Inverse Problems

Sommersemester 2023

Vesa Kaarnioja
vesa.kaarnioja@fu-berlin.de

FU Berlin, FB Mathematik und Informatik

Eleventh lecture, June 26, 2023

Recap: the linear Gaussian setting

Let the unknown $x \in \mathbb{R}^d$ and the data $y \in \mathbb{R}^k$ follow the relation

$$y = Ax + \eta, \tag{1}$$

where

1. The forward model is linear, i.e., $A \in \mathbb{R}^{k \times d}$.
2. The prior distribution is Gaussian: $x \sim \pi = \mathcal{N}(x_0, \Gamma_{\mathrm{pr}})$, where Γ_{pr} is symmetric and positive definite.
3. The noise is Gaussian: $\eta \sim \nu = \mathcal{N}(\eta_0, \Gamma_{\mathrm{n}})$, where Γ_{n} is symmetric and positive definite.
4. x and η are independent.

Theorem

Under assumptions 1–4, the posterior distribution corresponding to (1) is Gaussian with $x|y \sim \mathcal{N}(\mu_{\mathrm{post}}, \Gamma_{\mathrm{post}})$, where we have

$$\begin{aligned}
\mu_{\mathrm{post}} &= (A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}A + \Gamma_{\mathrm{pr}}^{-1})^{-1}(A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}(y - \eta_0) + \Gamma_{\mathrm{pr}}^{-1}x_0),\\
\Gamma_{\mathrm{post}} &= (A^{\mathrm{T}}\Gamma_{\mathrm{n}}^{-1}A + \Gamma_{\mathrm{pr}}^{-1})^{-1}.
\end{aligned}$$

Small noise limit of the posterior distribution

Now we assume that the observational noise has the distribution $\eta \sim \mathcal{N}(0, \gamma^2 \Gamma_0)$ with $\gamma > 0$ and Γ_0 is a fixed symmetric and positive definite matrix, and consider the limiting behavior of the posterior mean and covariance as $\gamma \to 0$.

Substituting $\Gamma_{\mathrm{n}} = \gamma^2 \Gamma_0$ in the expressions for the posterior mean and covariance yield

$$m(\gamma) := \left(A^{\mathrm{T}} \Gamma_0^{-1} A + \gamma^2 \Gamma_{\mathrm{pr}}^{-1}\right)^{-1} \left(A^{\mathrm{T}} \Gamma_0^{-1} y + \gamma^2 \Gamma_{\mathrm{pr}}^{-1} x_0\right), \tag{2}$$

$$C(\gamma) := \gamma^2 \left(A^{\mathrm{T}} \Gamma_0^{-1} A + \gamma^2 \Gamma_{\mathrm{pr}}^{-1}\right)^{-1}. \tag{3}$$

We distinguish between overdetermined, determined, and underdetermined problems.

Overdetermined and determined case

Recall that $x \in \mathbb{R}^d$ and $y \in \mathbb{R}^k$.

Theorem (Overdetermined and determined case)

Suppose in the linear Gaussian setting that $\Gamma_{\mathrm{n}} = \gamma^2 \Gamma_0$ *with* $\gamma > 0$*, and that* $\mathrm{Ker}(A) = \{0\}$.

1. *If* $d < k$*, then the posterior distribution* π^y *satisfies*

$$\pi^y \rightharpoonup \delta_{m^\dagger} \quad \text{as } \gamma \to 0,$$

where $m^\dagger$ *is the solution to the least squares problem*

$$m^\dagger = \operatorname*{arg\,min}_{u \in \mathbb{R}^d} \left\| \Gamma_0^{-\frac{1}{2}} (Au - y) \right\|^2.$$

2. *If* $d = k$*, then we have*

$$\pi^y \rightharpoonup \delta_{A^{-1}y} \quad \text{as } \gamma \to 0.$$

Proof. **(i):** As A has a trivial null space, $Au \neq 0$, and thus

$$(u, A^{\mathrm{T}}\Gamma_0^{-1}Au) = (Au, \Gamma_0^{-1}Au) > 0$$

for all $u \in \mathbb{R}^d \setminus \{0\}$. Therefore, the matrix $A^{\mathrm{T}}\Gamma_0^{-1}A$ is invertible. Now we can take γ to zero in (2) and (3) and get

$$m(\gamma) = \left(A^{\mathrm{T}}\Gamma_0^{-1}A + \gamma^2\Gamma_{\mathrm{pr}}^{-1}\right)^{-1}\left(A^{\mathrm{T}}\Gamma_0^{-1}y + \gamma^2\Gamma_{\mathrm{pr}}^{-1}m_0\right) \overset{\gamma \to 0}{\rightarrow} (A^{\mathrm{T}}\Gamma_0^{-1}A)^{-1}A^{\mathrm{T}}\Gamma_0^{-1}y =: m^*$$

as well as $C(\gamma) = \gamma^2\left(A^{\mathrm{T}}\Gamma_0^{-1}A + \gamma^2\Gamma_{\mathrm{pr}}^{-1}\right)^{-1} \overset{\gamma \to 0}{\rightarrow} 0$. This shows that $\pi^y = \mathcal{N}(m, C) \rightharpoonup \mathcal{N}(m^*, 0) = \delta_{m^*}$.

Due to the trivial null space of A, the minimizer $m^\dagger$ of

$$\left\|\Gamma_0^{-\frac{1}{2}}(Au - y)\right\|^2$$

is the unique solution to the normal equation

$$A^{\mathrm{T}}\Gamma_0^{-1}Am^\dagger = A^{\mathrm{T}}\Gamma_0^{-1}y,$$

which shows that $m^* = m^\dagger$.

