MATH 076 - Computational Inverse Problems

Carter Kruse, August 9, 2023

Homework 4

Instructions

For the non-computational questions, you are **not** permitted to use a calculator or any other computational tools, unless it is to check your work. For the computational questions, you are asked to use a coding language of your choice to perform the tasks requested. You may make use any code posted to our course's Canvas page. **Please show all of your work.** If you have any questions or uncertainties, please reach out to your instructor.

The label [**CQ**] indicates a computational question that is to be solved using a computational tool such as MATLAB or Python.

Questions

1) Consider the statistical inverse problem

$$\begin{bmatrix} 2 \\ 4 \end{bmatrix} = \begin{bmatrix} 1 & 7 \\ 2 & 3 \end{bmatrix} X + E$$

where $E \in \mathbb{R}^2$ is Gaussian with mean 0 and covariance matrix

$$\Gamma_{\text{noise}} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$

Suppose you know the prior density on X to be

$$X \sim \mathcal{N}\left(0, 2\mathbb{I}_2\right)$$

where $\mathcal{N}\left(\mu, \Gamma\right)$ indicates a normal density with mean μ and covariance Γ.

(a) What is the posterior density function?

(b) Compute the MAP estimate of the posterior density function.

(c) [**CQ**] Write code to generate samples from the posterior density function and plot your results. (View *Matlab* Code)

The following comes from *Statistical & Computational Inverse Problems*, Section 3.4. For further background on Gaussian densities, please see the appendix.

[Continued On Next Page]

Gaussian Densities: Let $x_0 \in \mathbb{R}^n$ and $\Gamma \in \mathbb{R}^{n \times n}$ be a symmetric positive definite matrix, denoted by $\Gamma > 0$. A Gaussian n-variate random variable X with mean x_0 and covariance Γ is a random variable with the probability density

$$\pi\left(x\right) = \left(\frac{1}{2\pi\left|\Gamma\right|}\right)^{n/2} \exp\Bigg\{\left(-\frac{1}{2}\left(x - x_0\right)^T \Gamma^{-1}\left(x - x_0\right)\right)\Bigg\}$$

where $\left|\Gamma\right| = \det\left(\Gamma\right)$. In such case, we use the notation

$$X \sim \mathcal{N}\left(x_0, \Gamma\right)$$

In this section, our aim is to derive closed formulas for the conditional means and covariances of Gaussian random variables. We start by recalling some elementary matrix properties.

Schur Complements: Let

$$\Gamma = \begin{bmatrix} \Gamma_{11} & \Gamma_{12} \\ \Gamma_{21} & \Gamma_{22} \end{bmatrix} \in \mathbb{R}^{n \times n}$$

be a positive definite symmetric matrix, where $\Gamma_{11} \in \mathbb{R}^{k \times k}$, $\Gamma_{22} \in \mathbb{R}^{(n-k)\times(n-k)}$, $k < n$, and $\Gamma_{21} = \Gamma_{12}^T$. We define the *Schur complements* $\tilde{\Gamma}_{jj}$ of $\Gamma_{jj}, j = 1, 2$ by the formulas

$$\tilde{\Gamma}_{22} = \Gamma_{11} - \Gamma_{12}\Gamma_{22}^{-1}\Gamma_{21} \qquad \tilde{\Gamma}_{11} = \Gamma_{22} - \Gamma_{21} - \Gamma_{11}^{-1}\Gamma_{12}$$

Observe that since the positive definiteness of Γ implies that $\Gamma_{jj}, j = 1, 2$, are also positive definite, the Schur complements are well defined.

Schur complements play an important role in calculating conditional covariances. We have the following matrix inversion lemma.

Matrix Inversion Lemma: Let Γ be a matrix satisfying the assumptions of the previous definition. Then the Schur complements $\tilde{\Gamma}_{jj}$ are invertible matrices and

$$\Gamma^{-1} = \begin{bmatrix} \tilde{\Gamma}_{22}^{-1} & -\tilde{\Gamma}_{22}^{-1}\Gamma_{12}\Gamma 22^{-1} \\ -\tilde{\Gamma}_{11}^{-1}\Gamma_{21}\Gamma 11^{-1} & \tilde{\Gamma}_{11}^{-1} \end{bmatrix}$$

Observe that since Γ is a symmetric matrix, so is Γ^{-1}, and thus by setting the off-diagonal blocks of Γ^{-1} equal up to transpose, we obtain the identity

$$\Gamma_{22}^{-1}\Gamma_{21}\tilde{\Gamma}_{22}^{-1} = \tilde{\Gamma}_{11}^{-1}\Gamma_{21}\Gamma_{11}^{-1}$$

We are now ready to prove the following result concerning the conditional probability densities of Gaussian random variables.

[Continued On Next Page]

Posterior Probability Density: Assume that $X : \Omega \to \mathbb{R}^n$ and $E : \Omega \to \mathbb{R}^m$ are mutually independent Gaussian random variables,

$$X \sim \mathcal{N}\left(x_0, \Gamma_{\text{prior}}\right) \qquad E \sim \mathcal{N}\left(e_0, \Gamma_{\text{noise}}\right)$$

and $\Gamma_{\text{prior}} \in \mathbb{R}^{n\times n}$ and $\Gamma_{\text{noise}} \in \mathbb{R}^{k\times k}$ are positive definite. Assume further that we have a linear model $Y = AX + E$ for a noisy measurement Y, where $A \in \mathbb{R}^{k\times n}$ is a known matrix. Then the posterior probability density of X given the measurement $Y = y$ is

$$\pi\left(x|y\right) \propto \exp\left\{\left(-\frac{1}{2}\left(x-\overline{x}\right)^T \Gamma_{\text{posterior}}^{-1}\left(x-\overline{x}\right)\right)\right\}$$

where

$$\overline{x} = x_0 + \Gamma_{\text{prior}} A^T \left(A\Gamma_{\text{prior}}A^T + \Gamma_{\text{noise}}\right)^{-1}\left(y - Ax_0 - e_0\right)$$

and

$$\Gamma_{\text{posterior}} = \Gamma_{\text{prior}} - \Gamma_{\text{prior}}A^T\left(A\Gamma_{\text{prior}}A^T + \Gamma_{\text{noise}}\right)^{-1} A\Gamma_{\text{prior}}$$

Remark: As mentioned, the posterior density can be derived directly from the Bayes' formula by arranging the quadratic form of the exponent according to the degree of x. Such a procedure gives an alternative representation for the posterior covariance matrix

$$\Gamma_{\text{posterior}} = \left(\Gamma_{\text{prior}}^{-1} + A^T\Gamma_{\text{noise}}^{-1}A\right)^{-1}$$

Furthermore, the posterior mean can be written as

$$\overline{x} = \left(\Gamma_{\text{prior}}^{-1} + A^T\Gamma_{\text{noise}}^{-1}A\right)^{-1}\left(A^T\Gamma_{\text{noise}}^{-1}\left(y - e_0\right) + \Gamma_{\text{prior}}^{-1}x_0\right)$$

The equivalence of formulas can be shown with a tedious matrix manipulation, and is sometimes referred to as the *matrix inversion lemma.* Note that there are several different formulas which are referred to as matrix inversion lemmas.

