replaced by the Radon-Nikodym derivative in cases where the prior distribution is not absolutely continuous with respect to the Lebesgue measure (or in case the space $\mathcal{X}$ does not admit a Lebesgue measure).

If μ is absolute continuous with respect to η with Radon-Nikodym derivative

$$\frac{\mathrm{d}\mu}{\mathrm{d}\eta} = f.$$

Then

$$\int_{\mathcal{X}} g(x)\mu(\mathrm{d}x) = \int_{\mathcal{X}} g(x)f(x)\eta(\mathrm{d}x)$$

for all sufficiently regular functions g.

Example (Empirical measures)

Consider as an example the case of an empirical measure μ_X, *i.e.*,

$$\mu_X(\mathrm{d}x) = \frac{1}{M}\sum_{i=1}^{M} \mu_{x_i}(\mathrm{d}x) = \frac{1}{M}\sum_{i=1}^{M} \delta(x - x_i)\mathrm{d}x.$$

Then the resulting posterior measure $\mu_X(\cdot|y_{\rm obs})$ is absolutely continuous with respect to μ_X and the associated Radon-Nikodym derivative is given by

$$\frac{\mathrm{d}\mu_X(x|y_{\rm obs})}{\mathrm{d}\mu_X(x)} \propto \pi_\Xi(y_{\rm obs} - h(x)).$$

The posterior measure is given by

$$\mu_X(\mathrm{d}x|y_{\rm obs}) = \sum_{i=1}^{M} w_i\, \mu_{x_i}(\mathrm{d}x) = \sum_{i=1}^{M} w_i \delta(x - x_i)\mathrm{d}x,$$

with weights $w_i \geq 0$ defined by $w_i \propto \pi_\Xi(h(x_i) - y_{\rm obs})$.

5.2 Sampling the posterior

We usually want to summarise our uncertainty, formally represented by the posterior PDF $\pi_X(x|y_{\rm obs})$, in terms of expectation values

$$\bar{g} = \int_{\mathcal{X}} g(x)\pi_X(x|y_{\rm obs}){\rm d}x,$$

where g could, for example, stand for the variance or correlation. Apart from a few special examples, these integrals are intractable, and it becomes necessary to use Monte Carlo methods to approximate them.

We introduce two alternative method for generating samples from a desired distribution π_X^*. The first method belongs to the general class of Markov chain Monte Carlo (MCMC) methods, while the second approach is based on Brownian dynamics.

Both approaches have in common that the lead to ergodic stochastic processes for which the posterior PDF, that is, $\pi_X^*(x) = \pi_X(x|y_{\rm obs})$ is an invariant measure invariant measures.

We first explain the idea of an MCMC method for a discrete state space $\mathcal{X} = \{a_1, a_2, \dots, a_M\}$. We will later return to $\mathcal{X} = \mathbb{R}^{N_x}$.

Definition (Discrete state space MCMC method)

Given a desired probability distribution $p^* \in \mathbb{R}^M$ and a symmetric stochastic transition matrix $\mathrm{P} \in \mathbb{R}^{M\times M}$ over a discrete state-space $\mathcal{X} = \{a_1, a_2, \dots, a_M\}$, the modified Markov chain with stochastic transition matrix $\tilde{\mathrm{P}}$ with entries

$$\tilde{p}_{ij} = (1 - c_j)\delta_{ij} + \alpha_{ij} p_{ij}, \quad \alpha_{ij} = 1 \wedge (p_i^*/p_j^*), \quad c_j = \sum_{i=1}^{M} \alpha_{ij} p_{ij} \tag{8}$$

gives rise to a MCMC method for sampling from p^*.

