Monte Carlo method

Vanilla Monte Carlo

Vanilla Monte Carlo

- Consider evaluating the statistical expectation $E_{x\sim p}[f(x)]$

$$I = E_{x\sim p}[f(x)] = \int_x f(x)p(x)dx$$

Note that any general integral $\int f(x)dx$ can be written as a statistical expectation.

$$\int f(x)dx = \int \frac{f(x)}{p(x)} p(x)dx = E_{x\sim p}\left[\frac{f(x)}{p(x)}\right]$$

- We can estimate the integral using Riemann discretisation. taking evenly (deterministic) spaced pts in the domain of x (pts are seperated by Δ)

$$\hat{I} = \sum_{i=1}^{N} \Delta_i f(x_i) p(x_i)$$

However, this has a few drawbacks

↳ It does not scale well w/ the dim. of x (req. many pts, hard to discretise / choose grid)

↳ The discretisation error is hard to characterise

↳ Wasteful in regions w/ small contributions to the integral.

- We can alternatively estimate the integral using a vanilla Monte Carlo w/ N samples

$$\hat{I} = \frac{1}{N}\sum_{i=1}^{N} f(x_i)$$

where x_i are drawn from $p(x)$, ie. $x_i \sim p$.

* Note that the MC estimate $\hat{I}$ is a RV → has properties like expectation and variance.

Expectation and variance of the MC estimator

- The expectation of the MC estimator $\hat{I}$ is given by

$$E[\hat{I}] = \frac{1}{N}\sum_{i=1}^{N} E_{x_i\sim p}[f(x_i)] = \frac{1}{N}\sum_{i=1}^{N} I = I$$

→ The MC estimator is unbiased ($E[\hat{I}] = I$)

- The variance of the MC estimator $\hat{I}$ is given by

$$Var[\hat{I}] = \frac{1}{N^2}\sum_{i=1}^{N} Var[f(x_i)] = \frac{1}{N^2}\sum_{i=1}^{N} \sigma_f^2 = \frac{\sigma_f^2}{N}$$

→ The variance of the MC estimator decreases linearly w/ the no. of samples N but also depends on σ_f^2 so if it is exhibitant, even large values of N will not help.

* Since $E[\hat{I}] = I$ and $\lim_{N\to\infty} Var[\hat{I}] = 0$, the MC estimator is consistent

Limit theorems for MC estimate

- The CLT states if $Y_1, Y_2 \dots$ be a sequence of iid RV w/ mean μ and variance $\sigma^2 < \infty$, and defining $S_n = \frac{1}{n}\sum_{i=1}^n Y_i$, then $\frac{S_n - \mu}{\sigma/\sqrt{n}} \to N(0,1)$ as $n \to \infty$
  - ↳ WLOG of generality, set $\mu = 0$. so $Z_n = \frac{S_n}{\sigma/\sqrt{n}} = \frac{S_n \sqrt{n}}{\sqrt{\sigma^2 n}}$

  consider the mgf of Z_n, $M_{Z_n}(t) = E[\exp(-tZ_n)]$

  $$M_{Z_n}(t) = E[\exp(-tZ_n)] = E\left[\left(\exp\left(-t\frac{S_n}{\sqrt{\sigma^2/n}}\right)\right)^n\right] = \left(E\left[\left(1 - \frac{t}{\sqrt{\sigma^2/n}}S_n + \frac{t^2}{\sigma^2/n}S_n^2 + \dots\right)\right]\right)^n$$
  $$= \left(1 - \frac{t}{\sqrt{\sigma^2/n}}\cancel{E[S_n]}^{\,0} + \frac{t^2}{\sigma^2/n}\cancel{E[S_n^2]}^{\,\sigma^2} + \dots\right)^n = \left(1 + \frac{t^2}{2n} + \dots\right)^n$$

  Noting that $\lim_{n\to\infty}(1+\frac{x}{n})^n = e^x$, $\lim_{n\to\infty} M_{Z_n}(t) = \lim_{n\to\infty}(1+\frac{t^2}{2n}+\dots)^n = \exp(\frac{t^2}{2})$

  which is the mgf of $N(0,1)$.