The practical feasibility of the different forms depends on the dimensions n and m and also on the existence of the matrices Γ_{prior} and Γ_{noise} or their respective inverses. Especially, if the standard smoothness priors are employed, Γ_{prior} does not exist and the form cannot be used.

Remark: In the purely Gaussian case, the center point $\overline{x}$ is simultaneously the maximum *a posteriori* estimator and the conditional mean, that is

$$\overline{x} = x_{CM} = c_{MAP}$$

Similarly, $\Gamma_{\text{posterior}}$ is the conditional covariance. Observe that in the sense of quadratic forms

$$\Gamma_{\text{posterior}} \leq \Gamma_{\text{prior}}$$

that is, the matrix $\Gamma_{\text{prior}} - \Gamma_{\text{posterior}}$ is positive semi-definite. Since the covariance matrix of a Gaussian probability density expresses the width of the density, this inequality means that *a measurement can never increase the uncertainty.*

In this problem, consider the statistical inverse problem $Y = AX + E$:

$$\begin{bmatrix} 2 \\ 4 \end{bmatrix} = \begin{bmatrix} 1 & 7 \\ 2 & 3 \end{bmatrix} X + E$$

We are provided the following information: $E \in \mathbb{R}^2$ is Gaussian with mean $e_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ and covariance matrix

$$\Gamma_{\text{noise}} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$

Further, the prior density on X is given. $X \in \mathbb{R}^2$ is Gaussian with mean $x_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ and covariance matrix

$$\Gamma_{\text{prior}} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

(This is determined from the expression $\mathcal{N}(\mu, \Gamma)$, which indicates a normal density with mean μ and covariance Γ.)

The variables/parameters $\overline{x}$ and $\Gamma_{\text{posterior}}$ are calculated as follows:

$$\overline{x} = x_0 + \Gamma_{\text{prior}} A^T \left(A \Gamma_{\text{prior}} A^T + \Gamma_{\text{noise}} \right)^{-1} (y - A x_0 - e_0)$$

$$\overline{x} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} 1 & 7 \\ 2 & 3 \end{bmatrix}^T \left(\begin{bmatrix} 1 & 7 \\ 2 & 3 \end{bmatrix} \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} 1 & 7 \\ 2 & 3 \end{bmatrix}^T + \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix} \right)^{-1} \left(\begin{bmatrix} 2 \\ 4 \end{bmatrix} - \begin{bmatrix} 1 & 7 \\ 2 & 3 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \end{bmatrix} \right)$$

$$\overline{x} = \begin{bmatrix} 1.7353 \\ 0.0319 \end{bmatrix}$$

$$\Gamma_{\text{posterior}} = \Gamma_{\text{prior}} - \Gamma_{\text{prior}} A^T \left(A \Gamma_{\text{prior}} A^T + \Gamma_{\text{noise}} \right)^{-1} A \Gamma_{\text{prior}}$$

$$\Gamma_{\text{posterior}} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} - \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} 1 & 7 \\ 2 & 3 \end{bmatrix}^T \left(\begin{bmatrix} 1 & 7 \\ 2 & 3 \end{bmatrix} \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} 1 & 7 \\ 2 & 3 \end{bmatrix}^T + \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 7 \\ 2 & 3 \end{bmatrix} \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

$$\Gamma_{\text{posterior}} = \begin{bmatrix} 0.2647 & -0.0319 \\ -0.0319 & 0.0239 \end{bmatrix}$$

Thus, with this information, the posterior density function is given as follows:

$$\pi(x|y) \propto \exp\left\{ \left(-\frac{1}{2} (x - \overline{x})^T \Gamma_{\text{posterior}}^{-1} (x - \overline{x}) \right) \right\}$$

$$\pi(x|y) \propto \exp\left\{ \left(-\frac{1}{2} \left(x - \begin{bmatrix} 1.7353 \\ 0.0319 \end{bmatrix} \right)^T \begin{bmatrix} 0.2647 & -0.0319 \\ -0.0319 & 0.0239 \end{bmatrix}^{-1} \left(x - \begin{bmatrix} 1.7353 \\ 0.0319 \end{bmatrix} \right) \right) \right\}$$

The MAP (maximum *a posteriori*) estimate of the posterior density function is equal to the center point $\overline{x}$ in the purely Gaussian case, as we have here. That is,

$$\overline{x} = x_{MAP} = \begin{bmatrix} 1.7353 \\ 0.0319 \end{bmatrix}$$

2) Let $X \in \mathbb{R}^2$ be a random variable with probability density given by

$$\pi(\boldsymbol{x}) \propto \exp\{\left(-\left|x_1^2 + x_2^2 - 5\right|\right)\}$$

(a) Describe a technique for generating samples from π.

(b) **[CQ]** Write code to generate samples from π and plot your results. (View *Matlab* Code)

The following comes from *Statistical & Computational Inverse Problems*, Section 3.6. For further background on Markov Chain Monte Carlo (MCMC) methods, please see the appendix.

Metropolis–Hastings Construction of the Kernel: Let μ denote the target probability distribution in $\mathbb{R}^n$ that we want to explore by the sampling algorithm. To avoid measure-theoretic notions, we assume that μ is absolutely continuous with respect to the Lebesgue measure, $\mu(dx) = \pi(x)\ dx$. We wish to determine a transition kernel $P(x, B)$ such that μ is its invariant measure.

Let P denote any transition kernel. When a point $x \in \mathbb{R}^n$ is given, we can postulate that the kernel either proposes a move to another point $y \in \mathbb{R}^n$ or it proposes no move away from x. This allows us to split the kernel into two parts,

$$P(x, B) = \int_B K(x, y)\ dy + r(x)\, \chi_B(x)$$

where χ_B is the characteristic function of the set $B \in \mathcal{B}$. Although $K(x, y) \geq 0$ is actually a density, we may think of $K(x, y)\ dy$ as the probability of the move from x to the infinitesimal set dy at y while $r(x) \geq 0$ is the probability of x remaining inert. The characteristic function χ_B of B appears since if $x \notin B$, the only way for x of reaching B is through a move.

