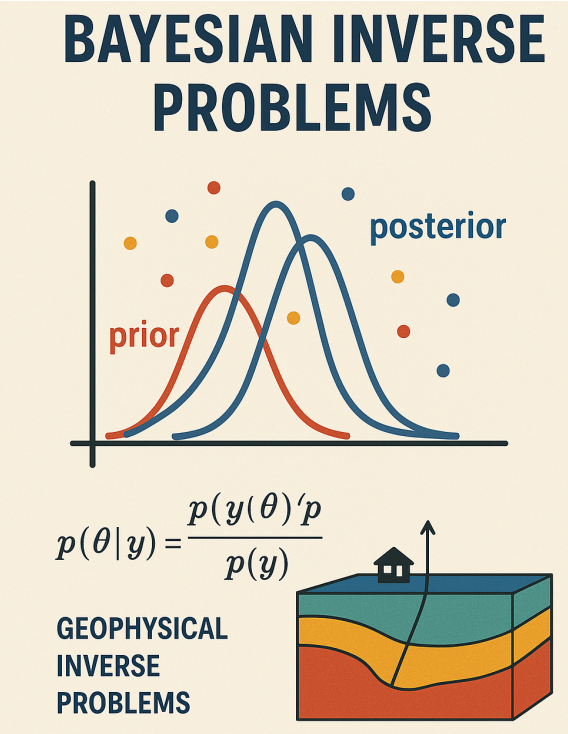
Posterior Estimation: MC, McMC, HMC, V.I.

Mark Asch - MAKUTU/2025



Outline of the course

1. Introduction to inverse problems and data assimilation: overview, setting, history, definitions, examples.

2. Bayesian inverse problems.

   (a) Bayesian inference.
   (b) Bayesian/Statistical inversion theory.
   (c) Full wave inversion example.
   (d) Point and interval estimates.

3. Posterior Estimation methods.

   (a) Monte Carlo methods.
   (b) Rejection Sampling. Importance Sampling.
   (c) McMC and variants for posterior estimation.
   (d) Metropolis Hastings, Gibbs and Hamiltonian McMC.
   (e) Introduction to Variational Inference (VI) for posterior estimation.

4. Statistical estimation, Kalman filters and sequential data assimilation.

   (a) Introduction to statistical DA.
   (b) The Kalman filter and Ensemble KF.
   (c) Ensemble Kalman Inversion (EKI).

Reference Textbooks

1. M. DeGroot, M. Schervisch. *Probability and Statistics*. Addison-Wesley, 2012.

2. P. Billingsley. *Probability and Measure*. John Wiley & Sons, 1995.

3. A. Gelman, et al. *Bayesian Data Analysis*. CRC Press. 2014.

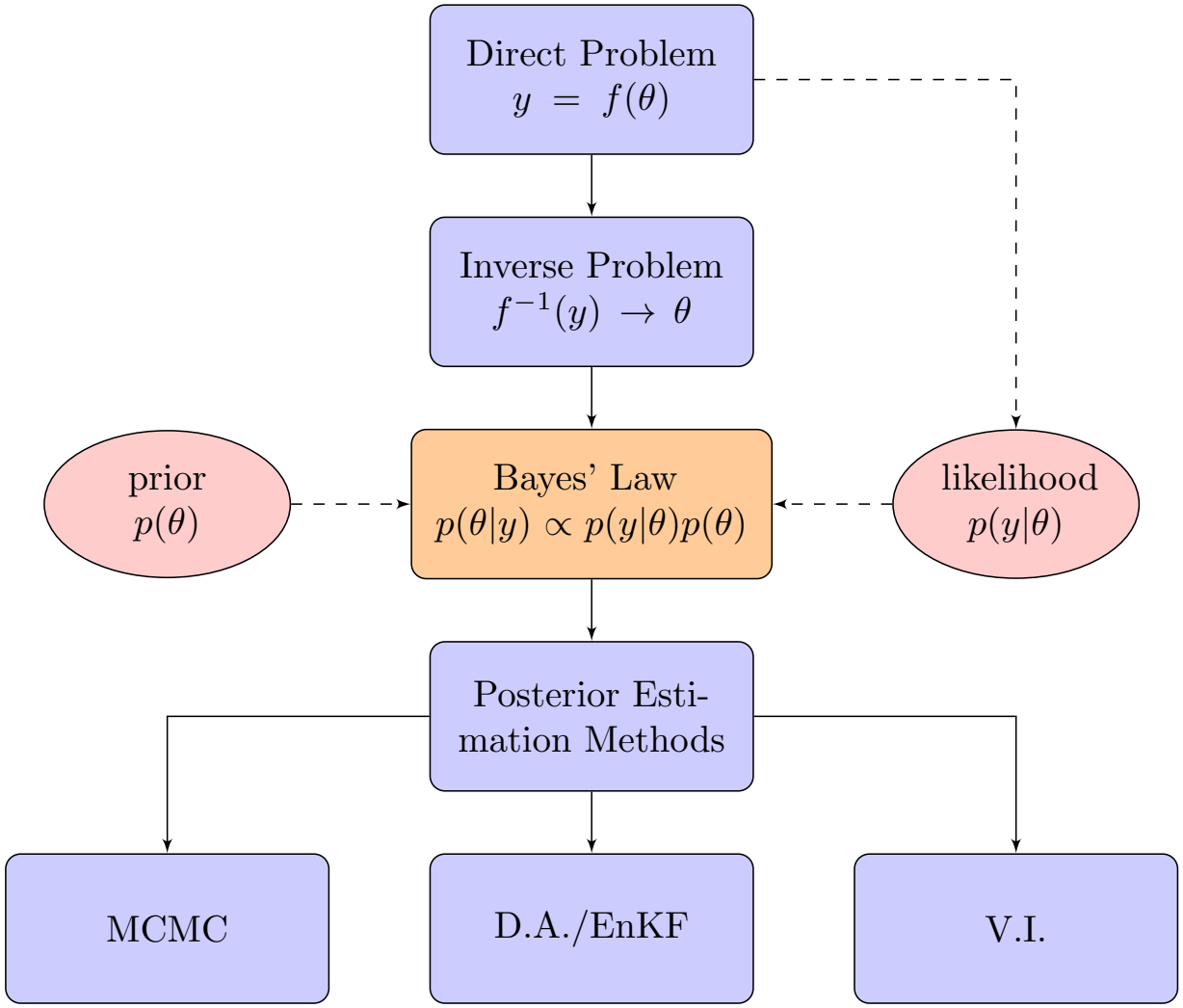
4. C. Robert, G. Casella. *Méthodes de Monte-Carlo avec R*. Springer, 2011. (256p.)

5. C. Robert, G. Casella. *Monte Carlo Statistical Methods*. Springer, 2004. (645p.)

6. M. Asch. *A Toolbox for Digital Twins*. SIAM, 2022.

INTRODUCTION

Recall: OVERVIEW



Recall: Bayes' Theorem

- We now formulate the general version of **Bayes' Theorem**.

Theorem 1. *Suppose that n random variables, $X_1, \ldots, X_n$, form a random sample from a distribution with density, or probability function in the case of a discrete distribution, $f(x \mid \theta)$. Suppose also that the unknown parameter, θ, has a prior pdf $f(\theta)$. Then the posterior pdf of θ is*

$$f(\theta \mid x) = \frac{f(x_1 \mid \theta) \cdots f(x_n \mid \theta) f(\theta)}{f_n(x)}, \quad (1)$$

where $f_n(x)$ is the marginal joint pdf of $X_1, \ldots, X_n$.

- In this theorem,

  $\Rightarrow$ the *prior*, $f(\theta)$, represents the credibility of, or belief in the values of the parameters

we seek, without any consideration of the data/observations;

$\Rightarrow$ the *posterior*, $f(\theta \mid x)$, is the credibility of the parameters with the data taken into account;

$\Rightarrow$ $f(x \mid \theta)$, considered as a function of θ, is the *likelihood* function, which is the probability that the data/observation could be generated by the model with a given value of the parameter;

$\Rightarrow$ the denominator, called the *evidence*, $f_n(x)$, is the total probability of the data taken over all the possible parameter values, also called the *marginal likelihood*, or the marginal, and can be considered as a normalization factor;

$\Rightarrow$ the posterior distribution is thus proportional to the product of the likelihood and the prior distribution, or, in applied terms,

$$f(\text{parameter} \mid \text{data}) \propto f(\text{data} \mid \text{parameter})\, f(\text{parameter}).$$

- What can one do with the posterior distribution thus obtained? The answer is a lot of things, in fact a complete quantification of the incertitude of the parameter's estimation is possible. We can compute:

$\Rightarrow$ Point estimates by summarizing the center of the posterior. Typically, these are the posterior mean or the posterior mode.