(ii): As in part (i), we have $m(\gamma) \to m^*$ and $C(\gamma) \to 0$. Since A is now invertible, we obtain

$$m^* = \left(A^{-1}\Gamma_0(A^{\mathrm{T}})^{-1}\right)A^{\mathrm{T}}\Gamma_0^{-1}y = A^{-1}y. \quad \square$$

Reminder: singular value decomposition (SVD)

Let $A \in \mathbb{R}^{k \times d}$ be *any* matrix. Then we can *always* write

$$A = U \Lambda V^{\mathrm{T}},$$

where $U \in \mathbb{R}^{k \times k}$, $\Lambda \in \mathbb{R}^{k \times d}$, and $V \in \mathbb{R}^{d \times d}$ are matrices such that

$$UU^{\mathrm{T}} = U^{\mathrm{T}} U = I_k \quad \text{and} \quad VV^{\mathrm{T}} = V^{\mathrm{T}} V = I_d \qquad (U \text{ and } V \text{ are } \textit{orthogonal} \text{ matrices})$$

and

$$\Lambda = \left(\begin{array}{ccc|c} \sigma_1 & & & \\ & \ddots & & O_{k \times (d-k)} \\ & & \sigma_k & \end{array} \right) \quad \text{if } k < d,$$

$$\Lambda = \left(\begin{array}{ccc} \sigma_1 & & \\ & \ddots & \\ & & \sigma_d \\ \hline & O_{(k-d) \times d} & \end{array} \right) \quad \text{if } k > d,$$

and $\Lambda = \mathrm{diag}(\sigma_1, \ldots, \sigma_k)$ if $k = d$, where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{\min\{k,d\}} \geq 0$ are called the *singular values* of matrix A.

Underdetermined case

Both in the overdetermined and the determined case, the small noise limit of the posterior distribution is a Dirac distribution. Note that the prior plays no role in the limit.

This case is of particular relevance because practical inverse problems are usually underdetermined. Here, we assume that the matrix $A \in \mathbb{R}^{k \times d}$ has $\mathrm{Rank}(A) = k < d$ and write

$$A \overset{(*)}{=} \begin{pmatrix} A_1 & 0 \end{pmatrix} Q^{\mathrm{T}} = \begin{pmatrix} A_1 & 0 \end{pmatrix} \begin{pmatrix} Q_1 & Q_2 \end{pmatrix}^{\mathrm{T}} = A_1 Q_1^{\mathrm{T}} \tag{4}$$

with an invertible matrix $A_1 \in \mathbb{R}^{k \times k}$ and an orthogonal matrix $Q = \begin{pmatrix} Q_1 & Q_2 \end{pmatrix} \in \mathbb{R}^{d \times d}$ (i.e., $Q^{\mathrm{T}} Q = Q Q^{\mathrm{T}} = I_d$).

To get an idea of what is going on in the underdetermined case, we first consider a basic example.

(*) To see this, consider the SVD $A = U \Lambda V^{\mathrm{T}}$. Since $k < d$, we have $\Lambda =: \begin{pmatrix} \Lambda_1 & 0 \end{pmatrix}$ with $\Lambda_1 = \mathrm{diag}(\sigma_1, \ldots, \sigma_k)$; thus $A = U \Lambda V^{\mathrm{T}} = U \begin{pmatrix} \Lambda_1 & 0 \end{pmatrix} V^{\mathrm{T}} = \begin{pmatrix} U \Lambda_1 & 0 \end{pmatrix} V^{\mathrm{T}}$. Finally, define $A_1 := U \Lambda_1$ (invertible) and $Q := V$ (orthogonal).

Example. Assume that $A = \begin{pmatrix} A_1 & 0 \end{pmatrix}$, $\eta \sim \mathcal{N}(0, \gamma^2 I_k)$, and $x \sim \mathcal{N}(0, I_d)$. Let

$$x =: \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

with $x_1 \in \mathbb{R}^k$ and $x_2 \in \mathbb{R}^{d-k}$. Then, the data satisfies

$$y = Ax + \eta = A_1 x_1 + \eta.$$

The posterior density is given by $\pi^y(x) = \frac{1}{Z} \exp(-J(x))$, where

$$\begin{aligned} J(x) &= \frac{1}{2\gamma^2} \|y - A_1 x_1\|^2 + \frac{1}{2} \|x\|^2 \\ &= \left(\frac{1}{2\gamma^2} \|y - A_1 x_1\|^2 + \frac{1}{2} \|x_1\|^2 \right) + \frac{1}{2} \|x_2\|^2, \end{aligned}$$

and Z is a normalization constant.

We can write it as a product

$$\pi^y(x_1, x_2) = \frac{1}{\tilde{\tilde{Z}}}\nu(y - A_1 x_1)\pi_1(x_1) \cdot \pi_2(x_2) =: \pi_1^y(x_1)\pi_2(x_2)$$

where $\pi_1(x_1) = \mathcal{N}(0, I_k)$ and $\pi_2(x_2) = \mathcal{N}(0, I_{d-k})$ are Gaussian densities. We can interpret the factor $\frac{1}{\tilde{\tilde{Z}}}\nu(y - A_1 x_1)\pi_1(x_1)$ as posterior density π_1^y resulting from the determined problem $y = A_1 x_1 + \eta$ with prior density $x_1 \sim \pi_1$. By the small noise limit in the determined case, we know that $\pi_1^y \rightharpoonup \delta_{A_1^{-1}y}$ as $\gamma \to 0$, whereas π_2 remains constant. Since x_1 and x_2 are independent, we would expect the posterior distribution to converge weakly towards

$$\pi^y(x_1, x_2) \rightharpoonup \delta_{A_1^{-1}y}(x_1)\pi_2(x_2).$$

This means that in the limit, the data determines the posterior distribution on a subspace of dimension k, whereas uncertainty remains in a subspace of dimension $d - k$.