The condition $P(x, \mathbb{R}^n) = 1$ implies that

$$r(x) = 1 - \int_{\mathbb{R}^n} K(x, y)\ dy$$

In order for $\pi(x)\ dx$ to be an invariant measure of P, we must have the identity

$$\begin{aligned}\mu P(B) &= \int_{\mathbb{R}^n} \left(\int_B K(x, y)\, dy + r(x)\, \chi_B(x)\right) \pi(x)\, dx \\ &= \int_B \left(\int_{\mathbb{R}^n} \pi(x)\, K(x, y)\, dx + r(y)\, \pi(y)\right) dy \\ &= \int_B \pi(y)\, dy\end{aligned}$$

for all $B \in \mathcal{B}$, implying that

$$\pi(y)\,(1 - r(y)) = \int_{\mathbb{R}^n} \pi(x)\, K(x, y)\, dx$$

By the formula above, this is tantamount to

$$\int_{\mathbb{R}^n} \pi(y)\, K(y, x)\, dx = \int_{\mathbb{R}^n} \pi(x) K(x, y)\, dx$$

This condition is called the *balance equation.* In particular, if K satisfies the *detailed balance equation*, which is given by

$$\pi(y)\, K(y, x) = \pi(x)\, K(x, y)$$

for all pairs $x, y \in \mathbb{R}^n$, then the balance equation holds *a fortiori.* These conditions constitute the starting point in constructing the Markov chain transition kernels used for stochastic sampling.

In the Metropolis–Hastings algorithm, the aim is to construct a transition kernel K that satisfies the detailed balance equation.

Let $q : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$ be a given function with the property $\int q(x, y)\, dy = 1$. The kernel q is called the *proposal distribution* or *candidate-generating kernel* for reasons explained later. Such a function q defines a transition kernel,

$$Q(x, A) = \int_A q(x, y)\, dy$$

If q happens to satisfy the detailed balance equation, we set simply $K(x, y) = q(x, y), r(x) = 0$ and we are done. Otherwise, we correct the kernel by a multiplicative factor and define

$$K(x, y) = \alpha(x, y)\, q(x, y)$$

where α is a correction term to be determined. Assume that for some $x, y \in \mathbb{R}^n$, instead of the detailed balance we have

$$\pi(y)\, q(y, x) < \pi(x)\, q(x, y)$$

Our aim is to choose α so that

$$\pi(y)\, \alpha(y, x) = \pi(x)\, \alpha(x, y)\, q(x, y)$$

This is achieved if we set

$$\alpha(y, x) = 1 \qquad \alpha(x, y) = \frac{\pi(y)\, q(y, x)}{\pi(x)\, q(x, y)} < 1$$

By reversing x and y, we see that the kernel K defined satisfies the *detailed balance equation* if we define

$$\alpha(x, y) = \min\left(1, \frac{\pi(y)\, q(y, x)}{\pi(x)\, q(x, y)}\right)$$

This transition kernel is called the Metropolis–Hastings kernel.

The above derivation does not shed much light on how to implement the method. Fortunately, the algorithm turns out to be relatively simple. Usually, it is carried out in practice through the following steps.

1) Pick the initial value $x_1 \in \mathbb{R}^n$ and set $k = 1$.
2) Draw $y \in \mathbb{R}^n$ from the proposal distribution $q(x_k, y)$ and calculate the acceptance ratio

$$\alpha(x_k, y) = \min\left(1, \frac{\pi(y)\, q(y, x_k)}{\pi(x_k)\, q(x_k, y)}\right)$$

3) Draw $t \in [0, 1]$ from uniform probability density.
4) If $\alpha(x_k, y) \geq t$, set $x_{k+1} = y$, else $x_{k+1} = x_k$. When $k = K$, the desired sample size, stop, else increase $k \to k + 1$ and go to step 2.

Before presenting some examples, a couple of remarks are in order. First, if the candidate-generating kernel is symmetric, that is,

$$q\left(x, y\right) = q\left(y, x\right)$$

for all $x, y \in \mathbb{R}^n$, then the acceptance ratio α simplifies to

$$\alpha\left(x, y\right) = \min\left(1, \frac{\pi\left(y\right)}{\pi\left(x\right)}\right)$$

Hence, we accept immediately moves that go towards higher probability and sometimes also moves that take us to lower probabilities. In particular, the symmetry condition is satisfied if the proposal distribution corresponds to a *random walk model*, that is,

$$q\left(x, y\right) = g\left(x - y\right)$$

for some non-negative even function $g : \mathbb{R}^n \to \mathbb{R}_+$, that is, $g\left(x\right) = g\left(-x\right)$.

An important but difficult issue is the stopping criterion, that is, how to decide when the sample is large enough. This question, as well as the convergence issues in general, will be discussed later in the light of examples.

In this problem, we test the Metropolis-Hastings algorithm with a low-dimensional density. Consider $X \in \mathbb{R}^2$, a random variable with probability density given by

$$\pi\left(\boldsymbol{x}\right) \propto \exp\left\{\left(-\left|x_1^2 + x_2^2 - 5\right|\right)\right\}$$

We construct a Metropolis-Hastings sequence using the random walk proposal distribution to explore this density. Let us define

$$q\left(x, y\right) \propto \exp\left\{\left(-\frac{1}{2\gamma^2}\left\|x - y\right\|^2\right)\right\}$$

In other words, we assume that the random step from x to y is distributed as white noise,

$$W = Y - X \sim \mathcal{N}\left(0, \gamma^2 I\right)$$

This choice of the proposal distribution gives rise to the following updating scheme. This is the technique used for generating samples from π.

```
Algorithm
Pick initial value $x_1$. Set $x = x_1$.
For $k = 2 : K$ do
Calculate $\pi\left(x\right)$.
Draw $w \sim \mathcal{N}\left(0, \gamma^2 I\right)$, set $y = x + w$.
Calculate $\pi\left(y\right)$.
Calculate $\alpha\left(x, y\right) = \min\left(1, \pi\left(y\right)/\pi\left(x\right)\right)$.
Draw $u \sim U\left(\left[0, 1\right]\right)$.
If $u < \alpha\left(x, y\right)$, accept. Set $x = y, x_k = x$.
Else, reject. Set $x_k = x$.
End
```

3) Let $X, Y \in \mathbb{R}^n$ be random variables with joint Gaussian density defined by

$$\pi\left(x,y\right) \propto \exp\left\{\left(-\frac{1}{2}\begin{bmatrix} x-3 \\ y-2 \end{bmatrix}^T \begin{bmatrix} 4 & 2 \\ 2 & 3 \end{bmatrix}^{-1} \begin{bmatrix} x-3 \\ y-2 \end{bmatrix}\right)\right\}$$

(a) Write the conditional probability densities, up to a normalization constant, of X conditioned on $Y = y$ and of Y conditioned on $X = x$.

(b) **[CQ]** Write code that generates samples of π using a Gibbs sampling technique. Plot your results. (View *Matlab* Code)

The following comes from *Statistical & Computational Inverse Problems*, Section 3.4. For further background on Gaussian densities, please see the appendix.