$\Rightarrow$ Interval estimates for a given level α—see below.

$\Rightarrow$ Estimates of the probability of an event, such as $\mathrm{P}(a < \theta < b)$ or $P(\theta > b)$.

$\Rightarrow$ Posterior quantiles.

Recall: Bayesian Framework

Advantages of Bayes

1. The solution to the Bayesian inverse problem, "*find* u *given* $y = G(u) + \eta$," is the posterior pdf $\pi^y(u)$. This allows for complete uncertainty quantification in the inferred parameter.

2. In the Bayesian framework, the inverse problem is well-posed:

   (a) There exists a unique posterior distribution for all $y \in \mathbb{R}^{d_y}$

   (b) If there are multiple minimisers of $\|y - G(u)\|_2^2$, then the posterior distribution has multiple modes.

   (c) The posterior distribution π^y depends continuously on y : if $L(y|u)$ is locally Lipschitz in y, then

$$d_{\mathrm{TV}}(\pi^y, \pi^{y'}) \leq C \|y - y'\|_2 .$$

Recall: Bayesian Framework

Challenges of Bayes

- The posterior distribution is typically not known in closed form—notable exception is the linear Gaussian case, which will be solved below and in exercises.

- Advanced sampling methods such as Markov chain Monte Carlo (McMC) methods are required for sampling from the posterior, e.g. for computing the posterior mean.

- Many variants of McMC exist, including so-called variational inference methods, where we seek an optimal posterior apprximation that minimizes a KL distance between two measures/pdf's.

- Finally, ensemble Kalman filters can be used for approximating the posterior.

Estimators

Recall: Bayesian Inverse Problem

Recall: Formulation

- Direct problem, with noise:

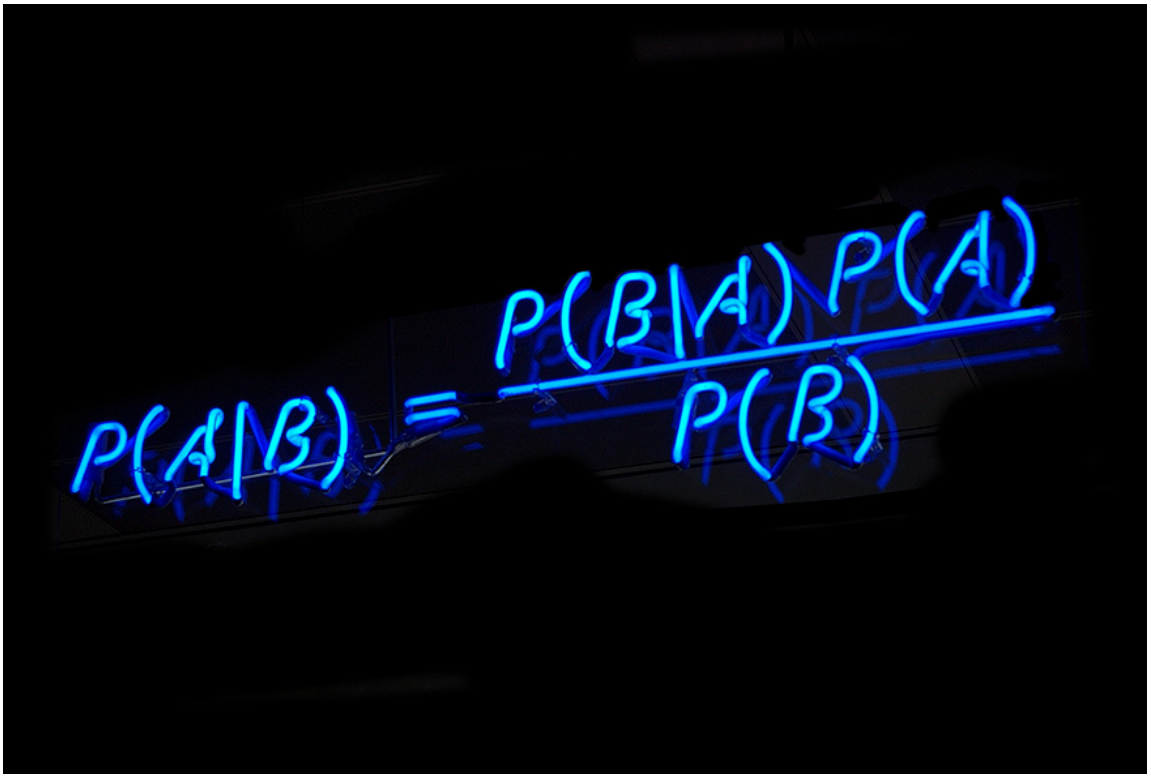
$$y = \mathcal{G}(u) + \eta, \quad \eta \sim \mathcal{N}(0, \gamma^2 I)$$

- Statistical (Bayesian) Inverse problem: find $\pi^y(u)$ on $\mathbb{R}^{d_u}$ such that

$$\underbrace{\pi^y(u)}_{p(u|y)} = \underbrace{\frac{1}{Z}}_{1/p(y)} \underbrace{\exp\left(-\Phi(u;y)\right)}_{p(y|u)} \underbrace{\pi_0(u)}_{p(u)},$$

where

$$\begin{aligned} Z &= \int_{\mathbb{R}^{d_u}} \exp\left(-\Phi(u;y)\right) \pi_0(u)\, \mathrm{d}u \\ &= \mathrm{E}_{\pi_0}\left[\exp\left(-\Phi(\cdot;y)\right)\right] \\ \Phi(u;y) &= \frac{1}{2\gamma^2} \left\|y - \mathcal{G}(u)\right\|_2^2 \end{aligned}$$


P(A|B) = P(B|A) P(A) / P(B)

Recall: Bayesian Inverse Problem

Estimators

- The dimension of the inverse problem can be very large and consequently the posterior distribution lives in a high dimensional space—this makes its computation and visualisation very difficult.

- However, we can calculate different point estimators and spread or region estimators.

  $\Rightarrow$ The point estimators approximate the most probable value of the unknown given the data and the prior.

  $\Rightarrow$ The spread estimators give a region that contain the unknown with some high probability.

- Estimation of the complete posterior is treated in detail below (McMC, VI, EnKF).

- **Remark** [Geyer]: "Bayesians have little interest in point estimates of parameters. To them a parameter is a *random variable*, and what is important

is its distribution. A point estimate is a meager bit of information as compared, for example, to a plot of the posterior density. Frequentists too have little interest in point estimates except as tools for constructing tests and confidence intervals."

Recall: Bayesian Inverse Problem

Point Estimators

- One of the most used statistical estimators is the maximum a posterior estimate (MAP)

  $\Rightarrow$ The MAP is the mode of the posterior distribution. That is, given the posterior density $\pi^y(u)$ the MAP estimate u_{MAP} satisfies

  $$u_{\mathrm{MAP}} = \arg\max_{u \in \mathbb{R}^n} \pi^y(u)$$

  if it exists. The MAP may not be unique—eg. for bi-modal distributions.

- Another widely used point estimate is the conditional mean (CM) of the unknown u given the data y, which is defined by

  $$u_{\mathrm{CM}} = \mathrm{E}(u|y) = \int_{\mathbb{R}^n} u\, \pi^y(u)\, \mathrm{d}u,$$

  proivided that the integral converges.

- **Remark** [Geyer]: The posterior mean and median are often "woofed" about using decision-theoretic terminology.

  $\Rightarrow$ The posterior mean is the *Bayes estimator that minimizes squared error loss*.

  $\Rightarrow$ The posterior median is the *Bayes estimator that minimizes absolute error loss*.

  $\Rightarrow$ The posterior mode is the *Bayes estimator that minimizes the loss* $t \mapsto \mathrm{E}\left[1 - I_{(-\epsilon,\epsilon)}(t-\theta)|y\right]$ when ϵ is infinitesimal.

Recall: Bayesian Inverse Problem

Spread Estimators

- The most common spread estimator for the BIP is the Bayesian Credible Interval or Set (BCI)

  $\Rightarrow$ A level $(1-\alpha)$ credible set, C_α, for given small $\alpha \in (0,1)$ is defined by

$$\Pi(C_\alpha|y) = \int_{C_\alpha} \pi^y(u)\,\mathrm{d}u = 1-\alpha.$$

  $\Rightarrow$ Thus, a credible set C_α is a region that contains a large proportion of the posterior mass.

  $\Rightarrow$ Alternative name is highest posterior density region (HPD region).

- The random sets C_α that frequently contain the 'true' unknown $u^\dagger$ such that

$$\mathrm{P}(u^\dagger \in C_\alpha) = 1-\alpha,$$

  are known as confidence regions of level $(1-\alpha)$.