In order to generalize these observations, we need the following decomposition of the identity.

Lemma

Let $\Gamma_{\mathrm{pr}} \in \mathbb{R}^{d\times d}$ be symmetric and positive definite and $Q = \begin{pmatrix} Q_1 & Q_2 \end{pmatrix}$ an orthogonal matrix with $Q_1 \in \mathbb{R}^{d\times k}$, $Q_2 \in \mathbb{R}^{d\times(d-k)}$. Then we have

$$I_d = \Gamma_{\mathrm{pr}} Q_1 (Q_1^{\mathrm{T}} \Gamma_{\mathrm{pr}} Q_1)^{-1} Q_1^{\mathrm{T}} + Q_2 (Q_2^{\mathrm{T}} \Gamma_{\mathrm{pr}}^{-1} Q_2)^{-1} Q_2^{\mathrm{T}} \Gamma_{\mathrm{pr}}^{-1}. \tag{5}$$

Proof. Let R denote the right-hand side of (5). Since Q is orthogonal, we have $Q_1^{\mathrm{T}} Q_2 = Q_2^{\mathrm{T}} Q_1 = 0$, and thus

$$Q_1^{\mathrm{T}}(R - I_d) = 0, \quad Q_2^{\mathrm{T}} \Gamma_{\mathrm{pr}}^{-1}(R - I_d) = 0.$$

If $B := \begin{pmatrix} Q_1 & \Gamma_{\mathrm{pr}}^{-1} Q_2 \end{pmatrix}$ has full rank, then the above identities, written as $B^{\mathrm{T}}(R - I_d) = 0$, imply $R = I$. B in turn is invertible, since

$$Q^{\mathrm{T}} B = \begin{pmatrix} Q_1^{\mathrm{T}} \\ Q_2^{\mathrm{T}} \end{pmatrix} \begin{pmatrix} Q_1 & \Gamma_{\mathrm{pr}}^{-1} Q_2 \end{pmatrix} = \begin{pmatrix} I_k & Q_1^{\mathrm{T}} \Gamma_{\mathrm{pr}}^{-1} Q_2 \\ 0 & Q_2^{\mathrm{T}} \Gamma_{\mathrm{pr}}^{-1} Q_2 \end{pmatrix}$$

is invertible and Q is orthogonal. □

Theorem (Underdetermined case)

Suppose in the linear Gaussian setting that $x \sim \mathcal{N}(x_0, \Gamma_{\mathrm{pr}})$*,* $\eta \sim \mathcal{N}(0, \gamma^2 \Gamma_0)$ *with* $\gamma > 0$*, and that* $\mathrm{Rank}(A) = k < d$*. Then*

$$\pi^y \rightharpoonup \mathcal{N}(m^*, C^*),$$

where

$$\begin{aligned} m^* &= \Gamma_{\mathrm{pr}} Q_1 (Q_1^{\mathrm{T}} \Gamma_{\mathrm{pr}} Q_1)^{-1} A_1^{-1} y + Q_2 (Q_2^{\mathrm{T}} \Gamma_{\mathrm{pr}}^{-1} Q_2)^{-1} Q_2^{\mathrm{T}} \Gamma_{\mathrm{pr}}^{-1} x_0, \\ C^* &= Q_2 (Q_2^{\mathrm{T}} \Gamma_{\mathrm{pr}}^{-1} Q_2)^{-1} Q_2^{\mathrm{T}}. \end{aligned}$$

Proof. Using the previous lemma, we can decompose x into

$$x = \underbrace{\Gamma_{\mathrm{pr}} Q_1 (Q_1^{\mathrm{T}} \Gamma_{\mathrm{pr}} Q_1)^{-1}}_{=:\, S} \underbrace{Q_1^{\mathrm{T}} x}_{=:\, x_1} + \underbrace{Q_2 (Q_2^{\mathrm{T}} \Gamma_{\mathrm{pr}}^{-1} Q_2)^{-1}}_{=:\, T} \underbrace{Q_2^{\mathrm{T}} \Gamma_{\mathrm{pr}}^{-1} x}_{=:\, x_2} = S x_1 + T x_2.$$

This way, $x_1 = Q_1^{\mathrm{T}} x$ and $x_2 = Q_2^{\mathrm{T}} \Gamma_{\mathrm{pr}}^{-1} x$ are Gaussian, and[†]

$$x_2 \sim \mathcal{N}(Q_2^{\mathrm{T}} \Gamma_{\mathrm{pr}}^{-1} x_0, Q_2^{\mathrm{T}} \Gamma_{\mathrm{pr}}^{-1} Q_2).$$

Now x_1 and x_2 are independent, since

$$\begin{aligned} \mathrm{Cov}(x_1, x_2) &= \mathbb{E}[(x_1 - \mathbb{E}\, x_1)(x_2 - \mathbb{E}\, x_2)^{\mathrm{T}}] \\ &= Q_1^{\mathrm{T}}\, \mathbb{E}[(x - \mathbb{E}\, x)(x - \mathbb{E}\, x)^{\mathrm{T}}] \Gamma_{\mathrm{pr}}^{-1} Q_2 \\ &= Q_1^{\mathrm{T}} Q_2 = 0, \end{aligned}$$

where we used $\mathrm{Cov}(x, x) = \mathbb{E}[(x - \mathbb{E}\, x)(x - \mathbb{E}\, x)^{\mathrm{T}}] = \Gamma_{\mathrm{pr}}$.(*)

(*) Note that, in general, *uncorrelated random variables are not necessarily independent*. However, this assertion is true for jointly Gaussian random variables.