Gaussian Densities: Let $x_0 \in \mathbb{R}^n$ and $\Gamma \in \mathbb{R}^{n\times n}$ be a symmetric positive definite matrix, denoted by $\Gamma > 0$. A Gaussian n-variate random variable X with mean x_0 and covariance Γ is a random variable with the probability density

$$\pi\left(x\right) = \left(\frac{1}{2\pi\left|\Gamma\right|}\right)^{n/2} \exp\left\{\left(-\frac{1}{2}\left(x-x_0\right)^T \Gamma^{-1}\left(x-x_0\right)\right)\right\}$$

where $|\Gamma| = \det\left(\Gamma\right)$. In such case, we use the notation

$$X \sim \mathcal{N}\left(x_0, \Gamma\right)$$

In this section, our aim is to derive closed formulas for the conditional means and covariances of Gaussian random variables. We start by recalling some elementary matrix properties.

Schur Complements: Let

$$\Gamma = \begin{bmatrix} \Gamma_{11} & \Gamma_{12} \\ \Gamma_{21} & \Gamma_{22} \end{bmatrix} \in \mathbb{R}^{n\times n}$$

be a positive definite symmetric matrix, where $\Gamma_{11} \in \mathbb{R}^{k\times k}$, $\Gamma_{22} \in \mathbb{R}^{(n-k)\times(n-k)}$, $k < n$, and $\Gamma_{21} = \Gamma_{12}^T$. We define the *Schur complements* $\tilde{\Gamma}_{jj}$ of $\Gamma_{jj}, j = 1,2$ by the formulas

$$\tilde{\Gamma}_{22} = \Gamma_{11} - \Gamma_{12}\Gamma_{22}^{-1}\Gamma_{21} \qquad \tilde{\Gamma}_{11} = \Gamma_{22} - \Gamma_{21} - \Gamma_{11}^{-1}\Gamma_{12}$$

Observe that since the positive definiteness of Γ implies that $\Gamma_{jj}, j = 1,2$, are also positive definite, the Schur complements are well defined.

Schur complements play an important role in calculating conditional covariances. We have the following matrix inversion lemma.

[Continued On Next Page]

Matrix Inversion Lemma: Let Γ be a matrix satisfying the assumptions of the previous definition. Then the Schur complements $\tilde{\Gamma}_{jj}$ are invertible matrices and

$$\Gamma^{-1} = \begin{bmatrix} \tilde{\Gamma}_{22}^{-1} & -\tilde{\Gamma}_{22}^{-1}\Gamma_{12}\Gamma 22^{-1} \\ -\tilde{\Gamma}_{11}^{-1}\Gamma_{21}\Gamma 11^{-1} & \tilde{\Gamma}_{11}^{-1} \end{bmatrix}$$

Observe that since Γ is a symmetric matrix, so is Γ^{-1}, and thus by setting the off-diagonal blocks of Γ^{-1} equal up to transpose, we obtain the identity

$$\Gamma_{22}^{-1}\Gamma_{21}\tilde{\Gamma}_{22}^{-1} = \tilde{\Gamma}_{11}^{-1}\Gamma_{21}\Gamma_{11}^{-1}$$

We are now ready to prove the following result concerning the conditional probability densities of Gaussian random variables.

Joint Gaussian Probability Density: Let $X : \Omega \to \mathbb{R}^n$ and $Y : \Omega \to \mathbb{R}^k$ be two Gaussian random variables whose joint probability density $\pi : \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R}_+$ is of the form

$$\pi\left(x, y\right) \propto \exp\left\{\left(-\frac{1}{2}\begin{bmatrix} x - x_0 \\ y - y_0 \end{bmatrix}^T \begin{bmatrix} \Gamma_{11} & \Gamma_{12} \\ \Gamma_{21} & \Gamma_{22} \end{bmatrix}^{-1} \begin{bmatrix} x - x_0 \\ y - y_0 \end{bmatrix}\right)\right\}$$

Then the probability distribution of X conditioned on $Y = y$, $\pi\left(x|y\right) : \mathbb{R}^n \to \mathbb{R}_+$ is of the form

$$\pi\left(x|y\right) \propto \exp\left\{\left(-\frac{1}{2}\left(x - \overline{x}\right)^T \tilde{\Gamma}_{22}^{-1}\left(x - \overline{x}\right)\right)\right\}$$

where

$$\overline{x} = x_0 + \Gamma_{12}\Gamma_{22}^{-1}\left(y - y_0\right)$$

Proof: By shifting the coordinate origin to $[x_0; y_0]$, we may assume that $x_0 = 0$ and $y_0 = 0$. By Bayes' formula, we have $\pi\left(x|y\right) \propto \pi\left(x, y\right)$, so we consider the joint probability density as a function of x. By the matrix inversion lemma and the associated identity, we have

$$\pi\left(x, y\right) \propto \exp\left\{\left(-\frac{1}{2}\left(x^T\tilde{\Gamma}_{22}^{-1}x - 2x^T\tilde{\Gamma}_{22}^{-1}\Gamma_{12}\Gamma_{22}^{-1}y + y^T\tilde{\Gamma}_{11}^{-1}y\right)\right)\right\}$$

Further, by completing the quadratic form in the exponential into squares, we can express the joint distribution as

$$\pi\left(x, y\right) \propto \exp\left\{\left(-\frac{1}{2}\left(\left(x - \Gamma_{12}\Gamma_{22}^{-1}y\right)^T \tilde{\Gamma}_{22}^{-1}\left(x - \Gamma_{12}\Gamma_{22}^{-1}y\right) + c\right)\right)\right\}$$

where

$$c = y^T\left(\tilde{\Gamma}_{11}^{-1} - \Gamma_{22}^{-1}\Gamma_{21}\tilde{\Gamma}_{22}^{-1}\Gamma_{12}\Gamma_{22}^{-1}\right)y$$

is mutually independent of x and thus can be factored out of the density. This completes the proof.

In this problem, let $X, Y \in \mathbb{R}^n$ be random variables with joint Gaussian density defined by

$$\pi\left(x, y\right) \propto \exp\left\{\left(-\frac{1}{2}\begin{bmatrix} x - 3 \\ y - 2 \end{bmatrix}^T \begin{bmatrix} 4 & 2 \\ 2 & 3 \end{bmatrix}^{-1} \begin{bmatrix} x - 3 \\ y - 2 \end{bmatrix}\right)\right\}$$

The conditional probability densities, up to a normalization constant, of X conditioned on $Y = y$ and of Y conditioned on $X = x$ are given as follows:

[Continued On Next Page]

The probability distribution of X conditioned on $Y = y$ is

$$\pi\left(x|y\right) \propto \exp\left\{\left(-\frac{1}{2}\left(x-\overline{x}\right)^{T}\tilde{\Gamma}_{22}^{-1}\left(x-\overline{x}\right)\right)\right\}$$

where

$$\overline{x} = x_0 + \Gamma_{12}\Gamma_{22}^{-1}\left(y - y_0\right) \text{ and } \tilde{\Gamma}_{22} = \Gamma_{11} - \Gamma_{12}\Gamma_{22}^{-1}\Gamma_{21}$$

Given the values $\begin{bmatrix} x_0 \\ y_0 \end{bmatrix} = \begin{bmatrix} 3 \\ 2 \end{bmatrix}$ and $\Gamma = \begin{bmatrix} \Gamma_{11} & \Gamma_{12} \\ \Gamma_{21} & \Gamma_{22} \end{bmatrix} = \begin{bmatrix} 4 & 2 \\ 2 & 3 \end{bmatrix}$, the following computation is true.