MONTE CARLO

Monte Carlo Method

- The "Monte Carlo method" refers to the theory and practice of learning about probability distributions by simulation rather than calculus.

- It is a "cute" name for

  $\Rightarrow$ computer simulation of probability distributions, and

  $\Rightarrow$ calculating probabilities and expectations by averaging over the simulations

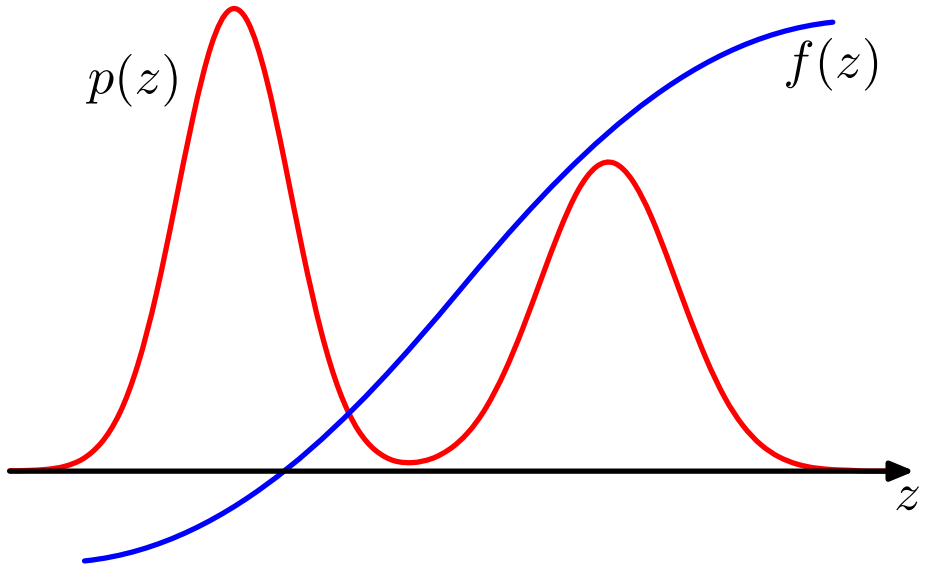
- It refers to a lot more than just simulation studies in statistics:

  $\Rightarrow$ Any integral or sum that cannot be done analytically, either by hand or by a computer algebra system can be put in the form of the expectation of some random variable with respect to some probability distribution.

$\Rightarrow$ So it is a general method of doing integrals or sums.

$\Rightarrow$ It is also the basis of the Ensemble Kalman Filter (see next lecture).

Ordinary Monte Carlo (OMC) Method



- In ordinary Monte Carlo (OMC) we use IID simulations from the distribution of interest.

  $\Rightarrow$ Suppose $X_1, X_2, \ldots$ are IID simulations from some distribution with density $p(x)$, and suppose we want to compute an expectation

$$\mu = \mathrm{E}\left[f\left(X\right)\right] = \int_{-\infty}^{\infty} f(x)p(x)\,\mathrm{d}x.$$

⇒ Then according to the law of large numbers (LLN), the estimate

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} f\left(X_i\right)$$

converges in probability to μ.

⇒ Now, the central limit theorem (CLT) says

$$\sqrt{n}\left(\hat{\mu}_n - \mu\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \sigma^2\right)$$

where

$$\sigma^2 = \mathrm{Var}\left[f\left(X_i\right)\right],$$

which can be estimated by the empirical variance

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(f\left(X_i\right) - \hat{\mu}_n\right)^2.$$

⇒ An asymptotic 95% confidence interval for μ is then

$$\hat{\mu}_n \pm 1.96\frac{\hat{\sigma}_n}{\sqrt{n}}.$$

- **Remark**: OMC works very well, but

  $\Rightarrow$ its convergence is very slow $\sim \mathcal{O}(n^{-1/2})$—to reduce the error by a factor of 10 (i.e. gain one decimal of accuracy), we need to increase the number of samples by a factor of 100.

  $\Rightarrow$ it is very expensive for vector functions,

  $\Rightarrow$ rejection and importance sampling fail in high dimensions (the curse of dimensionality)—see below.

  $\Rightarrow$ Hence Markov chain Monte Carlo (McMC)—see further below.

Monte Carlo Integration

Principles and Basics

Theorem 2 (SLLN for Monte Carlo Integration). *Let $\{X_i\}_{i\geq 1}$ be a sequence of independent identically distributed (i.i.d.) random variables and let $f\colon \mathbb{R} \to \mathbb{R}$ be an integrable function such that $\mathrm{E}[f(X_i)] < \infty$. Then $\{f(X_i)\}_{i\geq 1}$ is also a family of i.i.d. random variables, and*

$$\frac{1}{n}\sum_{i=1}^{n} f(X_i) \to \mathrm{E}[f(X_i)]$$

when $n \to \infty$. In particular, if the X_i are uniformly distributed on $[a,b]$ and $f\colon [a,b] \to \mathbb{R}$ is a continuous function, then

$$\frac{1}{n}\sum_{i=1}^{n} f(X_i) \to \frac{1}{b-a}\int_a^b f(x)\,\mathrm{d}x$$

almost surely.

- **Remarks**:

  $\Rightarrow$ Since $f(X_i) \nrightarrow f(x)$ on a set of measure zero, i.e. countably many points $x \in [a, b,]$, this implies that

$$\lim_{n\to\infty} \frac{b-a}{n} \sum_{i=1}^{n} f(X_i) = \int_a^b f(x)\,\mathrm{d}x.$$

  $\Rightarrow$ The result generalizes easily.

- The SLLN has the following, very general form:

  $\Rightarrow$ first (re)define expectation...

Definition 1 (Expected Value). The expected (or mean) value of a random variable X on $(\Omega, \mathcal{F}, \mathrm{P})$ is the integral of X with respect to the measure P,

$$\mathrm{E}[X] = \int X\,\mathrm{dP} = \int_\Omega X(\omega)\,\mathrm{P}(\mathrm{d}\omega).$$

For nonnegative X, the expected value is always defined (it may be infinite). For general X, either the

expected value of the positive or the negative part of X must be finite.

Theorem 3 (SLLN). *If $X_1, X_2, \ldots, X_n$ are IID and have finite mean, then the sample (empirical) mean*

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to \mathrm{E}\left[X_1\right]$$

with probability 1.

Monte Carlo Integration

Intuition...

- We want to compute an integral

$$I = \int_{\Omega} f(x)\,\mathrm{d}x.$$

- By definition, the expectation of a function $g(x)$ with respect to some distribution $p(x)$ is defined by

$$\mathrm{E}\left[g(x)\right] = \int g(x)p(x)\,\mathrm{d}x.$$

- If we choose $g(x) = f(x)/p(x)$, then

$$\begin{aligned}\mathrm{E}\left[g(x)\right] &= \int \frac{f(x)}{p(x)}p(x)\,\mathrm{d}x \\ &= \int f(x)\,\mathrm{d}x \\ &= I\end{aligned}$$

- By the LLN, the sample average converges to the expectation, so we have

$$I \approx \bar{g}_n = \frac{1}{n} \sum_{i=1}^{n} g(x_i),$$

  where $\{x_i\}_{i=1}^n$ are IID points drawn from the law of $p(x)$, that is $x_i \sim p$.

- If the integral of $f(x)$ is a proper integral (i.e. bounded), and $p(x)$ is the uniform distribution on $[0, 1]$, then $g(x) = f(x)$ and this is known as ordinary Monte Carlo (OMC).

- If the integral of $f(x)$ is improper, then we need to use another distribution with the same support as $f(x)$.

Monte Carlo Integration

Proof for 1D case

- Consider the 1D integral

$$I = \int_a^b f(x)\,\mathrm{d}x.$$

- Suppose we have a sample (IID) of n uniform random variables in the interval, $X_i \in [a, b]$

- The ordinary Monte Carlo estimator says that the expected value of

$$I_n = \frac{b-a}{n}\sum_{i=1}^{n} f(x_i),$$

  is exactly equal to the integral we seek.

- Explicit proof:

$$\begin{aligned}
\mathrm{E}\left[I_n\right] &= \mathrm{E}\left[\frac{b-a}{n}\sum_{i=1}^{n} f(x_i)\right] \\
&= \frac{b-a}{n}\sum_{i=1}^{n} \mathrm{E}\left[f(x_i)\right] \\
&= \frac{b-a}{n}\sum_{i=1}^{n} \int_a^b f(x)p(x)\,\mathrm{d}x \\
&= \frac{1}{n}\sum_{i=1}^{n} \int_a^b f(x)\,\mathrm{d}x \\
&= \int_a^b f(x)\,\mathrm{d}x
\end{aligned}$$

Monte Carlo Integration

General case

- The restriction to uniform random variables can be relaxed with a small generalization. This is an extremely important step, since carefully choosing the PDF from which samples are drawn is an important technique for reducing variance in Monte Carlo integration—see below, and also its use in McMC.