[†] Recall task 4 of exercise 6: if $z \sim \mathcal{N}(m, C)$, then $Lz + a \sim \mathcal{N}(Lm + a, LCL^{\mathrm{T}})$.

By (4), we have

$$y = Ax + \eta = A_1 Q_1^{\mathrm{T}} x + \eta = A_1 x_1 + \eta. \tag{6}$$

As $\eta \perp x$, this implies $x_2 \perp y, x_1$ and hence $\mathbb{P}(x_1, x_2|y) = \mathbb{P}(x_1|y)\,\mathbb{P}(x_2)$. The random variable x_1 is Gaussian, so problem (6) satisfies the assumptions of the linear Gaussian setting, and thus the posterior distribution $\mathbb{P}(x_1|y)$ is Gaussian. The small noise limit in the determined case in turn shows that $\mathbb{P}(x_1|y) \rightharpoonup \delta_{A_1^{-1}y}(x_1)$ as $\gamma \to 0$. As a consequence, the limiting posterior distribution of $(x_1, x_2)|y$ is

$$\mathbb{P}(x_1, x_2|y) \rightharpoonup \delta_{A_1^{-1}y}(x_1)\,\mathbb{P}(x_2).$$

Now, the mean and covariance of the limiting posterior distribution of $x|y$ are given by

$$\begin{aligned} m^* &= \mathbb{E}[Sx_1 + Tx_2|y] = SA_1^{-1}y + T\,\mathbb{E}[x_2] \\ &= SA_1^{-1}y + TQ_2^{\mathrm{T}}\Gamma_{\mathrm{pr}}^{-1}x_0, \\ C^* &= \mathrm{Var}(Sx_1 + Tx_2|y) = \mathrm{Var}(Sx_1|y) + \mathrm{Var}(Tx_2) \\ &= TQ_2^{\mathrm{T}}\Gamma_{\mathrm{pr}}^{-1}Q_2T^{\mathrm{T}} = Q_2(Q_2^{\mathrm{T}}\Gamma_{\mathrm{pr}}^{-1}Q_2)^{-1}Q_2^{\mathrm{T}}. \quad \square \end{aligned}$$

Q: How to interpret the limiting distribution in the underdetermined case?

A: Uncertainty remains in the subspace $\mathrm{Ker}(A) = \mathrm{Ran}(Q_2)$ of dimension $d - k$, where the posterior is fully described by the prior.

Monte Carlo and Importance Sampling

Suppose that we are interested in estimating the integral

$$\pi(f) := \mathbb{E}^{\pi}[f(x)] := \int_{\mathbb{R}^d} f(x)\pi(x)\,\mathrm{d}x, \tag{7}$$

where π is a probability density function and $f\colon \mathbb{R}^d \to \mathbb{R}$ is a quantity of interest.

In the Bayesian framework, we have $\pi(x) = \frac{1}{Z}g(x)\rho(x)$, where Z is a normalization constant, π is the posterior, $g(x) := \nu(y - F(x))$ is the likelihood, and ρ is the prior. Note that here we change the notations slightly to improve readability.

In a non-Gaussian setting, we usually have to resort to approximating the integral (7) by means of sampling. To this end, we will consider the following techniques:

- The Monte Carlo method (today's lecture)
- Importance sampling (today's lecture)
- Markov Chain Monte Carlo (MCMC) methods (next week's lecture)

The Monte Carlo method

A simple technique to approximate the integral

$$\pi(f) = \int_{\mathbb{R}^d} f(x)\pi(x)\,\mathrm{d}x, \quad d \in \mathbb{Z}_+,$$

is to use a sample average. If we are able to draw the i.i.d. samples $x_1, \ldots, x_n$ from the probability distribution corresponding to π, then one can consider the Monte Carlo estimate

$$\pi_n^{\mathrm{MC}}(f) := \frac{1}{n}\sum_{i=1}^{n} f(x_i).$$

Generally speaking, the Law of Large Numbers and the Central Limit Theorem imply that

$$\lim_{n\to\infty} \pi_n^{\mathrm{MC}}(f) = \pi(f) \quad \text{and} \quad \mathrm{Var}(\pi_n^{\mathrm{MC}}(f) - \pi(f)) \approx \frac{\mathrm{Var}(f(X))}{n},$$

provided that $f(X)$ has finite mean and variance with X distributed according to the probability distribution that corresponds to π.

Some properties of the Monte Carlo estimator

If we have the i.i.d. random samples $x_1, \ldots, x_n$ distributed according to π, then π can be estimated by

$$\pi_n^{\mathrm{MC}} := \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}.$$

Theorem ([Theorem 5.1, Sanz-Alonso, Stuart, and Taeb 2018])

For $f : \mathbb{R}^d \to \mathbb{R}$, denote $\|f\|_\infty := \sup_{x \in \mathbb{R}^d} |f(x)|$. Then

$$\sup_{\|f\|_\infty \le 1} \left| \mathbb{E}\left[\pi(f) - \pi_n^{\mathrm{MC}}(f)\right] \right| = 0 \quad \textit{and} \quad \sup_{\|f\|_\infty \le 1} \left| \mathbb{E}\left[(\pi(f) - \pi_n^{\mathrm{MC}}(f))^2\right] \right| \le \frac{1}{n}.$$

This shows that the Monte Carlo estimator π_n^{MC} is an unbiased estimator of π. While the convergence rate is slow with respect to n, the error is independent of the dimension d or the properties of f, its supremum notwithstanding.