- The SLLN states if $Y_1, Y_2 \dots$ be a sequence of iid RV w/ mean μ, and defining $S_n = \frac{1}{n}\sum_{i=1}^n Y_i$, then $S_n \to \mu$ w/ probability 1 as $n \to \infty$.

- Set $Y_i = f(X_i)$, then the MC estimate $\hat{I}$ becomes $\hat{I} = \frac{1}{N}\sum_{i=1}^N f(X_i) = \frac{1}{N}\sum_{i=1}^N Y_i = S_N$

  Applying CLT, for large N, $\boxed{\hat{I} \approx N\left(I, \frac{\sigma_f^2}{N}\right)}$

  Applying SLLN, $\boxed{\hat{I} \text{ is guaranteed to converge to } I \text{ w/ probability 1 as } N \to \infty}$.

Variance reduction

Importance sampling (IS)

- To estimate the integral $E_{x\sim p}[f(x)]$, we can use a proposal distribution $q(x)$ to sample from instead of $p(x)$

$$I = E_{x\sim p}[f(x)] = \int_x f(x)p(x)dx = \int_x f(x)\frac{p(x)}{q(x)}q(x)dx = E_{x\sim q}\left[f(x)\frac{p(x)}{q(x)}\right]$$

- Given a proposal function $q(x)$ s.t. $q(x) > 0$ whenever $p(x) > 0$, the IS estimate is

$$\boxed{\hat{I} = \frac{1}{N}\sum_{i=1}^N f(x_i)\frac{p(x_i)}{q(x_i)}}$$

  where x_i are drawn from $q(x)$, i.e. $x_i \sim q$

- The IS estimator is unbiased, $E[\hat{I}] = \frac{1}{N}\sum_{i=1}^N E_{x\sim q}\left[f(x_i)\frac{p(x_i)}{q(x_i)}\right] = \frac{1}{N}\sum_{i=1}^N I = I$

- Define $\tilde{f}(x) = \frac{f(x)p(x)}{q(x)}$. The vanilla MC estimate and IS estimate have errors:

  ↳ Vanilla MC: $\sigma_f^2 = \text{Var}[f(x)] = E_p[f(x)^2] - E_p[f(x)]^2 = E_p[f(x)^2] - I^2$

  ↳ IS: $\sigma_{\tilde{f}}^2 = \text{Var}[\tilde{f}(x)] = E_q[\tilde{f}(x)^2] - E_q[\tilde{f}(x)]^2 = E_q[\tilde{f}(x)^2] - I^2$

  ⇒ Diff in errors $\sigma_f^2 - \sigma_{\tilde{f}}^2 = E_p[f(x)^2] - E_q[\tilde{f}(x)^2] = \int f(x)^2 p(x)dx - \int \frac{f(x)^2 p(x)}{q(x)}p(x)dx$.

- By calculus of variations, the biggest decrease in error is achieved when $q(x) \propto |f(x)|p(x)$

$$\boxed{q_{opt}(x) = \frac{|f(x)|\,p(x)}{\int |f(x)|\,p(x)dx}}$$ ← intuitively, we want sample regions w/ large f → ↑ contributions

  where we have $E_{q_{opt}}\left[\left(f(x)\frac{p(x)}{q(x)}\right)^2\right] = \int_x \frac{f(x)^2 p(x)^2}{q(x)}dx = \left[\int_x f(x)p(x)dx\right]^2 = I^2 \to \boxed{\sigma_{\tilde{f}}^2 = 0}$.

* IS is useful for reducing estimator variance and is useful when we cannot draw samples from $p(x)$. (we don't even req. normalized distributions of p, q)

Control variates

- By we want to estimate $E[f(x)]$ and we have additional info in the form of another function $g(x)$ where we know the value of $E[g(x)]$.
- Denoting the MC estimates as

$$\hat{E}_f = \frac{1}{N}\sum_n f(x_n) \qquad \hat{E}_g = \frac{1}{N}\sum_n g(x_n)$$

An unbiased estimate follows as

$$\hat{I} = \hat{E}_f^c = \hat{E}_f + c(\hat{E}_g - E[g(x)])$$

where c controls the quality of the estimate.