- If the random variables X_i are drawn from some arbitrary PDF $p(x)$, then the estimator

$$I_n = \frac{1}{n} \sum_{i=1}^{n} \frac{f(x_i)}{p(x_i)}$$

  can be used to estimate the integral instead. The only limitation on $p(x)$ is that it must be nonzero for all x where $f(x)$ is non-zero.

- As above, it is not hard to see that the expected value of this estimator is the desired integral of f. Indeed,

$$\begin{aligned}
\mathrm{E}\left[I_n\right] &= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{f(x_i)}{p(x_i)}\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\int_a^b \frac{f(x)}{p(x)}p(x)\,\mathrm{d}x \\
&= \frac{1}{n}\sum_{i=1}^{n}\int_a^b f(x)\,\mathrm{d}x \\
&= \int_a^b f(x)\,\mathrm{d}x.
\end{aligned}$$

- **Remark**s:

  $\Rightarrow$ Extending this estimator to multiple dimensions or complex integration domains is straightforward. Nowhere was the dependence on dimension, or on the domain of integration relied upon.

⇒ The number of samples n can be chosen arbitrarily, regardless of the dimension of the integrand. This is another important advantage of Monte Carlo over traditional deterministic quadrature techniques. The number of samples taken in Monte Carlo is completely independent of the dimensionality of the integral, while with standard numerical quadrature techniques the number of samples required is exponential in the dimension.

⇒ Showing that the Monte Carlo estimator converges to the right answer is not enough to justify its use; a good knowledge of the rate of convergence is important too. Using the CLT, we can readily show that the error in the Monte Carlo estimator decreases at a rather slow rate of $\sim \mathcal{O}(n^{-1/2})$.

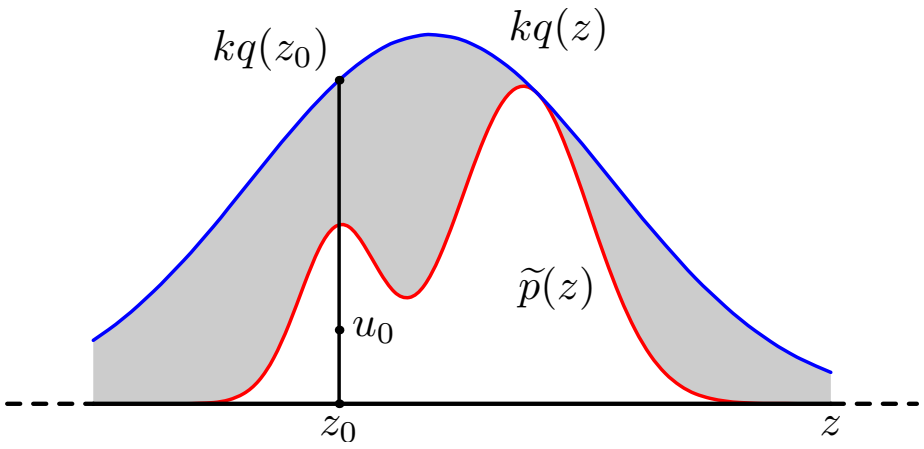
⇒ Although standard quadrature techniques converge faster than MC in one dimension, their performance becomes exponentially worse as the dimensionality of the integrand increases, while Monte Carlo's convergence rate is independent

of the dimension. This makes Monte Carlo the only practical numerical integration algorithm for high-dimensional integrals.

→ For example, Simpson's quadrature rule has 4th order convergence in 1D, but in d dimensions this becomes $\sim \mathcal{O}(n^{-4/d})$, so for $d > 8$ Monte Carlo integration has better convergence. Moreover, Simpson requires stringent smoothness conditions—4th order derivatives of f must be bounded. This is not the case in MC, where we only require measurability.

Monte Carlo Integration

Sampling Strategies: Rejection Sampling



- Problem statement:

  $\Rightarrow$ Sample from a distribution $p(\mathbf{z})$ that is unknown, and difficult to sample from (reminds us of the posterior in Bayes...).

  $\Rightarrow$ Suppose we can easily evaluate (eg. likelihood x prior) $p(\mathbf{z})$ for any given vlaue of $\mathbf{z}$ up to some (unknown) normalizing constant, Z_p, so that

$$p(z) = \frac{1}{Z_p}\tilde{p}(z)$$

where $\tilde{p}(z)$ is the quantity that can be (readily) evaluated, but Z_p is unknown.

In the rejection sampling method, samples are drawn from a simple distribution $q(z)$, known as the proposal distribution, and rejected if they fall in the grey area between the unnormalized distribution $\tilde{p}(z)$ (red) and the scaled distribution $kq(z)$ (blue), known as the comparison function. *Claim*: the resulting samples are distributed according to $p(z)$, which is the normalized version of $\tilde{p}(z)$. (see Exercise in Practical #3)

- **Intuition**: we just have to construct a probability density function $q(z)$, such that $kq(z) > p(z)$ for all z. In other words, we need a PDF that when multiplied by a constant k, is everywhere larger than $p(z)$.

Monte Carlo Integration

Rejection Sampling: Algorithm

- **Algorithm**:

  $\Rightarrow$ Define a simple distribution $q(z)$ and find a k such that for all z

  $$kq(z) \geq \tilde{p}(z).$$

  $\Rightarrow$ Draw $z_0 \sim q(z)$.
  $\Rightarrow$ Draw $u_0 \sim \mathcal{U}\left[0, kq(z_0)\right]$.
  $\Rightarrow$ Discard if $u_0 > \tilde{p}(z_0)$, otherwise retain u_0.

- **Remarks**:

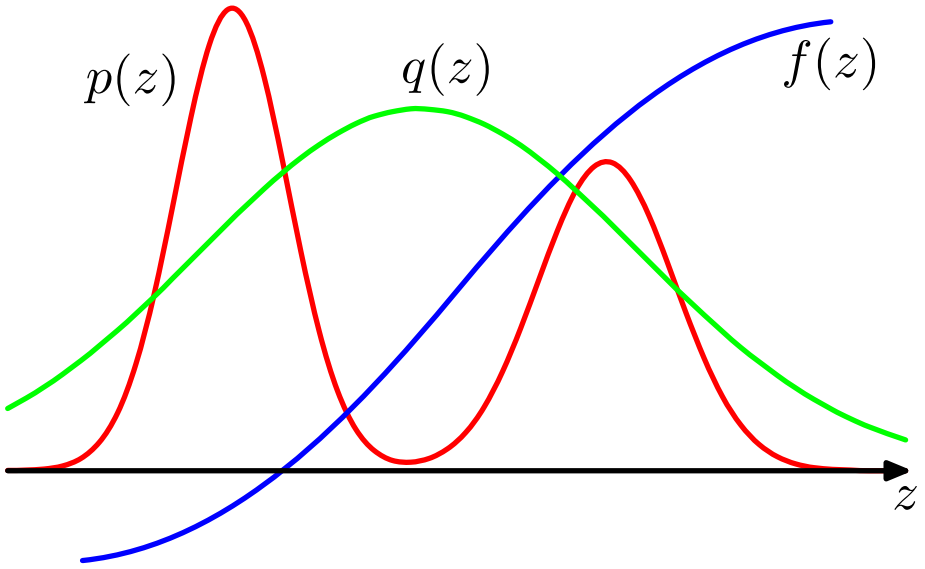
  $\Rightarrow$ The retained pairs (z_0, u_0) then have a uniform distribution under the curve of $\tilde{p}(z)$, and hence the corresponding z values are distributed according to $p(z)$, as desired. See Exercises for the proof.

⇒ The fraction of points that are rejected by this method depends on the ratio of the area under the unnormalized distribution $\tilde{p}(z)$ to the area under the curve $kq(z)$. We therefore see that the constant k should be as small as possible subject to the limitation that $kq(z)$ must be nowhere less than $\tilde{p}(z)$. See Python exercise for the extreme sensitivity of the method to the value of k.

⇒ Rejection sampling ⇒ Importance sampling ⇒ Metropolis-Hastings McMC. That's why it is important to understand rejection sampling!

Monte Carlo Integration

Sampling Strategies: Importance Sampling



Importance sampling addresses the problem of evaluating the expectation of a function $f(z)$ with respect to a distribution $p(z)$ from which it is difficult to draw samples directly. Instead, samples z are drawn from a simpler distribution $q(z)$, and the corresponding terms in the summation are weighted by ratios $p(z)/q(z)$, where we assume that we can readily evaluate $p(z)$ for any given value of z.