Proof. Let $x_1, \dots, x_n$ be i.i.d. according to π. Define

$$\bar{f}(x) = f(x) - \pi(f).$$

To prove the first result, namely that the estimator is unbiased, note that

$$\mathbb{E}\left[\pi_n^{\mathrm{MC}}(f) - \pi(f)\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[f\left(x_i\right) - \pi(f)] = \frac{1}{n}\sum_{i=1}^{n}\left(\pi\left(f\right) - \pi\left(f\right)\right)$$

$$= \frac{1}{n}\cdot 0 = 0.$$

Therefore the supremum of its absolute value is also zero. For the second result, which bounds the variance of the estimator, we observe that $\mathbb{E}[\bar{f}] = 0$ and, then,

$$\mathbb{E}\left[\left(\pi_n^{\mathrm{MC}}(f) - \pi(f)\right)^2\right] = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}\left[\bar{f}\left(x_i\right)\bar{f}\left(x_j\right)\right]$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left[\bar{f}\left(x_i\right)^2\right] = \frac{1}{n}\mathbb{E}\left[\bar{f}\left(x_1\right)^2\right] = \frac{1}{n}\mathrm{Var}_\pi[f]$$

since x_i are i.i.d.

In particular we have

$$\mathbb{E}\Big[\big(\pi_n^{\mathrm{MC}}(f)-\pi(f)\big)^2\Big]=\frac{1}{n}\mathrm{Var}_\pi[f]\leq\frac{1}{n}\pi(f^2) \tag{8}$$

since

$$\mathrm{Var}_\pi[f]=\pi(f^2)-\pi(f)^2\leq\pi(f^2).$$

Therefore

$$\sup_{\|f\|_\infty\leq 1}\left|\mathbb{E}\Big[\big(\pi_n^{\mathrm{MC}}(f)-\pi(f)\big)^2\Big]\right|=\sup_{\|f\|_\infty\leq 1}\left|\frac{1}{n}\mathrm{Var}_\pi[f]\right|\leq\frac{1}{n}. \quad \square$$

Example

Suppose that we have the PDF $\pi(x) := (6x - 6x^2)\chi_{(0,1)}(x)$ and $f(x) = x$. We can design the following simple scheme based on inverse transform sampling to draw samples from this distribution.

MATLAB implementation:

```
n = 1e5; % sample size
x = linspace(0,1);
p = @(x) 6*x-6*x.^2; % PDF
P = cumsum(p(x)); P = P/P(end); % "empirical" CDF of p
samples = [];
for iter = 1:n
  u = rand; % realization of U(0,1)
  ind = find(u <= P,1,'first'); % inverse CDF rule
  samples = [samples,x(ind)]; % store sample
end
histogram(samples,'Normalization','pdf'); % draw a histogram
hold on, plot(x,p(x),'LineWidth',3), legend('samples','pdf');
hold off;
```

Python implementation:

```
import numpy as np
import matplotlib.pyplot as plt
n = int(1e5) # sample size
x = np.linspace(0,1,1000)
p = lambda x: 6*x-6*x**2 # PDF
P = np.cumsum(p(x)); P = P/P[-1] # "empirical" CDF of p
samples = []
for iter in range(n):
    u = np.random.uniform() # realization of U(0,1)
    ind = np.where(u<=P)[0][0] # inverse CDF rule
    samples.append(x[ind]) # store sample
plt.hist(samples,bins=’auto’,
         density=True,label=’samples’) # draw a histogram
plt.plot(x,p(x),linewidth=2,label=’pdf’)
plt.legend()
plt.show()
# Thanks to Subodh Khanger for the Python implementation!
```

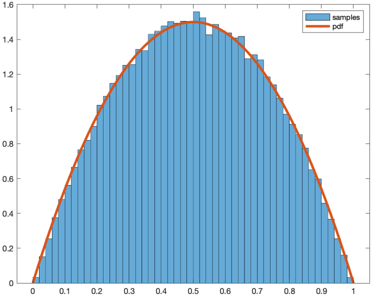


Figure: 10^5 samples drawn from the distribution given on the previous page organized as a histogram.

MATLAB:

```
>> mean(samples) % Monte Carlo estimate of the mean


ans =


    0.5001
```

Python: `np.mean(samples) # Monte Carlo estimate of the mean`

Example

Use Monte Carlo to estimate the value of $\int_{\mathbb{R}^2} \chi_{\{x^2+y^2<1\}}(x,y)\,\mathrm{d}x\,\mathrm{d}y$.

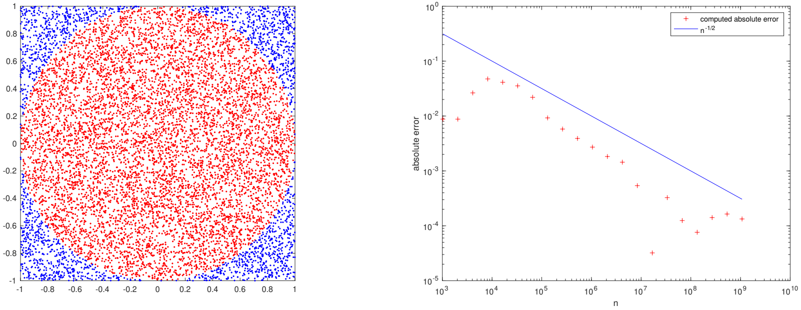


Figure: Left: 2^{13} samples drawn from $U((-1,1)^2)$. We calculate the value of the integral as $4 \cdot \frac{\#\text{samples inside unit disk}}{\#\text{samples inside unit disk} + \#\text{samples outside unit disk}}$. Right: the absolute integration error for $n = 2^k$, $k \in \{10, \ldots, 30\}$.