$\pi\left(x \mid y\right) \propto \exp\left\{\left(-\frac{1}{2}\left(x-\overline{x}\right)^{T}\tilde{\Gamma}_{22}^{-1}\left(x-\overline{x}\right)\right)\right\}$	$\overline{x} = \left(3\right) + \left(2\right)\left(3\right)^{-1}\left(y - \left(2\right)\right)$	$\tilde{\Gamma}_{22} = \left(4\right) - \left(2\right)\left(3\right)^{-1}\left(2\right)$
$\pi\left(x \mid y\right) \propto \exp\left\{\left(-\frac{1}{2}\left(x-\overline{x}\right)^{T}\tilde{\Gamma}_{22}^{-1}\left(x-\overline{x}\right)\right)\right\}$	$\overline{x} = 3 + \frac{2}{3}\left(y - 2\right)$	$\tilde{\Gamma}_{22} = 4 - \left(\frac{4}{3}\right)$
$\pi\left(x \mid y\right) \propto \exp\left\{\left(-\frac{1}{2}\left(x-\overline{x}\right)^{T}\tilde{\Gamma}_{22}^{-1}\left(x-\overline{x}\right)\right)\right\}$	$\overline{x} = \frac{2}{3}y + \frac{5}{3}$	$\tilde{\Gamma}_{22} = \frac{8}{3}$

Thus,

$$\pi\left(x|y\right) \propto \exp\left\{\left(-\frac{1}{2}\left(x - \left(\frac{2}{3}y + \frac{5}{3}\right)\right)^{T}\left(\frac{3}{8}\right)\left(x - \left(\frac{2}{3}y + \frac{5}{3}\right)\right)\right)\right\}$$

which simplifies to

$$\pi\left(x|y\right) \propto \exp\left\{\left(-\frac{3}{16}\left(x - \frac{2}{3}y - \frac{5}{3}\right)^{T}\left(x - \frac{2}{3}y - \frac{5}{3}\right)\right)\right\}$$

Similarly, the probability distribution of Y conditioned on $X = x$ is

$$\pi\left(y|x\right) \propto \exp\left\{\left(-\frac{1}{2}\left(y-\overline{y}\right)^{T}\tilde{\Gamma}_{11}^{-1}\left(y-\overline{y}\right)\right)\right\}$$

where

$$\overline{y} = y_0 + \Gamma_{21}\Gamma_{11}^{-1}\left(x - x_0\right) \text{ and } \tilde{\Gamma}_{11} = \Gamma_{22} - \Gamma_{21}\Gamma_{11}^{-1}\Gamma_{12}$$

This follows from the circumstance of *symmetry*, which allows us to "swap" x and y, along with the associated Γ matrix values (and Shur complements).

Given the values $\begin{bmatrix} x_0 \\ y_0 \end{bmatrix} = \begin{bmatrix} 3 \\ 2 \end{bmatrix}$ and $\Gamma = \begin{bmatrix} \Gamma_{11} & \Gamma_{12} \\ \Gamma_{21} & \Gamma_{22} \end{bmatrix} = \begin{bmatrix} 4 & 2 \\ 2 & 3 \end{bmatrix}$, the following computation is true.

$\pi\left(y \mid x\right) \propto \exp\left\{\left(-\frac{1}{2}\left(y-\overline{y}\right)^{T}\tilde{\Gamma}_{11}^{-1}\left(y-\overline{y}\right)\right)\right\}$	$\overline{y} = \left(2\right) + \left(2\right)\left(4\right)^{-1}\left(x - \left(3\right)\right)$	$\tilde{\Gamma}_{11} = \left(3\right) - \left(2\right)\left(4\right)^{-1}\left(2\right)$
$\pi\left(y \mid x\right) \propto \exp\left\{\left(-\frac{1}{2}\left(y-\overline{y}\right)^{T}\tilde{\Gamma}_{11}^{-1}\left(y-\overline{y}\right)\right)\right\}$	$\overline{y} = 2 + \frac{1}{2}\left(x - 3\right)$	$\tilde{\Gamma}_{11} = 3 - 1$
$\pi\left(y \mid x\right) \propto \exp\left\{\left(-\frac{1}{2}\left(y-\overline{y}\right)^{T}\tilde{\Gamma}_{11}^{-1}\left(y-\overline{y}\right)\right)\right\}$	$\overline{y} = \frac{1}{2}x + \frac{1}{2}$	$\tilde{\Gamma}_{11} = 2$

Thus,

$$\pi\left(y|x\right) \propto \exp\left\{\left(-\frac{1}{2}\left(y - \left(\frac{1}{2}x + \frac{1}{2}\right)\right)^{T}\left(\frac{1}{2}\right)\left(y - \left(\frac{1}{2}x + \frac{1}{2}\right)\right)\right)\right\}$$

which simplifies to

$$\pi\left(y|x\right) \propto \exp\left\{\left(-\frac{1}{4}\left(y - \frac{1}{2}x - \frac{1}{2}\right)^{T}\left(y - \frac{1}{2}x - \frac{1}{2}\right)\right)\right\}$$

4) Consider the linear inverse problem $\boldsymbol{y} = A\boldsymbol{x} + \varepsilon$, where $\boldsymbol{y}$ is data, $\boldsymbol{x}$ is the unknown of interest, and $\boldsymbol{\varepsilon}$ is noise. Answer each of the following questions thoughtfully:

(a) Describe scenarios when the solution to the linear least squares problem

$$\boldsymbol{x}_{LS} = \arg\min_{\boldsymbol{x}} \|A\boldsymbol{x} - \boldsymbol{y}\|_2^2$$

does not provide a meaningful solution.

When considering the linear inverse problem $\boldsymbol{y} = A\boldsymbol{x} + \varepsilon$, where $\boldsymbol{y}$ is data, $\boldsymbol{x}$ is the unknown of interest, and $\boldsymbol{\varepsilon}$ is noise, the linear least squares problem

$$\boldsymbol{x}_{LS} = \arg\min_{\boldsymbol{x}} \|A\boldsymbol{x} - \boldsymbol{y}\|_2^2$$

seeks to find the value of $\boldsymbol{x}$ for which the sum of squared differences (between the observed data $\boldsymbol{y}$ and the product $A\boldsymbol{x}$) is minimized. There are various scenarios when the solution obtained from the linear least squares method does not provide a meaningful solution, as we have seen in this class.