- As in the case of rejection sampling, importance sampling is based on the use of a proposal distribution $q(z)$ from which it is easy to draw samples.

- We can then express the expectation as a finite sum (thanks to the LLN) over samples $\{\mathbf{z}_i\}$ drawn from $q(\mathbf{z})$

$$\begin{aligned}\mathrm{E}\left[f(\mathbf{z})\right] &= \int f(\mathbf{z})p(\mathbf{z})\,\mathrm{d}\mathbf{z} \\ &= \int f(\mathbf{z})\frac{p(\mathbf{z})}{q(\mathbf{z})}q(\mathbf{z})\,\mathrm{d}\mathbf{z} \\ &\approx \frac{1}{n}\sum_{i=1}^{n}\frac{p(\mathbf{z}_i)}{q(\mathbf{z}_i)}f(\mathbf{z}_i)\end{aligned}$$

- The quantities $r_i = p(\mathbf{z}_i)/q(\mathbf{z}_i)$ are known as importance weights, and they correct the bias introduced by sampling from the wrong distribution. Note that, unlike rejection sampling, all of the samples generated are retained.

- As with rejection sampling, the success of the importance sampling approach depends crucially on how well the sampling distribution $q(\mathbf{z})$ matches the desired distribution $p(\mathbf{z})$.

  $\Rightarrow$ A good choice for q is one that is similar to p and has heavier (thicker) tails. For example, a t-distribution.

- If, as is often the case, $f(\mathbf{z})p(\mathbf{z})$ is strongly varying and has a significant proportion of its mass concentrated over relatively small regions of $\mathbf{z}$-space, then the set of importance weights $\{r_i\}$ may be dominated by a few weights having large values, with the remaining weights being relatively insignificant. Thus the effective sample size can be much smaller than the apparent sample size n.

- The principal advantage of IS is the drastic reduction of variance that it can provide—see Practical session.

IS Algorithm

Input:

- Target distribution $p(x)$,
- Function of interest $f(x)$,
- Proposal distribution $q(x)$,
- Number of samples N.

1. **Initialize**: Set sum $S = 0$
2. **For** $i = 1$ to N:
   (a) **Sample**: Generate $x_i \sim q(x)$
   (b) **Compute weight** : $r_i = \frac{p(x_i)}{q(x_i)}$
   (c) **Evaluate**: $y_i = f(x_i) \cdot r_i$
   (d) **Accumulate**: $S = S + y_i$
3. **Estimate**: $\hat{\mu}_{\text{IS}} = \frac{S}{N}$

MARKOV CHAIN MONTE CARLO

McMC

Markov Chains

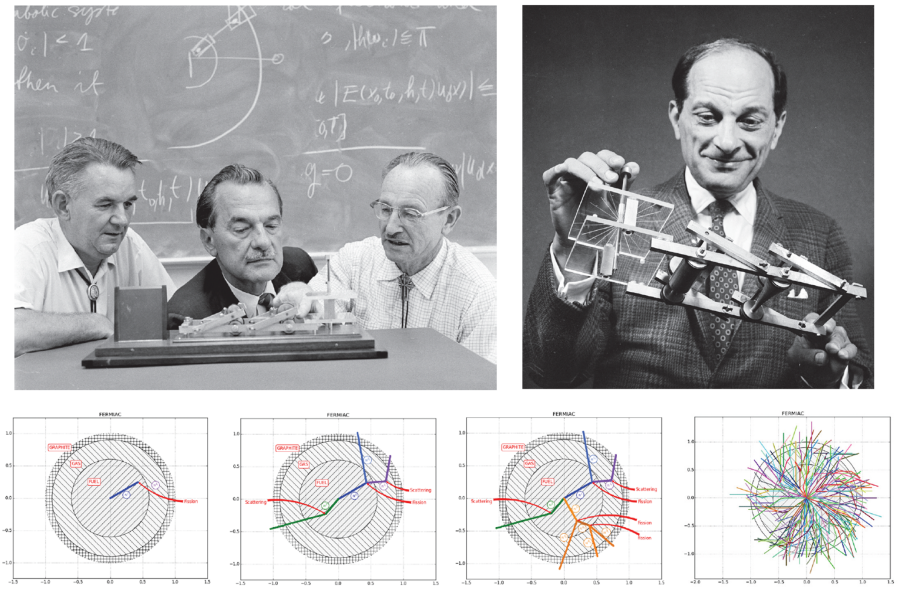
Principles and Basics

- **Motivation**:

  $\Rightarrow$ Rejection sampling and importance sampling strategies for evaluating expectations of functions suffer from severe limitations particularly in spaces of high dimensionality.

  $\Rightarrow$ Markov chain Monte Carlo (McMC) is a very general and powerful framework that allows sampling from a large class of distributions and scales well with the dimensionality of the sample space.

  $\Rightarrow$ Markov chain Monte Carlo methods have their origins in physics (Metropolis and Ulam, 1949), and it was only towards the end of the 1980s that they started to have a significant impact in the field of statistics.



Left: Bengt Carlson, Nick Metropolis, and Perc King, 1966; Right: Stanislaw Ulam with the FERMIAC, 1966. The FERMIAC (invented by E. Fermi) used the Monte Carlo method to model neutron transport in various types of nuclear systems.

- **Principles**:

  $\Rightarrow$ As with rejection and importance sampling, we again sample from a "friendly" proposal distribution, $q(x)$.

  $\Rightarrow$ McMC has the built-in capacity to automatically place its sampling points preferentially where the

pdf is large, in direct proportion to it.

$\Rightarrow$ In high-dimensional domains, or when the pdf is expensive to compute, both being prevalent in inverse parameter estimation and UQ problems, the McMC will then have orders of magnitude less computational work to perform.

$\Rightarrow$ McMC works by constructing and simulating a Markov Chain whose equilibrium distribution is the distribution of interest.

- **Theory**:

  $\Rightarrow$ based on existence and ergodic properties of limits of Markov chains;

  $\Rightarrow$ see references for details.

Markov Chains

V. brief Theory

- A Markov chain is a sequence of dependent random variables,

$$X^{(0)}, X^{(1)}, \ldots, X^{(t-1)}, X^{(t)}, X^{(t+1)}, \ldots$$

- The probability distribution of $X^{(t)}$ given the past variables depends only on $X^{(t-1)}$.

- This conditional probability distribution is called a transition matrix/kernel or a Markov kernel M,

$$X^{(t+1)} | X^{(0)}, X^{(1)}, \ldots, X^{(t)} \sim M\left(X^{(t)}, X^{(t+1)}\right).$$

- For the most part, all Markov chains encountered in Markov chain Monte Carlo (McMC) have a very

strong stability property: a stationary probability distribution exists by construction, such that

$$\lim_{t\to\infty} M^t(X,Y) = \pi(Y)$$

for all states X and Y.

- To reach this state, we need to attain a mixing time, defined as the time until the TV distance to π is less than ϵ,

$$\tau_{\text{mix}}(\epsilon) = \max_X \min\left\{t\colon \left\|M^t(X) - \pi\right\| \leq \epsilon\right\} \approx \ln \epsilon^{-1}.$$

- This ergodic property ensures that the Law of Large Numbers (Ergodic Theorem) can also be applied in McMC settings, giving

$$\frac{1}{T}\sum_{t=1}^{T} h\left(X^{(t)}\right) \xrightarrow{\text{P}} \text{E}_\pi\left[h(X)\right].$$

McMC

Algorithmic Principles

The principle on which all the McMC algorithms are based is the following.

- We want to sample from a complicated, unknown distribution $p(x)$ that can be written as

$$p(x) = \frac{1}{Z_p}\tilde{p}(x),$$

where $\tilde{p}(x)$ can be easily evaluated for all x, and Z_p is the unknown, intractable part.

- To perform the sampling, we will employ the same three steps as for importance/rejection sampling:

1. Choose a proposal distribution, $q(x \mid x_t)$.
2. Sample a candidate point y from the proposal.

3. Choose a new point, x_{t+1}, according to an acceptance probability ratio, r, that depends on q and p.

- **Remarks**:

  $\Rightarrow$ In the acceptance ratio the hard, intractable part, Z_p, cancels.

  $\Rightarrow$ In McMC, we choose the new point to ensure that we generate a Markov Chain that will converge to a stationary distribution π (see above) that is then a good approximation to *any* target distribution p.

  $\Rightarrow$ The good news is that this is valid universally, not only for all p, but also for any choice of the proposal q, though in practice we will tune q to improve convergence (see mixing time).