Sample average at $n = 2^{30}$: `3.141725998371840`.

Importance sampling

Let us focus on the setting

$$\pi(x) = \frac{1}{Z} g(x)\rho(x), \tag{9}$$

where Z is a normalization constant. Unless π is some well-understood distribution (e.g., Gaussian), the basic Monte Carlo method is generally infeasible due to the difficulties associated with drawing samples from π directly in the high-dimensional setting.

An alternative tactic is to use ρ as a *proposal density*, drawing samples from it instead of π. By substituting the identity (9) into $\pi(f)$, we obtain

$$\pi(f) = \int_{\mathbb{R}^d} f(x)\pi(x)\,\mathrm{d}x = \frac{\int_{\mathbb{R}^d} (f(x)g(x))\,\rho(x)\,\mathrm{d}x}{\int_{\mathbb{R}^d} g(x)\,\rho(x)\,\mathrm{d}x}.$$

If the samples $x_1, \ldots, x_n$ are now distributed i.i.d. according to ρ, we can replace the numerator and denominator by their respective Monte Carlo estimates:

$$\pi_n^{\mathrm{IS}}(f) := \sum_{i=1}^{n} w_i f(x_i), \quad w_i := \frac{g(x_i)}{\sum_{j=1}^{n} g(x_j)} \quad (\text{“importance weights”}).$$

Similarly to the Monte Carlo estimator, we can define the *particle approximation measure*

$$\pi_n^{\mathrm{IS}} := \sum_{i=1}^{n} w_i \delta_{x_i}, \quad w_i := \frac{g(x_i)}{\sum_{j=1}^{n} g(x_j)}.$$

Theorem ([Theorem 5.4, Sanz-Alonso, Stuart, and Taeb 2018])

$$\sup_{\|f\|_\infty \le 1} \left| \mathbb{E}\left[\pi_n^{\mathrm{IS}}(f) - \pi(f)\right] \right| \le 2\, \frac{1 + d_{\chi^2}(\pi \| \rho)}{n},$$

$$\sup_{\|f\|_\infty \le 1} \left| \mathbb{E}\left[(\pi_n^{\mathrm{IS}}(f) - \pi(f))^2\right] \right| \le 4\, \frac{1 + d_{\chi^2}(\pi \| \rho)}{n},$$

where the χ^2 divergence of two probability distributions $\pi, \pi' > 0$ is defined as

$$d_{\chi^2}(\pi \| \pi') := \int_{\mathbb{R}^d} \left(\frac{\pi(x)}{\pi'(x)} - 1 \right)^2 \pi'(x) \,\mathrm{d}x.$$

Unlike Monte Carlo, π_n^{IS} is biased for π. The χ^2 divergence between π and ρ should not be too large for importance sampling to be accurate.

Proof. Let $x_1, \dots, x_n$ be i.i.d. according to ρ. Given

$$\pi(x) = \frac{1}{Z} g(x)\rho(x) = \frac{1}{\rho(g)} g(x)\rho(x),$$

we obtain

$$d_{\chi^2}(\pi \| \rho) = \int_{\mathbb{R}^d} \left(\frac{\pi(x)}{\rho(x)} - 1 \right)^2 \rho(x)\,\mathrm{d}x = \int_{\mathbb{R}^d} \left(\frac{g(x)}{Z} - 1 \right)^2 \rho(x)\,\mathrm{d}x$$
$$= \underbrace{\int_{\mathbb{R}^d} \frac{g(x)^2 \rho(x)}{Z^2}\,\mathrm{d}x}_{=\frac{\rho(g^2)}{\rho(g)^2}} - 2\frac{1}{Z} \underbrace{\int_{\mathbb{R}^d} g(x)\rho(x)\,\mathrm{d}x}_{=Z} + \underbrace{\int_{\mathbb{R}^d} \rho(x)\,\mathrm{d}x}_{=1} = \frac{\rho(g^2)}{\rho(g)^2} - 1.$$

Let $\zeta := \frac{\rho(g^2)}{\rho(g)^2}$. Noting that

$$\pi(f) = \frac{\rho(gf)}{\rho(g)} \approx \frac{\rho_n^{\mathrm{MC}}(gf)}{\rho_n^{\mathrm{MC}}(g)} = \pi_n^{\mathrm{IS}}(f),$$

it follows that

$$\begin{aligned} \pi_n^{\mathrm{IS}}(f) - \pi(f) &= \pi_n^{\mathrm{IS}}(f) - \frac{\rho(gf)}{\rho(g)} \\ &= \frac{\pi_n^{\mathrm{IS}}(f)\left(\rho(g) - \rho_n^{\mathrm{MC}}(g)\right)}{\rho(g)} - \frac{\left(\rho(gf) - \rho_n^{\mathrm{MC}}(gf)\right)}{\rho(g)}. \end{aligned} \tag{10}$$