The coefficients $\alpha_{ij} \in [0,1]$ define the Metropolis accept/rejection criterion. Here δ_{ij} denotes the Kronecker symbol with values $\delta_{ii} = 1$ and $\delta_{ij} = 0$ for $i \neq j$ and

$$a \wedge b = \min\{a, b\}.$$

Remark (Detailed balance)

The invariance of p^* under (8), that is $\tilde{\mathrm{P}} p^* = p^*$, follows from the detailed balance condition

$$\tilde{p}_{ij} p_j^* = \tilde{p}_{ji} p_i^* \tag{9}$$

for all pairs $i, j = 1, \ldots, M$ since (9) implies

$$\sum_j \tilde{p}_{ij} p_j^* = \sum_j \tilde{p}_{ji} p_i^* = p_i^*.$$

Detailed balance is satisfied by (8) since $p_{ij} = p_{ji}$.

Remark (Implementation)

In terms of practical implementation of (8) we proceed as follows: If the chain is in state a_j, draw a proposal state a_i from $\mathcal{X}$ with probability p_{ij}. Next draw a uniform random number ξ and accept the proposal if $\alpha_{ij} > \xi$ otherwise remain in the current state a_j.

Example (MCMC on discrete state space)

Let us consider a model with three states $\{1, 2, 3\}$. We start from the Markov chain defined by the following process: If in state i jump to the other two state with equal probability $1/2$. If we follow this process for long enough, then all three states will be visited with equal probability $p = 1/3$. The Markov chain is characterised by the stochastic matrix

$$\mathrm{P} = \begin{pmatrix} 0 & 1/2 & 1/2 \\ 1/2 & 0 & 1/2 \\ 1/2 & 1/2 & 0 \end{pmatrix}$$

Let us now assume that we wish to sample the state 1 with probability $p_1^* = 1/2$ and the other two states with probability $p_i^* = 1/4$, $i = 2, 3$.

This can be achieved via MCMD as follows: If currently in state j, pick a proposal state i according to P. Accept this proposal with probability

$$\alpha = 1 \wedge (p_i^*/p_j^*) \tag{10}$$

otherwise remain in state j. The thus generated sequence of numbers will produce a histogram with relative frequencies approaching $(1/2, 1/4, 1/4)$.

We now return to the state space $\mathcal{X} = \mathbb{R}^{N_x}$ and formally extend the MCMC methodology to a method for sampling from a PDF π^*_X.

Definition (Random walk MCMC)

We consider the random walk proposal step

$$\hat{x} = x + \sqrt{2\gamma}\xi \tag{11}$$

where $x \in \mathbb{R}^{N_x}$ denotes the current state of the chain, $\gamma > 0$ is a parameter (step-size), and ξ is the realisation of a mean-zero Gaussian random variable in $\mathbb{R}^{N_x}$ with covariance $P = I$.
Accept the proposal $\hat{x}$ with probability

$$\alpha = 1 \wedge \exp(-U(\hat{x}) + U(x)) \tag{12}$$

otherwise remain in state x.
Here $U : \mathbb{R}^{N_x} \to \mathbb{R}$ is an appropriate potential defined by

$$\pi^*_X(x) = C^{-1}\exp(-U(x)), \qquad C = \int_{\mathbb{R}^{N_x}} \exp(-U(x))\mathrm{d}x, \tag{13}$$

provided $C < \infty$.

Remark (Choice of γ)

The choice of the step-size $\gamma > 0$ in RW MCMC is crucial; if chosen too small the samples move to little, if too large the rejection rate becomes too high.

Remark (Sampling the posterior)

The posterior PDF $\pi_X(x|y_{\rm obs})$ becomes the stationary distribution of random walk MCMC provided that the potential U is chosen according to

$$\begin{aligned} U(x) &= -\ln \pi_X(x|y_{\rm obs}) \\ &= -\ln \pi_X(x) - \ln \pi_Y(y_{\rm obs}|x) + \ln \pi_Y(y_{\rm obs}). \end{aligned}$$

Since the dynamics does not depend on a constant added or subtracted from $U(x)$, we may actually use

$$U(x) = -\ln \pi_X(x) - \ln \pi_Y(y_{\rm obs}|x), \tag{14}$$

which provides a huge simplification in practice since it avoids the computation of the evidence $\pi_Y(y_{\rm obs})$.