Ill-Conditioned Matrix: In particular, if the matrix A is ill-conditioned (nearly singular), small changes in the data $\boldsymbol{y}$, or noise $\boldsymbol{\varepsilon}$ may result in significant changes in the solution $\boldsymbol{x}$. This is a result of lack of "stability" in the solution vector, i.e. the solution is highly sensitive to perturbations, and thus, may not accurately represent the true underlying signal.

Rank Deficient Matrix: Further, if the matrix A is not full rank, there may be multiple solutions $\boldsymbol{x}$ that fit the data $\boldsymbol{y}$, given the forward operator A, combined with the noise $\boldsymbol{\varepsilon}$. In these cases, the least squares solution may identify a solution that is not close to the true (underlying) solution.

(b) When the least squares solution fails to be adequate, what are other methods (not using statistical inversion theory) for obtaining meaningful solutions?

Ill-Posed Problems: Computational inverse problems deal with those where the output is known (with errors), though either the input or system is unknown. Typically, these problems belong to a class of *ill-posed* problems. A linear problem is well-posed if it satisfies three conditions.

- Existence: The problem has a solution.
- Uniqueness: The problem has at most one solution.
- Stability: The solution depends *continuously* on the data.

A problem is *ill-posed* if it fails any of these. Typically, we consider lack of "stability", which involves the consequence that arbitrarily small perturbations of the data can produce arbitrarily large perturbations of the solution. The way we approach these problems is to reformulate (stabilize/regularize the problem).

The reason for lack of "stability" in a linear system $A\boldsymbol{x} = \boldsymbol{b}$ is that A is *ill-conditioned* (nearly singular), meaning the problem is effectively *under-determined.* This allows for small changes to the residual despite changes of a (nearly) null vector the vector solution.

Regularization: As such, we modify the problem so that the new solution is more stable. For example, we can enforce an upper bound δ on the norm of the solution.

$$\min_{\boldsymbol{x}} ||A\boldsymbol{x} - \boldsymbol{b}||_2 \text{ subject to } ||\boldsymbol{x}||_k < \delta$$

The solution x_δ depends in a unique, but *non-linear* way on δ.

When the least squares solution fails to be adequate, there are various other (regularization) methods (not using statistical inversion theory) for obtaining meaningful solutions, as follows.

Truncated SVD: Truncated SVD is a regularization technique used to approximate a matrix A by retaining only a subset of its most important singular values and corresponding singular vectors. The singular value decomposition is given as follows.

$$\boldsymbol{x} = \frac{\boldsymbol{u}_i^T \boldsymbol{b}}{\sigma_i} \boldsymbol{v}_i$$

The truncated SVD only considers the first k values of i. To determine the value of k, we use the *Picard* log plot, demonstrated later in this paper. When σ_i is small, the contribution to $\boldsymbol{x}$ can be very large.

This regularization method is particularly useful when dealing with high-dimensional data or when trying to reduce the dimensionality while preserving significant information from the original signal. Further, the method is used to approximate a matrix with lower-rank, which is useful in removing noise from data (as in this case).

Tikhonov/Ridge (L2) Regularization: Tikhonov/Ridge (L2) regularization is a technique used to stabilize and regularize ill-posed problems, by adding a regularization term to control the solution's sensitivity to noise/numerical instability.

Given the linear inverse problem $A\boldsymbol{x} = \boldsymbol{b}$, were A is a given matrix and b is observed, issues are encountered when the matrix A is ill-conditioned, as the solution is sensitive to small perturbations in the data. To address this, we introduce a regularization term that encourages a solution with the desirable property of *smoothness.* This solution is found by minimizing the following objective function:

$$\boldsymbol{x}_\lambda = \arg\min_{\boldsymbol{x}} \|A\boldsymbol{x} - \boldsymbol{b}\|^2 + \lambda^2 \|\boldsymbol{x}\|_2^2$$

The value λ is the regularization parameter that controls the trade-off between fitting the data and the regularization term. Tikhonov/Ridge (L2) regularization penalizes solutions with large magnitudes by adding a term that considers the square of the solution's norm, thus encouraging solutions that are close fitting to the data while adhering to the regularization properties (smallness).

The value of λ determines the strength of the regularization, and is typically determined using the L-curve for Tikhonov/Ridge (L2) regularization. The choice of this λ highlights the trade-off between fitting the data and regularization.

Lasso (L1) Regularization: Lasso (L1) regularization is a further technique used to stabilize and regularize ill-posed problems, by adding a regularization term to control the solution's sensitivity to noise/numerical instability.

Given the linear inverse problem $A\boldsymbol{x} = \boldsymbol{b}$, were A is a given matrix and b is observed, issues are encountered when the matrix A is ill-conditioned, as the solution is sensitive to small perturbations in the data. To address this, we introduce a regularization term that encourages a solution with the desirable property of *sparsity.* This solution is found by minimizing the following objective function:

$$\boldsymbol{x}_\lambda = \arg\min_{\boldsymbol{x}} \|A\boldsymbol{x} - \boldsymbol{b}\|^2 + \lambda \|\boldsymbol{x}\|_1$$

The value λ is the regularization parameter that controls the trade-off between fitting the data and the regularization term. Lasso (L1) regularization penalizes solutions that are not sparse by adding a term that considers this, thus encouraging solutions that are close fitting to the data while adhering to the regularization properties (sparsity).

The value of λ determines the strength of the regularization, and the choice of this λ highlights the trade-off between fitting the data and regularization.

Typically, a sparsifying transform matrix L is used, which acts as a derivative operator on $\boldsymbol{x}$. To compute the $L1$ solution, we may use the matrix L as follows. Let $\boldsymbol{s}$ in $\mathbb{R}^n$ such that $\boldsymbol{x} = L^{-1}\boldsymbol{s}$. We find the solution $\boldsymbol{s}_{L1}$ to the following $L1$ regularization problem:

$$\boldsymbol{s}_{L1} = \min_{\boldsymbol{s}} \left\| AL^{-1}\boldsymbol{s} - \boldsymbol{b} \right\|_2^2 + \lambda \|\boldsymbol{s}\|_1$$

Then, we compute $\boldsymbol{x}_{L1} = L^{-1}\boldsymbol{s}_{L1}$. The value $\boldsymbol{s}_{L1}$ is sparse, and $\boldsymbol{x}_{L1}$ is recovered from this.

Principle Component Analysis (PCA): By transforming the data into a lower-dimensional space (capturing the most significant variations), we are able to reconstruct a signal with a reduced set of "principle components". This allows the solution to be more robust to noise.

(c) Compared with the classical techniques you described above, what are the benefits of using statistical approaches to solving linear inverse problems?

Compared with the classical (regularization) techniques described above, there are various benefits of using statistical approaches to solving linear inverse problems, as follows.