Let us prove the second inequality first. We use the splitting of $\pi_n^{\mathrm{IS}}(f) - \pi(f)$ into the sum of two terms from the previous slide together with $\mathbb{E}[(\rho(f) - \rho_n^{\mathrm{MC}}(f))^2] \leq \frac{1}{n}\rho(f^2)$ (see (8)) and the inequality $(a-b)^2 \leq 2(a^2+b^2)$ such that for all $\|f\|_\infty \leq 1$ we have $|\pi_n^{\mathrm{IS}}(f)| \leq 1$ and

$$\begin{aligned}
&\left|\mathbb{E}\left[\left(\pi_n^{\mathrm{IS}}(f) - \pi(f)\right)^2\right]\right| \\
&\leq \frac{2}{\rho(g)^2}\left(\mathbb{E}\left[\left(\pi_n^{\mathrm{IS}}(f)\right)^2\left(\rho(g) - \rho_n^{\mathrm{MC}}(g)\right)^2\right] + \mathbb{E}\left[\left(\rho(gf) - \rho_n^{\mathrm{MC}}(gf)\right)^2\right]\right) \\
&\leq \frac{2}{\rho(g)^2}\left(\mathbb{E}\left[\left(\rho(g) - \rho_n^{\mathrm{MC}}(g)\right)^2\right] + \mathbb{E}\left[\left(\rho(gf) - \rho_n^{\mathrm{MC}}(gf)\right)^2\right]\right) \\
&= \frac{2}{\rho(g)^2 n}\left(\mathrm{Var}_\rho\left[g\right] + \mathrm{Var}_\rho\left[gf\right]\right) \\
&\leq \frac{2}{\rho(g)^2 n}\left(\rho(g^2) + \rho(g^2 f^2)\right) \leq \frac{4}{n}\frac{\rho(g^2)}{\rho(g)^2} = \frac{4\zeta}{n}.
\end{aligned}$$

Therefore, since $\zeta = d_{\chi^2}(\pi\|\rho) + 1$, we obtain

$$\sup_{|f|_\infty \leq 1}\left|\mathbb{E}\left[\left(\pi_n^{\mathrm{IS}}(f) - \pi(f)\right)^2\right]\right| \leq 4\,\frac{1 + d_{\chi^2}(\pi\|\rho)}{n}.$$

To prove the first inequality, we start again with the splitting (10), i.e.,

$$\pi_n^{\mathrm{IS}}(f) - \pi(f) = \frac{\pi_n^{\mathrm{IS}}(f)\left(\rho(g) - \rho_n^{\mathrm{MC}}(g)\right)}{\rho(g)} - \frac{\left(\rho(gf) - \rho_n^{\mathrm{MC}}(gf)\right)}{\rho(g)}.$$

The expectation of the second term vanishes since

$$\left|\mathbb{E}\left[\frac{\rho(gf) - \rho_n^{\mathrm{MC}}(gf)}{\rho(g)}\right]\right| = \frac{1}{\rho(g)}\left|\mathbb{E}\left[\rho(gf) - \rho_n^{\mathrm{MC}}(gf)\right]\right| = 0.$$

The Cauchy–Schwarz inequality together with $\mathbb{E}[(\rho(g) - \rho_n^{\mathrm{MC}}(g))^2] \leq \frac{1}{n}\rho(g^2)$ (see (8)) and the previous result yield that

$$\begin{aligned}
\left|\mathbb{E}\left[\pi_n^{\mathrm{IS}}(f) - \pi(f)\right]\right| &= \frac{1}{\rho(g)}\left|\mathbb{E}\left[\pi_n^{\mathrm{IS}}(f)\left(\rho(g) - \rho_n^{\mathrm{MC}}(g)\right)\right]\right| \\
&\leq \frac{1}{\rho(g)}\left|\mathbb{E}\left[\left(\pi_n^{\mathrm{IS}}(f) - \pi(f)\right)\left(\rho(g) - \rho_n^{\mathrm{MC}}(g)\right)\right] + \pi(f)\underbrace{\mathbb{E}\left[\left(\rho(g) - \rho_n^{\mathrm{MC}}(g)\right)\right]}_{=0}\right| \\
&\leq \frac{1}{\rho(g)}\left(\mathbb{E}\left[\left(\pi_n^{\mathrm{IS}}(f) - \pi(f)\right)^2\right]\right)^{1/2}\left(\mathbb{E}\left[\left(\rho(g) - \rho_n^{\mathrm{MC}}(g)\right)^2\right]\right)^{1/2} \\
&\leq \frac{1}{\rho(g)}\left(\frac{4\zeta}{n}\right)^{1/2}\left(\frac{\rho(g^2)}{n}\right)^{1/2} = \frac{2\zeta}{n} = 2\,\frac{d_{\chi^2}(\pi\|\rho) + 1}{n}. \qquad \square
\end{aligned}$$

Case study: source localization

Suppose that a particle with unit charge is located at some (unknown) point $x^* \in (0,1)$ and our goal is to locate it based on measurements of voltage at the interval end points $x = 0$ and $x = 1$. The mathematical model for the voltage at any point $x \in [0,1]$ is given by

$$y(x) = \frac{1}{|x^* - x|}.$$

Our noisy measurements are modeled by $y_1 = \frac{1}{|x^*-0|} + \eta_1$ and $y_2 = \frac{1}{|x^*-1|} + \eta_2$, where η_1 and η_2 are i.i.d. realizations of $\mathcal{N}(0,\sigma^2)$ with $\sigma = 0.2$.