Definition (Brownian dynamics and canonical distribution)

We consider Brownian dynamics

$$\mathrm{d}X = -\nabla_x U(X)\mathrm{d}\tau + \sqrt{2}\mathrm{d}W, \tag{15}$$

where $U : \mathbb{R}^{N_x} \to \mathbb{R}$ is given by (14), the independent variable is denoted by τ, and $W(\tau)$ denotes standard Brownian motion.

Note that

$$\pi_X^*(x) = \pi_X(x|y_{\rm obs}) \propto \exp(-U(x)). \tag{16}$$

We also define the linear operator

$$\mathcal{L}\pi := \nabla_x \cdot (\pi \nabla_x U) + \nabla_x \cdot \nabla_x \pi, \tag{17}$$

and write the Fokker-Planck equation for Brownian dynamics in the abstract operator form

$$\frac{\partial \pi_X}{\partial \tau} = \mathcal{L}\pi_X. \tag{18}$$

Lemma (Stationary distribution)

The canonical distribution (16) satisfies $\mathcal{L}\pi_X^* = 0$ which implies that the canonical PDF is stationary under the associated Fokker-Planck equation.

Proof.

The identify

$$\mathcal{L}\pi_X^* = 0$$

can be verified by direct calculation. □

Remark (Gradient Flow)

Note that the Fokker-Planck equation can also be reformulated as a gradient flow towards the stationary solution π_X^*:

$$\begin{aligned}\frac{\partial \pi_X}{\partial \tau} &= \nabla_x \cdot (\pi_X \nabla_x U(x) + \nabla_x \pi_X), \\ &= \nabla_x \cdot \left(\pi_X \nabla_x \log\left(\frac{\pi_X}{\pi_X^*}\right)\right). \end{aligned} \tag{19}$$

Remark (Ergodicity)

If the Brownian dynamics is ergodic, then expectations from the posterior can be computed by using time integrals according to

$$\begin{aligned}\mathbb{E}[g(X)] &= \int_{\mathbb{R}^{N_x}} g(x)\pi_X^*(x)\mathrm{d}x \\ &= \lim_{T\to\infty} \frac{1}{T}\int_0^T g(x(\tau))\mathrm{d}\tau \approx \lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^N g(x^n).\end{aligned}$$

Here $\{x^n\}_{n\geq 0}$ denotes a particular realisation of Brownian dynamics and $\tau_n = n\,\Delta\tau$ with step-size $\Delta\tau > 0$ produced by the Euler-Maruyama method

$$x^{n+1} = x^n - \Delta\tau\, U'(x^n) + \sqrt{2\Delta\tau}\xi^n, \quad \xi^n \sim \mathrm{N}(0,1).$$

This is an example of a generalised Monte Carlo method, as defined in Chapter 3, with equal weights $w_i = 1/N$. It is generalised since the sample points $\{x^n\}_{n\geq 0}$ are not independent.

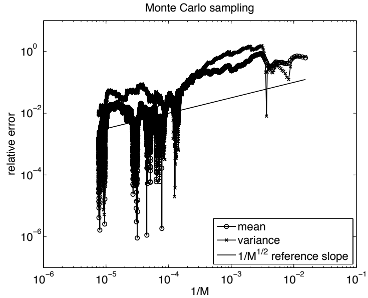
Example (Nonlinear forward map, continued)

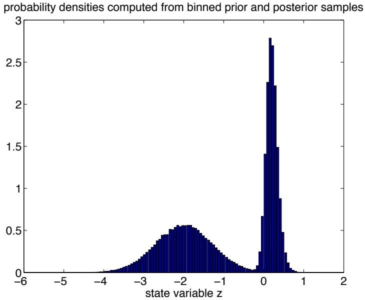
The posterior from the previous example leads to the potential

$$U(x) = (x+2)^2 + \frac{1}{2}(h(x)-2)^2.$$

We now conduct a long simulation with the Euler-Maruyama method:

$$x^{n+1} = x^n - \Delta\tau U'(x^n) + \sqrt{2\Delta\tau}\xi^n, \quad \xi^n \sim \mathrm{N}(0,1).$$





To determine whether solutions of the Fokker-Planck equation do indeed convergence to the stationary PDF, we need to study the spectral properties of $\mathcal{L}$.