Prior Information: Statistical approaches allow us to incorporate prior information (knowledge or assumptions) about a given problem, particularly with respect to the distribution of variables. This allows us to develop solutions that may be more representative of the real-world.

Uncertainty: Using statistical methods provides a framework for considering and quantifying uncertainty associated with solutions. This may be helpful in determining confidence levels/intervals, alongside characterizing the reliability of a certain method/solution. This is particularly useful considering noise or uncertain data.

Numerical Methods: Statistical (Bayesian) techniques allow us to develop and implement numerical methods (including Markov Chain Monte Carlo and Gibbs Sampling) that computers are able to compute with relatively high speed.

Model Selection: By selecting the most appropriate model (alongside various parameters), we are able to tailor our approach to statistical inverse problems, with regularization included. This allows us to make decisions regarding a solution's trade-offs with regularization and over-fitting.

Overall, statistical approaches are useful over classical techniques when solving linear inverse problems, particularly when there is prior information that we can include. In this sense, a statistical method allows for a framework that addresses the challenges posed by the instability of linear inverse problems, resulting in more accurate/meaningful solutions.

(d) Was there a topic you enjoyed learning about the most this term? If so, what was it?

Yes! While I enjoyed learning about the regularization techniques in the first section of the course, the most engaging aspects were the topics covered near the end of the course, specifically the Metropolis-Hasting and Gibbs Sampling algorithms.

These Markov Chain Monte Carlo methods were interesting to consider, given that they allow us to explore/visualize a posterior distribution with just a little information about the prior distributions

Appendix

Gaussian Densities - Remark: In the literature, Gaussian random variables are often defined through the Fourier transform, or *characteristic function* as it is called in probability theory, as follows. A random variable X is Gaussian if

$$E\left\{\exp\left\{\left(-i\zeta^T X\right)\right\}\right\} = \exp\left\{\left(-i\zeta^T x_0 - \frac{1}{2}\zeta^T \Gamma \zeta\right)\right\}$$

where $x_0 \in \mathbb{R}^n$ and $\Gamma \in \mathbb{R}^{n\times n}$ is positive semi-definite, that is, $\Gamma \geq 0$. In this definition, it is not necessary to require that Γ is positive definite. The definition of a Gaussian density is not more restrictive in the following sense. Let $d_1 \geq d_2 \geq \cdots \geq d_n$ denote the eigenvalues of the matrix Γ, and let $\{v_1, \ldots, v_n\}$ be the corresponding eigenbasis. Assume that the p first eigenvalues are positive and $d_{p+1} = \cdots = d_n = 0$. If X is a random variable with the above property, one can prove that

$$X - x_0 \in \operatorname{span}\{v_1, \ldots, v_p\}$$

almost certainly. To see this, observe that for any vector u perpendicular to the eigenvectors $v_j, 1 \leq j \leq p$,

$$E\left\{\left(u^T\left(X - x_0\right)\right)^2\right\} = u^T \Gamma u = 0$$

Therefore, by defining the orthogonal projection

$$P_p : \mathbb{R}^n \to \operatorname{span}\{v_1, \ldots, v_p\} \approx \mathbb{R}^p$$

the random variable $P_p X$ is a p-variate Gaussian random variable in the sense of the previous definition with mean $x_0' = P_p x_0$ and covariance $\Gamma' = P_p^T \Gamma P_p$. Using the notation $x = [P_p x; (1 - P_p)x] = [x'; x'']$ and $x_0 = [P_p x_0; (1 - P_p)x_0] = [x_0'; x_0'']$ we have in this case

$$\pi(x) = \left(\frac{1}{2\pi\left|\Gamma'\right|}\right)^{\frac{p}{2}} \exp\left\{\left(-\frac{1}{2}\left(x' - x_0'\right)^T \left(\Gamma'\right)^{-1}\left(x' - x_0'\right)\right)\right\}\delta\left(x'' - x_0''\right)$$

where δ denotes the Dirac delta in $\mathbb{R}^{n-p}$.

Schur Complements - Proof: Consider the determinant of Γ

$$|\Gamma| = \left| \begin{bmatrix} \Gamma_{11} & \Gamma_{12} \\ \Gamma_{21} & \Gamma_{22} \end{bmatrix} \right| \neq 0$$

By subtracting from the second row from the first one multiplied by $\Gamma_{21}\Gamma_{11}^{-1}$ from the left we find that

$$|\Gamma| = \left| \begin{bmatrix} \Gamma_{11} & \Gamma_{12} \\ 0 & \Gamma_{22} - \Gamma_{11}^{-1}\Gamma_{12} \end{bmatrix} \right| = |\Gamma_{11}| \left|\tilde{\Gamma}_{11}\right|$$

implying that $\left|\tilde{\Gamma}_{11}\right| \neq 0$. Similarly we can prove that $\tilde{\Gamma}_{22}$ is invertible. This is referred to as the Schur identity.

The proof follows from Gaussian elimination. Consider the linear system

$$\begin{bmatrix} \Gamma_{11} & \Gamma_{12} \\ \Gamma_{21} & \Gamma_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$$

By eliminating x_2 from the second equation we get

$$x_2 = \Gamma_{22}^{-1} \left(y_2 - \Gamma_{21} x_1 \right)$$

and substituting into the first equation we obtain

$$\left(\Gamma_{11} - \Gamma_{12}\Gamma_{22}^{-1}\Gamma_{21} \right) x_1 = y_1 - \Gamma_{12}\Gamma_{22}^{-1} y_2$$

or, equivalently

$$x_1 = \tilde{\Gamma}_{22}^{-1} y_1 - \tilde{\Gamma}_{22}^{-1}\Gamma_{12}\Gamma_{22}^{-1} y_2$$

Similarly we solve the second equation for x_2 and the claim follows.

Markov Chain Monte Carlo Methods: The abstract definition of the solution of an inverse problem as the posterior probability distribution is not very helpful in practice, if we have no means of exploring it. Various random sampling methods have been proposed, and we discuss here an effective class of these, known as the Markov Chain Monte Carlo techniques.

The maximum *a posteriori* estimate leads to an optimization problem, while the conditional mean and conditional covariances require integration over the space $\mathbb{R}^n$ where the posterior density is defined. It is clear that if the dimension n of the parameter space $\mathbb{R}^n$ is large, the use of numerical quadrature rules is out of the question. An m-point rule for each direction would require m^n integration points, exceeding the computational capacity of most computers. Another problem with quadrature rules is that they require a relatively good knowledge of the support of the probability distribution, which is usually part of the information that we are actually looking for.

An alternative way to look at the problem is the following. Instead of evaluating the probability density at given points, let the density itself determine a set of points, a sample, that supports well the distribution. These sample points can then be used for approximate integration. The MCMC methods are at least on the conceptual level relatively simple algorithms to generate sample ensembles for Monte Carlo integration.