- The likelihood is given by $\mathbb{P}(y|x) \propto \exp\left(-\frac{1}{2\sigma^2}\sum_{j=0}^{1}\left(y_j - \frac{1}{|x-j|}\right)^2\right)$.
- We consider the prior $\pi(x) = \chi_{(0,1)}(x) = \begin{cases} 1 & \text{if } x \in (0,1), \\ 0 & \text{otherwise.} \end{cases}$

Then the posterior density is given by Bayes' formula

$$\pi^y(x) \propto \chi_{(0,1)}(x) \exp\left(-\frac{1}{2\sigma^2}\sum_{j=0}^{1}\left(y_j - \frac{1}{|x-j|}\right)^2\right).$$

Computation of the CM estimate (MATLAB)

First, let us generate the measurements.

MATLAB:

```
format long
x_ast = 1/pi; % Fix "ground truth", i.e., particle location
sigma = .2; % Std for noise
v = 1./abs(x_ast-[0,1]); % Measurements at end points
v = v+sigma*randn(1,2); % Add noise
x = linspace(0,1); % Discretize the unit interval
% Define the (unnormalized) posterior density
p = @(x) exp(-1/(2*sigma^2)*((v(1)-1./abs(x-0)).^2+ ...
         (v(2)-1./abs(x-1)).^2));
```

```
%% Monte Carlo
n = 1e5;
P = cumsum(p(x)); P = P/P(end); % "empirical" CDF
% For the Monte Carlo method, we need to sample the posterior.
% We do this using inverse transform sampling.
samples = [];
for ii = 1:n
   u = rand; % realization of U(0,1)
   ind = find(u <= P,1,’first’); % inverse CDF rule
   samples = [samples,x(ind)]; % store sample
end
% Sanity check: plot samples in histogram.
histogram(samples,’Normalization’,’probability’, ...
                 ’BinWidth’,.01), axis([0,1,0,.25]);
hold on;
plot(x,p(x)/sum(p(x)),’LineWidth’,2), hold off;
title([num2str(n),’ samples from the posterior density’]);
mean(samples) % Monte Carlo estimate
```

```
%% Importance sampling
n = 1e5;
samples = rand(1,n); % Sample our prior, i.e., U(0,1)
weights = p(samples); % Compute the importance weights
weights = weights/sum(weights); % Normalize the weights

% Compute the IS estimate
dot(weights,samples)
```

Computation of the CM estimate (Python)

First, let us generate the measurements.

Python:

```
import numpy as np
x_ast = 1/np.pi # Fix "ground truth", i.e., particle location
sigma = .2 # Std for noise
v = 1/np.abs(x_ast-np.array([0,1])) # Measurements at
                                    # end points
v = v+sigma*np.random.normal(size=v.shape) # Add noise
x = np.linspace(0,1) # Discretize the unit interval
x = x[1:-1] # Drop end points to avoid numerical issues...
# Define the (unnormalized) posterior density
p = lambda x: (x > 0) * (x < 1) *\
             np.exp(-1/(2*sigma**2)*((v[0]-1/np.abs(x-0))**2\
             +(v[1]-1/np.abs(x-1))**2))
```

```
## Monte Carlo
n = int(1e5)
P = np.cumsum(p(x)); P = P/P[-1] # "empirical" CDF
# For the Monte Carlo method, we need to sample the posterior.
# We do this using inverse transform sampling.
samples = []
for ii in range(n):
    u = np.random.uniform() # realization of U(0,1)
    ind = np.where(u<=P)[0][0] # inverse CDF rule
    samples.append(x[ind]) # store sample

# Compute the Monte Carlo estimate
print(np.mean(samples))
```

```
## Importance sampling
n = int(1e5)
samples = np.random.uniform(size=(1,n)) # Sample our prior,
                                        # i.e., U(0,1)
weights = p(samples) # Compute the importance weights
weights = weights/np.sum(weights) # Normalize the weights

# Compute the IS estimate
print(np.sum(weights*samples))
```

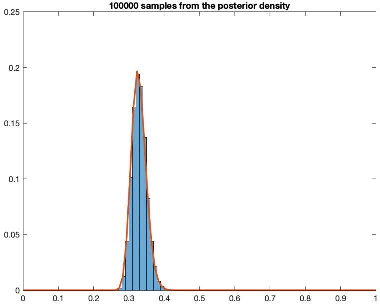


Figure: Histogram of the samples drawn from the posterior density.

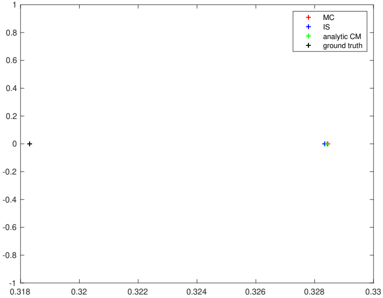
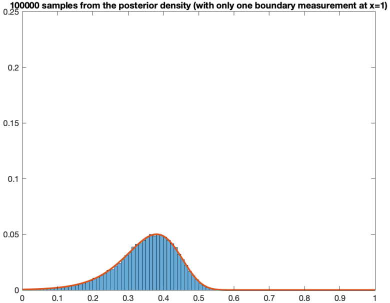


Figure: Comparison of MC and IS estimates vs. the analytic CM estimate and ground truth.

Monte Carlo estimate
`0.328444646464649`
Importance sampling estimate
`0.328340981036045`

Ground truth
`0.318309886183791`
Analytic CM estimate
`0.328421554655529`

What if we modify the problem so that we have access to only one boundary measurement at $x = 1$?



Monte Carlo estimate
0.349233333333324
Importance sampling estimate
0.349743141888635
Analytic CM estimate
0.349675613936670
Ground truth
0.318309886183791

The problem becomes substantially more ill-posed!

N.B. In the implementation above, a discretized version of the inverse transform sampling rule was used to obtain the MC estimate. The repeating digits are an artifact of the relatively coarse discretization used in the actual implementation.