Remark

The following material is not examinable for data science students.

To analyse the eigenvalues of $\mathcal{L}$, we introduce the weighted inner product

$$\langle \pi_1, \pi_2 \rangle_* = \int_{\mathbb{R}^{N_x}} \pi_X^*(x)^{-1} \, \pi_1(x) \, \pi_2(x) \, \mathrm{d}x$$

in the space of all integrable functions such that $\|\pi\|_* = \langle \pi, \pi \rangle_*^{1/2} < \infty$.

Since

$$
\begin{aligned}
\langle \mathcal{L}\pi_1, \pi_2\rangle_* &= \int_{\mathbb{R}^{N_x}} (\pi_X^*)^{-1}\pi_2 \nabla_x \cdot (\pi_1 \nabla_x U + \nabla_x \pi_1)\mathrm{d}x \\
&= -\int_{\mathbb{R}^{N_x}} \nabla_x((\pi_X^*)^{-1}\pi_2)\cdot (\pi_1 \nabla_x U + \nabla_x \pi_1)\mathrm{d}x \\
&= -\int_{\mathbb{R}^{N_x}} (\pi_X^*)^{-1}(\nabla_x \pi_2 + \pi_2 \nabla_x U)\cdot (\pi_1 \nabla_x U + \nabla_x \pi_1)\mathrm{d}x \\
&= -\int_{\mathbb{R}^{N_x}} \nabla_x((\pi_X^*)^{-1}\pi_1)\cdot (\pi_2 \nabla_x U + \nabla_x \pi_2)\mathrm{d}x \\
&= \int_{\mathbb{R}^{N_x}} (\pi_X^*)^{-1}\pi_1 \nabla_x \cdot (\pi_2 \nabla_x U + \nabla_x \pi_2)\mathrm{d}x = \langle \pi_1, \mathcal{L}\pi_2\rangle_*,
\end{aligned}
$$

we may conclude that $\mathcal{L}$ is self-adjoint with respect to the inner product $\langle\cdot,\cdot\rangle_*$. Hence the spectrum $\sigma(\mathcal{L})$ of $\mathcal{L}$ is on the real axis.

Since

$$\begin{aligned}
\langle \mathcal{L}\pi, \pi\rangle_* &= \int_{\mathbb{R}^{N_x}} (\pi_X^*)^{-1}\pi\nabla_x \cdot (\pi\nabla_x U + \nabla_x \pi)\mathrm{d}x,\\
&= -\int_{\mathbb{R}^{N_x}} (\pi_X^*)^{-1}\left(\nabla_x \pi + \pi\nabla_x U\right)\cdot\left(\pi\nabla_x U + \nabla_x \pi\right)\mathrm{d}x,\\
&= -\|\pi\nabla_x U + \nabla_x\pi\|_*^2,\\
&\le 0,
\end{aligned}$$

all eigenvalues of $\mathcal{L}$ have to be non-positive. We express this by writing $\sigma(\mathcal{L}) \subset \{\lambda \in \mathbb{R} : \lambda \le 0\}$.

Theorem (Spectral gap)

If (i) U is smooth, (ii) $\int_{\mathcal{X}} \exp(-U(x))\mathrm{d}x < \infty$, and (iii) there is a constant $c > 0$ such that the Hessian matrix, $D^2U(x)$, of second-order derivatives satisfies

$$v^T D^2 U(x) v \geq c\|v\|^2,$$

for all $v \in \mathbb{R}^{N_x}$ and all $x \in \mathbb{R}^{N_x}$, then the canonical PDF is the unique invariant density and

$$\sup[\sigma(\mathcal{L}) \setminus \{0\}] \leq -c. \tag{20}$$

Proof.