Monte Carlo Integration: Before going into detailed analysis, we discuss first the basic idea behind Monte Carlo integration.

Let μ denote a probability measure over $\mathbb{R}^n$. Further, let f be a scalar or vector-valued measurable function, integrable over $\mathbb{R}^n$ with respect to the measure μ, that is, $f \in L^{-1}(\mu(dx))$. Assume that the objective is to estimate the integral of f with respect to the measure μ. In numerical quadrature methods, one defines a set of support points $x_j \in \mathbb{R}^n, 1 \leq j \leq N$ and the corresponding weights w_j to get an approximation

$$\int_{\mathbb{R}^n} f(x)\,\mu(dx) \approx \sum_{j=1}^{N} w_j f(x_j)$$

Typically, the quadrature methods are designed so that they are accurate for given functions spanning a finite-dimensional space. Typically, these functions are polynomials of restricted degree.

In Monte Carlo integration, the support points x_j are generated randomly by drawing from some probability density and the weights w_j are then determined from the distribution μ. Ideally, the support points are drawn from the probability distribution determined by the measure μ itself. Indeed, let $X \in \mathbb{R}^n$ denote a random variable such that μ is its probability distribution. If we had a random generator such that repeated realizations of X could be produced, we could generate a set of points distributed according to μ. Assume that $\{x_1, x_2, \ldots x_N\} \subset \mathbb{R}^n$ is such a representative ensemble of samples distributed according to the distribution μ. We could then seek to approximate the integral of f by the so-called ergodic average,

$$\int_{\mathbb{R}^n} f(x)\,\mu(dx) = E\{f(X)\} \approx \frac{1}{N}\sum_{j=1}^{N} f(x_j)$$

The MCMC methods are systematic ways of generating a sample ensemble such that this holds. We need some basic tools from probability theory to do this.

Let $\mathcal{B} = \mathcal{B}(\mathbb{R}^n)$ denote the Borel sets over $\mathbb{R}^n$. A mapping $P : \mathbb{R}^n \times \mathcal{B} \to [0,1]$ is called a probability transition kernel, if for each $B \in \mathcal{B}$, the mapping $\mathbb{R}^n \to [0,1], x \to P(x,B)$ is a measurable function and for each $x \in \mathbb{R}^n$, the mapping $\mathcal{B} \to [0,1], B \to P(x,B)$ is a probability distribution.

A *discrete time stochastic process* is an ordered set $\{X_j\}_{j=1}^{\infty}$ of random variables $X_j \in \mathbb{R}^n$. A *time-homogeneous Markov chain* with the transition kernel P is a stochastic process $\{X_j\}_{j=1}^{\infty}$ with the properties

$$\mu_{X_{j+1}}(B_{j+1}|x_1,\ldots,x_j) = \mu_{X_{j+1}}(B_{j+1}|x_j) = P(x_j, B_{j+1})$$

In words, the first equality states that the probability for $X_{j+1} \in B_{j+1}$ conditioned on observations $X_1 = x_1, \ldots, X_j = x_j$ equals the probability conditioned on $X_j = x_j$ alone. This property is stated often by saying that "the future depends on the past only through the present." The second equality says that time is homogeneous in the sense that the dependence of adjacent moments does not vary in time. Let us emphasize that the kernel P does not depend on time j.

More generally, we define the transition kernel that propagates k steps forward in time, setting

$$\begin{aligned} P^{(k)}\left(x_j, B_{j+k}\right) &= \mu_{X_{j+k}}\left(B_{j+k} | x_j\right) \\ &= \int_{\mathbb{R}^n} P\left(x_{j+k-1}, B_{j+k}\right) P^{(k-1)}\left(x_j, dx_{j+k-1}\right) \end{aligned}$$

where it is understood that $P^{(1)}\left(x_j, B_{j+1}\right) = P\left(x_j, B_{j+1}\right)$. In particular, if μ_{X_j} denotes the probability distribution of X_j, the distribution of X_{j+1} is given by

$$\mu_{X_{j+1}}\left(B_{j+1}\right) = \mu_{X_j} P\left(B_{j+1}\right) = \int_{\mathbb{R}^n} P\left(x_j, B_{j+1}\right) u_{X_j}\left(dx_j\right)$$

The measure μ is an invariant measure of $P\left(x_j, B_{j+1}\right)$ if

$$\mu P = mu$$

that is, the distribution of the random variable X_j before the time step $j \to j+1$ is the same as the variable X_{j+1} after the step.

We still need to introduce few concepts concerning the transition kernels. Given a probability measure μ, the transition kernel P is irreducible (with respect to μ) if for each $x \in \mathbb{R}^n$ and $B \in \mathcal{B}$ with $\mu\left(B\right) > 0$ there exists an integer k such that $P^{(k)}\left(x, B\right) > 0$. Thus, regardless of the starting point, the Markov chain generated by the transition kernel P visits with a positive probability any set of positive measure.

Let P be an irreducible kernel. We say that P is periodic if, for some integer $m \geq 2$, there is a set of disjoint nonempty sets $\{E_1, \ldots, E_m\} \subset \mathbb{R}^n$ such that for all $j = 1, \ldots, m$ and all $x \in E_j, P\left(x, E_{j+1}\left(\mathrm{mod} m\right)\right) = 1$. In other words, a periodic transition kernel generates a Markov chain that remains in a periodic loop forever. A kernel P is an aperiodic kernel if it is not periodic.

The following result is of crucial importance for MCMC methods. The proof of this theorem will be omitted.

Proposition: Let μ be a probability measure in $\mathbb{R}^n$ and $\{X_j\}$ a time-homogeneous Markov chain with a transition kernel P. Assume further that μ is an invariant measure of the transition kernel P, and that P is irreducible and aperiodic. Then for all $x \in \mathbb{R}^n$

$$\lim_{N \longrightarrow \infty} P^{(N)}(x, B) = \mu(B) \text{ for all } B \in \mathcal{B}$$

and for $f \in L^{-1}(\mu(dx))$, almost certainly

$$\lim_{N \longrightarrow \infty} \frac{1}{N} \sum_{j=1}^{N} f(X_j) = \int_{\mathbb{R}^n} f(x)\,\mu(dx)$$

The above theorem gives a clear indication how to explore a given probability distribution. One needs to construct an invariant, aperiodic and irreducible transition kernel P and draw a sequence of sample points $x_1, x_2, \ldots$ using this kernel, that is, one needs to calculate a realization of the Markov chain.

The property is the important ergodicity property used in Monte Carlo integration. The convergence stating that μ is a limit distribution for the transition kernel P, can be stated also in a slightly stronger form.

In the following sections, we discuss how to construct transition kernels with the desired properties. The two most common procedures are the Metropolis–Hastings algorithm and the Gibbs sampler. A large number of variants of these basic methods have been proposed.