  $\Rightarrow$ The common choice for $q(y \mid x)$ is $\mathcal{N}(x, \delta^2)$, where δ is chosen to ensure good mixing—rule-of-thumb: choose δ so that proposals are accepted approximately 50% of the time.

McMC

Algorithm Principles

- General principles of McMC algorithms:

  $\Rightarrow$ the proposal distribution, q, combined with the acceptance ratio, r, together ensure that the resulting Markov chain explores regions of higher posterior probability more frequently than zones of lower posterior probability;

  $\Rightarrow$ moreover, this chain converges to a stationary, or equilibrium state that produces the posterior probability distribution that we seek.

McMC

Metropolis Hastings

There are several variants of the original Metropolis algorithm, but they all have the following basic structure.

- Fix a tractable distribution $\tilde{p}(x)$ and choose a proposal distribution $q(x)$ and an initial point x
- For each t

  $\Rightarrow$ sample a candidate point y from the proposal distribution q

  $\Rightarrow$ accept y with probability $\alpha = \min(1, r)$
- Next t
- Check convergence and output if satisfied.

McMC

Metropolis Hastings - details

- The most general, Metropolis-Hasting algorithm uses the acceptance ratio

$$r_{\mathrm{MH}} = \frac{p(y)}{p(x)} \frac{q(x \mid y)}{q(y \mid x)}, \tag{2}$$

- What does it mean, and how is it done, to "accept with probability ρ"?

  $\Rightarrow$ Draw a uniform random number $u \sim \mathcal{U}[0,1]$.

  $\Rightarrow$ Compare u to ρ :

  $\rightarrow$ if $u \leq \rho$ then ACCEPT new point y,

  $\rightarrow$ otherwise, $u > \rho$, REJECT i.e. keep x. $(y = x)$

  $\Rightarrow$ Explanation: $\mathrm{P}(u \leq \rho) = \rho$, by definition of a $\mathcal{U}[0,1]$ random variable, where $0 \leq \rho \leq 1$.

McMC

Gibbs Sampling

- This is a special case of the more general Metropolis-Hasting Algorithm particularly adapted to computing (or sampling from) high-dimensional multivariate distributions, since it proceeds by sequentially sampling from univariate conditional distributions, often available in explicit forms.

- The Gibbs algorithm for McMC sampling produces a Markov chain from a joint distribution of interest (the target) that must have a special form. Fixing all but one variable, the joint pdf must be of the same type and easy to simulate pseudo-random variables from it.

- The Gibbs algorithm then cycles through the coordinate variables, simulating each one in turn, conditional on all the others.

- As before, the algorithm requires a burn-in period

for the Markov chain to converge sufficiently closely to its stationary distribution. To evaluate the convergence and the standard errors of the simulated values, one must run several independent Markov chains simultaneously.

- Gibbs should not be used in general as it often exhibits very slow convergence. However, its simplicity compared to MH is often a determining factor. In any case, it should be used only if it seems to work well.

McMC

Hamiltonian Monte Carlo (HMC)

- Inherent inefficiency in Metropolis and Gibbs sampling is their random walk behavior. Simulations are long, due to "zigging and zagging" while exploring the target distribution.

- HMC[1] is a more efficient sampling algorithm that takes advantage of gradient information about the posterior in order to make sampling more efficient – exploring more of the space, while also accepting more proposed samples.

  $\Rightarrow$ It does so by associating each location in parameter space with a position in a physical simulation space, position-momentum, where the momentum is an auxiliary variable.

[1] S Duane, AD Kennedy, BJ Pendleton, D Roweth - Physics letters B, 1987. R. Neal, Mcmc using hamiltonian dynamics, in Handbook of Markov Chain Monte Carlo, edited by S. Brooks, A. Gelman, G. L. Jones, and X.-L. Meng, CRC Press, New York, 2011.

$\Rightarrow$ We then run a physics simulation based on Hamiltonian dynamics formulated in terms of position and momentum of a particle traveling on a surface.

$\Rightarrow$ At any point on the surface the particle will have a potential energy U and kinetic energy K that are traded back and forth due to conservation of energy.

$\Rightarrow$ HMC uses this energy conservation to simulate the dynamics, based on gradients of U and K.

- Hamiltonian Monte Carlo (HMC) uses physics (position and momentum) to suppress the local random walk behavior, yielding much more rapid movement through the target distribution.

  $\Rightarrow$ The posterior density $\pi(\theta|y)$ is augmented by an independent distribution $\pi(\phi)$ on the momentum, giving a joint distribution

$$\pi(\theta, \phi|y) = \pi(\phi)\pi(\theta|y),$$

  where the auxiliary variable ϕ ensures that the

algorithm moves faster through the parameter space.

$\Rightarrow$ We usually give ϕ a multivariate normal distribution, with mean zero and covariance M, a "mass" matrix, usually diagonal.

- HMC requires more effort to program and tune.

McMC

HMC—Formulation

1. Concatenate all the parameters into a single position variable, $\mathbf{q}$. We are trying to sample from the probability density function $\pi(\mathbf{q})$.

2. Add a momentum variable, $\mathbf{p}$, of the same dimension as $\mathbf{q}$, and consider the joint probability distribution

$$\pi(\mathbf{q}, \mathbf{p}) = \pi(\mathbf{p}|\mathbf{q})\pi(\mathbf{q}),$$

   where we get to choose $\pi(\mathbf{p}|\mathbf{q})$.

   (a) In practice, we will choose $\pi(\mathbf{p}|\mathbf{q}) = \mathcal{N}(\mathbf{0}, M)$, and often, we will choose a diagonal $M = mI$.

3. Define the Hamiltonian as[2]

$$H(\mathbf{q}, \mathbf{p}) = -\log \pi(\mathbf{q}, \mathbf{p})$$

[2] WLOG, a joint distribution can be expressed as $\pi(z) = (1/Z_\pi)\exp(-H(z))$

and expand

$$\begin{aligned} H(\mathbf{q}, \mathbf{p}) &= -\log \pi(\mathbf{p}|\mathbf{q})\pi(\mathbf{q}) \\ &= -\log \pi(\mathbf{p}|\mathbf{q}) - \log \pi(\mathbf{q}) \\ &= \quad K(\mathbf{p}, \mathbf{q}) \quad + \quad V(\mathbf{q}), \end{aligned}$$

where, using an analogy to physical systems,

(a) $K(\mathbf{p}, \mathbf{q})$ is the kinetic energy,
(b) $V(\mathbf{q})$ is the potential energy.

4. Evolve the system $(\mathbf{q}, \mathbf{p})$ according to Hamilton's equations,

$$\begin{aligned} \frac{\mathrm{d}\mathbf{q}}{\mathrm{d}t} &= \frac{\partial H}{\partial \mathbf{p}} \quad\quad = \frac{\partial K}{\partial \mathbf{p}} + \frac{\partial V}{\partial \mathbf{p}} \\ \frac{\mathrm{d}\mathbf{p}}{\mathrm{d}t} &= -\frac{\partial H}{\partial \mathbf{q}} = -\frac{\partial K}{\partial \mathbf{q}} - \frac{\partial V}{\partial \mathbf{q}} \end{aligned}$$

Note that $\frac{\partial V}{\partial \mathbf{p}} = \mathbf{0}$ and that the RHS is a (skewed) gradient of H.

- Special case: chose the kinetic energy to be a Gaussian, which lets us calculate the gradients by hand instead of recalculating them. Specifically, $\mathbf{q} \sim \mathcal{N}(0, M)$ and $\pi(\mathbf{p}|\mathbf{q}) = \mathcal{N}(\mathbf{p}|0, M)$, implying

$$K(\mathbf{p}, \mathbf{q}) = \frac{1}{2}\mathbf{p}^T M^{-1}\mathbf{p} + \log|M| + \text{const.},$$

and with our choice of $M = I$,

$$K(\mathbf{p}, \mathbf{q}) = \frac{1}{2}\mathbf{p}^T\mathbf{p} + \text{const.},$$

so

$$\frac{\partial K}{\partial \mathbf{p}} = \mathbf{p}$$

and

$$\frac{\partial K}{\partial \mathbf{q}} = \mathbf{0}$$

- We can then simplify Hamilton's equations to:

$$\frac{\mathrm{d}\mathbf{q}}{\mathrm{d}t} = \mathbf{p}$$
$$\frac{\mathrm{d}\mathbf{p}}{\mathrm{d}t} = -\frac{\partial V}{\partial \mathbf{q}}$$

- **Algorithm**:

  $\Rightarrow$ Sample a $\mathbf{p} \sim \mathcal{N}(0, I)$.
  $\Rightarrow$ Simulate $\mathbf{q}(t)$ and $\mathbf{p}(t)$ for some amount of time T using the simplified equations above.
  $\Rightarrow$ New sample is $\mathbf{q}^* = \mathbf{q}(T)$.