See textbook by Pavliotis for a derivation of this result. □

5.3 Optimal coupling approach to Bayesian inference

Given a prior or forecast random variable $X^f : \Omega \to \mathbb{R}^{N_x}$, we denote its PDF by $\pi_{X^f}(x)$, $x \in \mathbb{R}^{N_x}$, and consider the assimilation of an observed $y_{\rm obs} \in \mathbb{R}^{N_y}$ with likelihood function $\pi_Y(y|x)$. The posterior or analysis PDF is given by

$$\pi_{X^a}(x|y_{\rm obs}) = \frac{\pi_Y(y_{\rm obs}|x)\pi_{X^f}(x)}{\int_{\mathbb{R}^{N_x}} \pi_Y(y_{\rm obs}|x)\pi_{X^f}(x){\rm d}x}, \qquad (21)$$

according to Bayes' theorem.

Remark

The material of this section will not be examined but a basic understanding of the material will be helpful for understanding the sequential Monte Carlo method as introduced in the following Chapter 6.

Typically, the forecast random variable X^f and its PDF are not available explicitly. Instead we assume that an ensemble of forecasts $x_i^f \in \mathbb{R}^{N_x}$, $i = 1, \ldots, M$, is given, which are considered as realisations $X_i^f(\omega)$, $\omega \in \Omega$, of M independent (or dependent) random variables $X_i^f : \Omega \to \mathbb{R}^{N_x}$ with law π_{X^f}.

Applying the importance sampling technique, we obtain the following estimator for $\mathbb{E}[g(X^a)]$ with respect to the posterior PDF $\pi_{X^a}(x|y_{\rm obs})$ using the forecast ensemble:

$$\bar{g}_M^a = \sum_{i=1}^{M} w_i g(x_i^f),$$

with weights

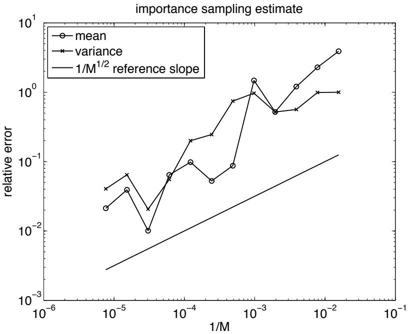
$$w_i = \frac{\pi_Y(y_{\rm obs}|x_i^f)}{\sum_{j=1}^M \pi_Y(y_{\rm obs}|x_j^f)}. \tag{22}$$

Example (Nonlinear forward map, continued)

We return to the previous example from this section. Instead of the Brownian dynamics sampling approach we now consider importance sampling by drawing M samples x_i from the Gaussian prior $\mathrm{N}(-2, 1/2)$ with posterior weights

$$w_i \propto e^{-\frac{1}{2}(h(x_i)-2)^2}.$$

Below we display the relative errors in the posterior mean and variance.



Instead of using weighted forecast samples, an alternative is to attempt to transform the samples $x_i^f = X_i^f(\omega)$ with $X_i^f \sim \pi_{X^f}$ into samples x_i^a from the posterior distribution $\pi_{X^a}(x|y_{\rm obs})$. Then we can use the estimator

$$\bar{g}_M^a = \frac{1}{M}\sum_{i=1}^{M} g(x_i^a),$$

with equal weights.

In other words, we are looking for a coupling between the prior and posterior PDFs as discussed in Chapter 2 and then again in Chapter 3 under resampling.

Recall that for univariate random variables $X_1 = X^f$ and $X_2 = X^a$ with PDFs π_{X^f} and π_{X^a} respectively, the transformation is characterised by

$$F_{X^a}(x_i^a) = F_{X^f}(x_i^f), \tag{23}$$

where F_{X^f} and F_{X^a} denote the cumulative distribution functions of X^f and X^a, respectively.

Equation (23) requires knowledge of the associated PDFs; the extension to multivariate random variables is non-trivial.

Instead, we propose an alternative approach which combines importance sampling with the idea of coupling of measures & optimal transportation. We essentially have done this already in Chapter 3 (resampling).

Remark (monomial resampling)

We have already discussed monomial resampling in Chapter 3, which can be used to generate posterior samples $\{x_i^a\}$ from weighted prior samples $\{x_i^f, w_i\}$.

Monomial resampling effectively defines a coupling between the two discrete random variables $X_M^f : \Omega \to \mathcal{X}_M$ and $X_M^a : \Omega \to \mathcal{X}_M$ with realisations in $\mathcal{X}_M = \{x_1^f, \dots, x_M^f\}$ and probability vector $p^f = (1/M, \dots, 1/M)^T$ for X_M^f and $p^a = (w_1, \dots, w_M)^T$ for X_M^a, respectively.

Here a coupling between p^f and p^a is an $M \times M$ matrix T with non-negative entries $t_{ij} = (T)_{ij} \geq 0$ such that

$$\sum_{i=1}^{M} t_{ij} = \frac{1}{M}, \qquad \sum_{j=1}^{M} t_{ij} = w_i. \tag{24}$$

Instead of defining a coupling T through monomial resampling, we seek the coupling T^* that minimises the expected Euclidean distance

$$\mathbb{E}[\|X_M^f - X_M^a\|^2] = \sum_{i,j=1}^{M} t_{ij}\|x_i^f - x_j^f\|^2. \tag{25}$$

The desired coupling T^* is obtained by solving a linear transport problem.

Since (24) leads to $2M-1$ independent constraints, the matrix T^* contains at most $2M-1$ non-zero entries.

Having computed T^*, the stochastic transition matrix $\mathrm{P} = M\,T^* \in \mathbb{R}^{M\times M}$ on $\mathcal{X}_M$ then has the property that the probability vectors p^f and p^a satisfy $p^a = \mathrm{P}p^f$.

Given a set of M realisations x_j^f, $j = 1, \ldots, M$, from the prior PDF and importance weights $w_i \propto \pi_Y(y_{\rm obs}|x_i^f)$, a Monte Carlo resampling step proceeds now as follows.

1. Compute the coupling matrix T^* which is optimal under the cost function (25) and define discrete random variables $\hat{X}_j^a$, $j = 1, \ldots, M$, with law

$$\hat{X}_j^a \sim \begin{pmatrix} p_{1j} \\ \vdots \\ p_{Mj} \end{pmatrix}. \tag{26}$$

   Here p_{ij} denotes the (i, j)th entry of $\mathrm{P} = MT^*$ and each column vector in P defines a probability vector, *i.e.*, $\sum_{i=1}^M p_{ij} = 1$.

2. An analysis ensemble $\{x_j^a\}$ of size M is obtained by collecting a single realisation from each random variable $\hat{X}_j^a$, *i.e.*, $x_j^a := \hat{X}_j^a(\omega)$ for $j = 1, \ldots, M$. This ensemble of equally weighted samples allows for the approximation of expectation values with respect to the posterior distribution $\pi_{X^a}(x|y_{\rm obs})$.

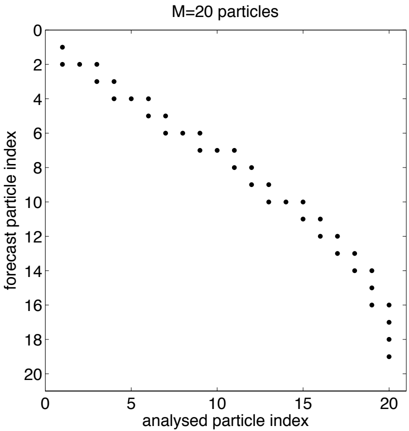


Figure: Non-zero entries in the matrix P for $M = 20$, which indicate the support of the coupling. Here the univariate samples x_i^{f} have been sorted. There are a total of $2M - 1 = 39$ non-zero entries in P. The banded structure reveals the spatial locality and the cyclical monotonicity of the resampling step and indicates the graph of an underlying deterministic transformation.

The previous figure suggests a further modification where we replace the random resampling step by a linear transformation.