- After each application of a symplectic leapfrog algorithm for evolving the Hamiltonian dynamics, the resulting candidate state is accepted or rejected according to a Metropolis criterion based on the value of the Hamiltonian H

$$\min\left(1, \exp\left[H(\mathbf{p}, \mathbf{q}) - H(\mathbf{p}^*, \mathbf{q}^*)\right]\right).$$

  See Practical #3 for details.

McMC

Practical Implementation Aspects

- There are a LARGE number of practical issues involved in the implementations of McMC methods.

- These concern:

  $\Rightarrow$ distributions, in particular priors,
  $\Rightarrow$ noise models,
  $\Rightarrow$ discretization (for HMC),
  $\Rightarrow$ burn-in, convergence,
  $\Rightarrow$ predictive checking,
  $\Rightarrow$ model comparison.

- ALL details are given in the Practical #3.

EXAMPLES

HMC academic
1-D Example

The simplest example in 1D, where q and p are scalars, is based on the Hamiltonian

$$H(q,p) = K(p) + V(q)$$

with

$$V(q) = \frac{q^2}{2}, \quad K(p) = \frac{p^2}{2}.$$

This corresponds to a Gaussian distribution for q where $q \sim \mathcal{N}(0,1)$. The solutions are

$$q(t) = r\cos(a+t), \quad p(t) = -r\sin(a+t),$$

where a and r are constants. Hence a mapping T_s over s time-steps is a rotation by s radians around the origin in the (q,p) phase plane. We can readily see that the Markov chain produced by T_s satisfies the essential properties for McMC:

1. Reversibility.

2. Conservation of the Hamiltonian.

3. Volume preservation.

4. Symplectic.

Leapfrog method: (with time-step τ)

$$p_i\left(t+\frac{\tau}{2}\right) = p_i(t) - \frac{\tau}{2}\frac{\partial V}{\partial q_i}\left(q(t)\right)$$

$$q_i(t+\tau) = q_i(t) + \tau\frac{p_i\left(t+\frac{\tau}{2}\right)}{m_i}$$

$$p_i(t+\tau) = p_i\left(t+\frac{\tau}{2}\right) - \frac{\tau}{2}\frac{\partial V}{\partial q_i}\left(q(t+\tau)\right)$$

FWI using HMC

- A. Fichtner. HMCLab: a framework for solving diverse geophysical inverse problems using the Hamiltonian Monte Carlo method, *GJI* 235 (2023).

  $\Rightarrow$ 2-D inversion of a seismic data set based on the acoustic approximation, where the forward problem is represented by the constant-density acoustic wave equation,

  $$\frac{1}{c^2(x,z)}\frac{\partial^2 u(x,z,t)}{\partial t^2} = \frac{\partial^2 u(x,z,t)}{\partial x^2} + \frac{\partial^2 u(x,z,t)}{\partial z^2} + s(x,z,t),$$

  where u is the pressure field, c the acoustic velocity and s the source.

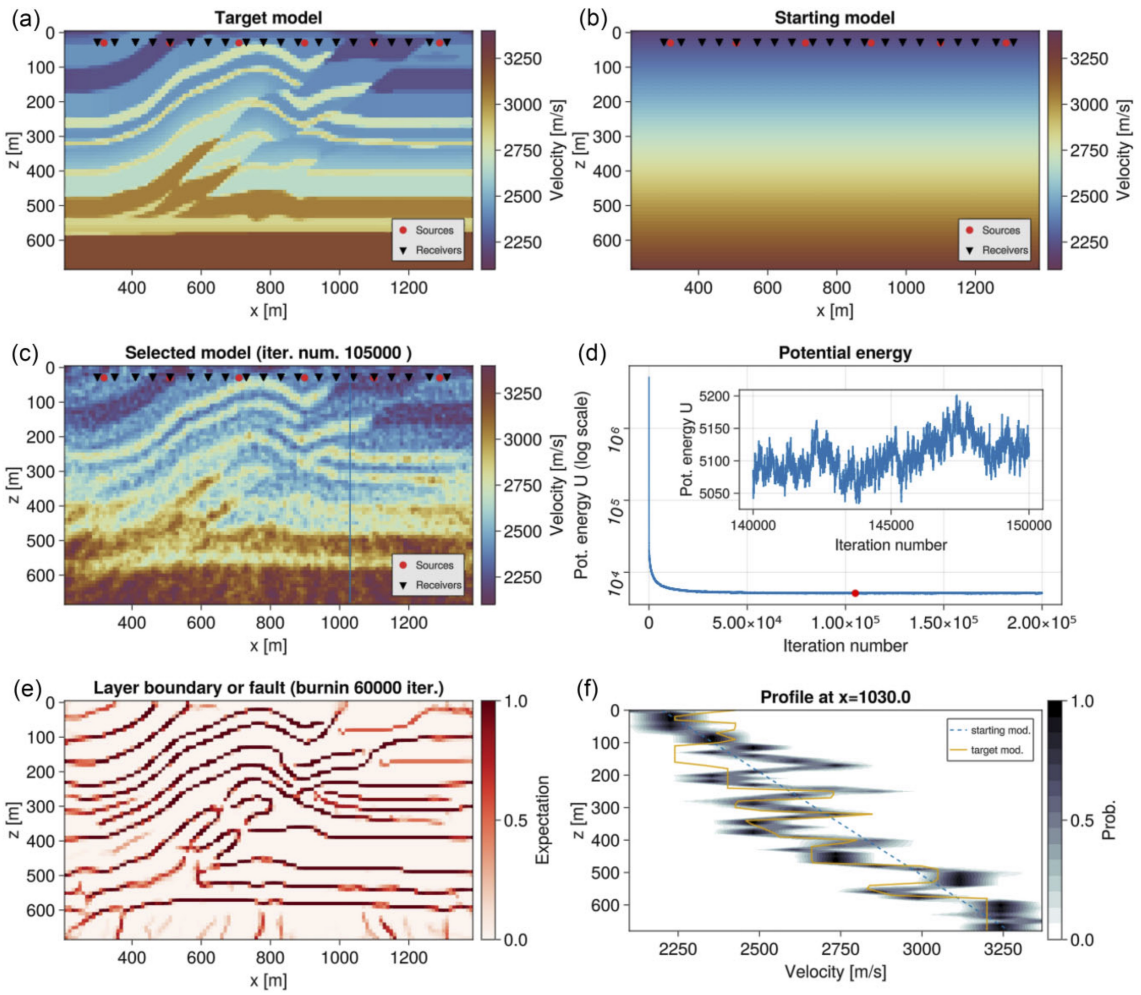
  $\Rightarrow$ HMC method based on the NUTS algorithm using an adjoint method for gradient computation.

  $\Rightarrow$ Other applications:

  $\rightarrow$ Travel-time hypocentre location using fibre-optic sensing.

  $\rightarrow$ First arrival tomography based on the eikonal equation.

  $\rightarrow$ Magnetic anomaly inversion with polygonal bodies.


(a)
Target model
(b)
Starting model
(c)
Selected model (iter. num. 105000)
(d)
Potential energy
(e)
Layer boundary or fault (burnin 60000 iter.)
(f)
Profile at x=1030.0
Sources
Receivers
z [m]
x [m]
Velocity [m/s]
Pot. energy U (log scale)
Iteration number
Expectation
Prob.
starting mod.
target mod.

VARIATIONAL INFERENCE

(V.I.)

Motivation and Overview

- McMC:

  $\Rightarrow$ Due to its random-walk behavior the method becomes inefficient in high dimensional space (e.g., >1000).

  $\Rightarrow$ Other more advanced McMC methods have been introduced in geophysics to solve high dimensional problems such as Hamiltonian Monte Carlo (HMC). Nevertheless, these methods remain intractable for large datasets and high dimensionality because of their extremely high computational cost.

- Variational inference solves Bayesian inference problems in a different way:

  $\Rightarrow$ Seek an optimal approximation to the posterior pdf within a predefined family of (simplified) probability distributions.

  $\Rightarrow$ This is achieved by minimizing a measure of the difference between the posterior pdf and

the approximating pdf, for example, the Kullback–Leibler (KL) divergence.

⇒ Since the method uses optimization rather than random sampling, it can be computationally more efficient than McMC and provide better scaling to high dimensionality.

⇒ The methods can also be applied to large datasets by dividing the dataset into minibatches and using stochastic optimization techniques that cannot be applied to McMC because they break the detailed balance which is required by most McMC methods.