The modification is based on the observation that the expectation values of the random variables (26) are given by

$$\bar{x}_j^a = \mathbb{E}[\hat{X}_j^a] = \sum_{i=1}^{M} x_i^f p_{ij}. \tag{27}$$

We use this result to propose the deterministic transformation

$$x_j^a := \bar{x}_j^a = \sum_{i=1}^{M} x_i^f p_{ij}, \tag{28}$$

$j = 1, \ldots, M$.

Example (transform method)

We take the univariate Gaussian with mean $\bar{x} = 1$ and variance $\sigma^2 = 2$ as the PDF for the prior random variable X^f. Realisations of X^f are generated using

$$x_i^f = 1 + 2\mathrm{erf}^{-1}(2u_i - 1), \quad u_i = \frac{1}{2M} + \frac{i-1}{M}$$

for $i = 1, \ldots, M$. The likelihood function is

$$\pi_Y(y_{\mathrm{obs}}|x) = \frac{1}{\sqrt{4\pi}} \exp\left(\frac{-(y_{\mathrm{obs}} - x)^2}{4}\right)$$

with assumed observed value $y_{\mathrm{obs}} = 0.1$. Bayes' formula yields a posterior distribution which is Gaussian with mean $\bar{x} = 0.55$ and variance $\sigma^2 = 1$.

Example (continued)

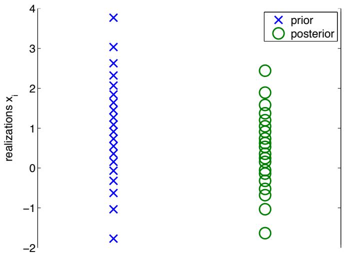


Figure: Prior x_i^f and posterior x_i^a realisations from the transform method for $M = 20$.

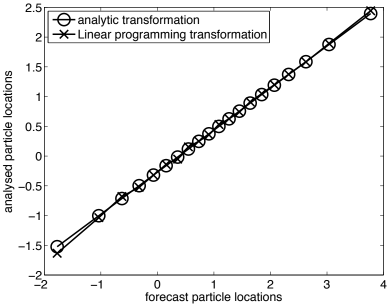


Figure: Exact and numerical ensemble transform map for $M = 20$. The Gaussian case leads to the exact transformation being linear. The numerical approximation deviates from linearity mostly in its both tails.

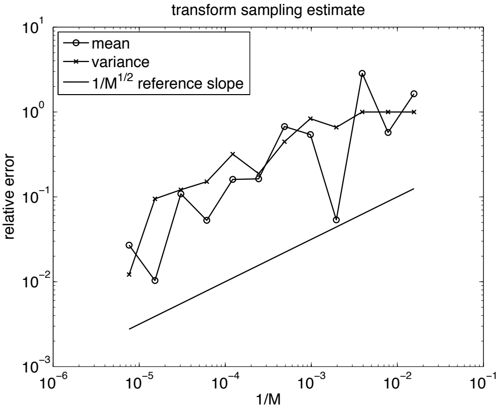
Example (continued)

Table: Estimated posterior first to fourth-order moments from the ensemble transform method applied to a Gaussian scalar Bayesian inference problem. We observe first-order convergence in $1/M$!

	$\bar{x}$	σ^2	$\mathbb{E}[(X-\bar{x})^3]$	$\mathbb{E}[(X-\bar{x})^4]$
$M = 10$	0.5361	1.0898	-0.0137	2.3205
$M = 40$	0.5473	1.0241	0.0058	2.7954
$M = 100$	0.5493	1.0098	-0.0037	2.9167

Example (Nonlinear forward map, continued)

We return to the problem set out earlier in this section. Instead of estimating posterior expectation values from the samples x_i^f with weights w_i, we now apply the linear transform method in order to generate posterior samples x_i^a with uniform weights $w_i = 1/M$. The resulting estimates for the posterior mean and variance are displayed below.


transform sampling estimate
mean
variance
1/M^{1/2} reference slope
relative error
10^1
10^0
10^{-1}
10^{-2}
10^{-3}
10^{-6}
10^{-5}
10^{-4}
10^{-3}
10^{-2}
10^{-1}
1/M