- In variational inference the choice of the variational family determines the accuracy of the approximation and the complexity of the optimization

  ⇒ A good choice should be rich enough to approximate complex distributions and simple enough such that the optimization problem can be efficiently solved.

  ⇒ A common choice is to use a mean-field approximation in which the parameters are assumed to

be mutually independent.

⇒ To make variational inference applicable to general inverse problems, a variety of "black box" methods have been proposed based on different variational families, for example, the mean-field approximation, Gaussian distributions and probability transforms.

⇒ These methods are quite general and can be applied to a wide range of applications, for example, in geophysics to

→ travel time tomography

→ full waveform inversion

→ seismic image denoising

- Four different variational methods:

⇒ mean-field variational inference,

⇒ automatic differential variational inference (ADVI),

⇒ normalizing flows, and

⇒ Stein variational gradient descent (SVGD).

- We emphasize that McMC and VI (variational in-

ference) are different approaches to solving the same problem.

⇒ McMC algorithms sample a Markov chain, whereas
⇒ variational algorithms solve an optimization problem.
⇒ McMC algorithms approximate the posterior with samples from the chain, whereas
⇒ variational algorithms approximate the posterior with the result of the optimization.

- **Conclusion**:

⇒ Variational inference is suited to large data sets and scenarios where we want to quickly explore many models.
⇒ McMC is suited to smaller data sets and scenarios where we are ready to pay a heavier computational cost for more precise samples.
⇒ Ensemble Kalman Inversion (EKI) is (possibly) the best compromise between accuracy and cost.

Method

- Define a family of known probability distributions (eg. Gaussians, or sums of Gaussians)

$$Q = \{q(\mathbf{u})\}\,.$$

- Find the best approximation to the posterior pdf $p(\mathbf{u}|\mathbf{y})$ within Q by minimizing the KL divergence between $q(\mathbf{u})$ and $p(\mathbf{u}|\mathbf{y})$

$$q^*(\mathbf{u}) = \arg\min_{q \in Q} \mathrm{KL}\left[q(\mathbf{u}) \| p(\mathbf{u}|\mathbf{y})\right],$$

  where the KL divergence is a measure of the difference between two pdf's, defined as

$$\mathrm{KL}\left[q(\mathbf{u}) \| p(\mathbf{u}|\mathbf{y})\right] = \mathrm{E}_q\left[\log q(\mathbf{u})\right] - \mathrm{E}_q\left[\log p(\mathbf{u}|\mathbf{y})\right].$$

- We cannot compute the KL since it depends on the intractable evidence term $p(\mathbf{y})$. This can be seen

by expanding the second term according to Bayes' formula,

$$\mathrm{E}_q\left[\log p(\mathbf{u}|\mathbf{y})\right] = \mathrm{E}_q\left[\log p(\mathbf{u},\mathbf{y})\right] - \log p(\mathbf{y}).$$

So we define the evidence lower bound (ELBO)

$$\mathrm{ELBO}(q) = \mathrm{E}_q\left[\log p(\mathbf{u},\mathbf{y})\right] - \mathrm{E}_q\left[\log q(\mathbf{u})\right]$$

whose maximization is equivalent to minimizing the KL divergence.

$\Rightarrow$ ELBO provides a lower bound for the evidence.

$\Rightarrow$ Note also that the first term is the expected log-likelihood which is optimized by the well-known EM Algorithm.

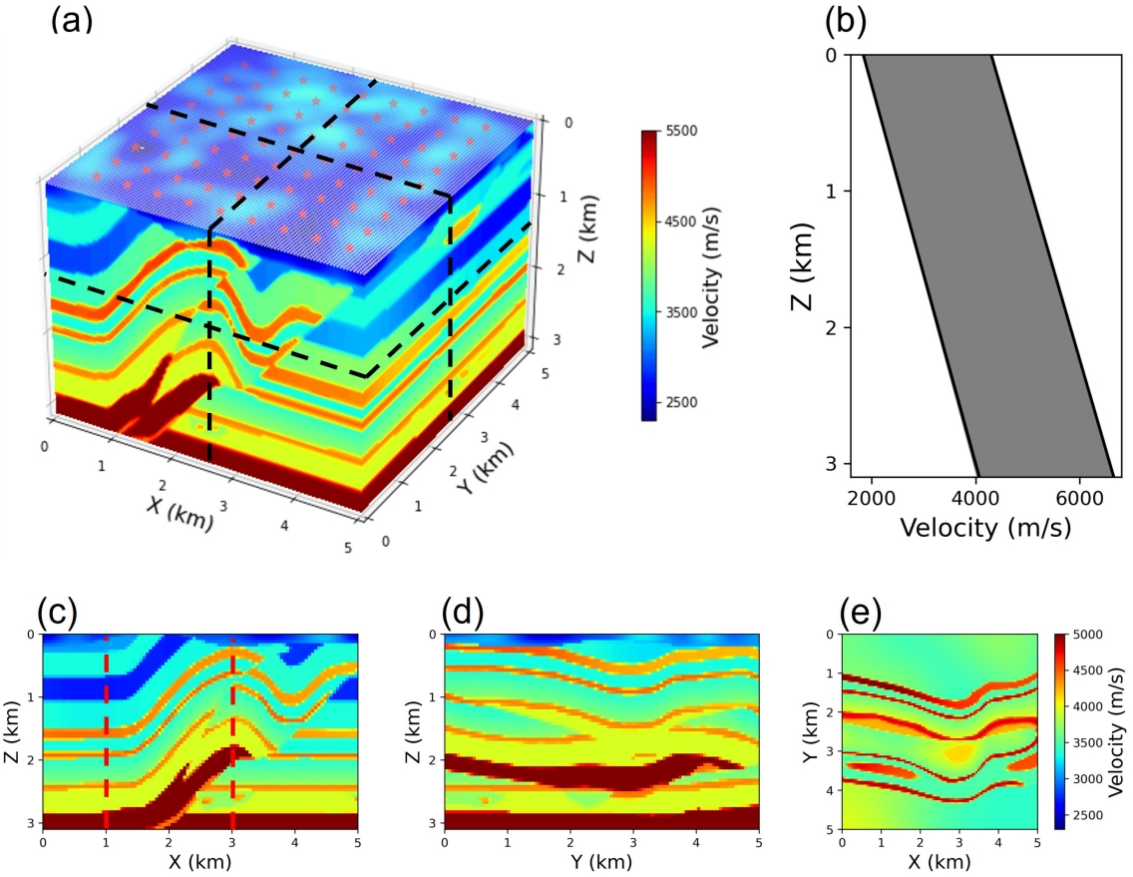
FWI using VI

- See Zhang, Curtis:

  $\Rightarrow$ An introduction to variational inference in geophysical inverse problems, *Advances in Geophysics*, 2021.

  $\Rightarrow$ VIP package.

  $\Rightarrow$ 3D FWI has computational cost only $\sim 10X$ that of deterministic FWI.


(a)
X (km)
Y (km)
Z (km)
Velocity (m/s)
5500
4500
3500
2500
(b)
Z (km)
Velocity (m/s)
2000
4000
6000
(c)
Z (km)
X (km)
(d)
Z (km)
Y (km)
(e)
Y (km)
X (km)
Velocity (m/s)
5000
4500
4000
3500
3000
2500

CONCLUSIONS

References

1. M. DeGroot, M. Schervisch. *Probability and Statistics*. Addison-Wesley. 2012.

2. P. Billingsley. *Probability and Measure*. John Wiley & Sons.1995.

3. A. Gelman, et al. *Bayesian Data Analysis*. CRC Press. 2014.

4. C. Robert, G. Casella. *Méthodes de Monte-Carlo avec R*. Springer, 2011.

5. M. Asch. *A Toolbox for Digital Twins*. SIAM. 2022.

6. A.M. Stuart. Inverse problems: a Bayesian perspective, *Acta Numerica*, 19 , pp. 451–559 , 2010.

7. M. Dashti, A.M. Stuart. The Bayesian approach to inverse problems, *Handbook of Uncertainty Quantification*, pp. 1–118, 2016

8. A. Tarantola. *Inverse problem theory and methods for model parameter estimation.* SIAM. 2005.

9. Zunino, Gebradd, Ghirotto, A. Fichtner. HMCLab: a framework for solving diverse geophysical inverse problems using the Hamiltonian Monte Carlo method, *Geophys J Int,* 235 (2023).

10. Xin Zhang, Muhammad Atif Nawaz, Xuebin Zhao, Andrew Curtis, Chapter Two - An introduction to variational inference in geophysical inverse problems, Editor(s): Cedric Schmelzbach, *Advances in Geophysics*, Volume 62, 2021, Pages 73-140, `https://doi.org/10.1016/bs.agph.2021.06.003`.