- The estimator is unbiased $E[\hat{E}_f^c] = E[\hat{E}_f] + c(E[\hat{E}_g] - E[g(x)]) = E_f$

The estimator has variance $\mathrm{Var}[\hat{E}_f^c] = \mathrm{Var}[\hat{E}_f] + c^2\mathrm{Var}[\hat{E}_g] + 2c\,\mathrm{Cov}[\hat{E}_f, \hat{E}_g]$.

- The estimator has min. variance when

$$\frac{d\mathrm{Var}[\hat{E}_f^c]}{dc} = 2c\,\mathrm{Var}[\hat{E}_g] + 2\,\mathrm{Cov}[\hat{E}_f, \hat{E}_g] = 0 \rightarrow c_{opt} = -\frac{\mathrm{Cov}[\hat{E}_f, \hat{E}_g]}{\mathrm{Var}[\hat{E}_g]}$$

The variance of the estimator is thus

$$\mathrm{Var}[\hat{E}_f^{c_{opt}}] = \mathrm{Var}[\hat{E}_f] - \frac{\mathrm{Cov}[\hat{E}_f, \hat{E}_g]^2}{\mathrm{Var}[\hat{E}_g]}$$

i.e. if $\mathrm{Cov}[\hat{E}_f, \hat{E}_g] \neq 0$, the variance of the estimator is reduced.

- Several control variates can be used at the same time

$$\hat{I} = \hat{E}_f^c = \hat{E}_f + \sum_{i=1}^{M} c_i(\hat{E}_{g_i} - E[g_i(x)])$$

- The estimator is unbiased $E[\hat{E}_f^c] = E[\hat{E}_f] + \sum_{i=1}^{M} c_i(E[\hat{E}_{g_i}] - E[g_i(x)]) = E_f$

The estimator has variance $\mathrm{Var}[\hat{E}_f^c] = \mathrm{Var}[\hat{E}_f] + 2\underline{c}^T\underline{b} + \underline{c}^T\underline{\underline{C}}\,\underline{c}$

where $\underline{c} \in \mathbb{R}^M$ has elements c_i, $\underline{b} \in \mathbb{R}^M$ has elements $\mathrm{Cov}[\hat{E}_f, \hat{E}_{g_i}]$, $\underline{\underline{C}} \in \mathbb{R}^{M\times M}$ has elements $\mathrm{Cov}[\hat{E}_{g_i}, \hat{E}_{g_j}]$

- The estimator has min. variance when

$$\frac{d\mathrm{Var}[\hat{E}_f^c]}{d\underline{c}} = 2\underline{b} + 2\underline{\underline{C}}\,\underline{c} = 0 \rightarrow \underline{c}_{opt} = -\underline{\underline{C}}^{-1}\underline{b}$$

The variance of the estimator is thus

$$\mathrm{Var}[\hat{E}_f^{c_{opt}}] = \mathrm{Var}[\hat{E}_f] - \underline{b}^T\underline{\underline{C}}^{-1}\underline{b}$$

Measure theory and integration

Measure theory

Power set and σ-algebras

- Let Ω be a set. The power set $\mathcal{P}$ on Ω is the collection of all subsets of Ω, i.e.

$$\mathcal{P} = \{A : A \subseteq \Omega\}$$

↳ e.g. $\Omega = \{a,b,c\} \rightarrow \mathcal{P} = \{\phi, \Omega, \{a\}, \{b\}, \{c\}, \{a,b\}, \{b,c\}, \{c,a\}\}$.

- Let Ω be a set. A σ-algebra $\mathcal{F}$ on Ω is a collection of subsets of Ω that satisfies

(i) $\phi, \Omega \in \mathcal{F}$

(ii) $A \in \mathcal{F} \Rightarrow A' = \Omega \setminus A \in \mathcal{F}$

(iii) For any countable seq (A_n), $n \in \mathbb{N}$ in $\mathcal{F}$, we have that $\bigcup_{n\in\mathbb{N}} A_n \in \mathcal{F}$.

⇒ Any set $A \in \mathcal{F}$ is a measurable set, and

The pair $(\Omega, \mathcal{F})$ is a measurable space

- Let Ω be a set, and $M \subseteq \mathcal{P}(\Omega)$ be a collection of subsets of Ω. The σ-algebra generated by M, (the smallest σ-algebra $\mathcal{F}(\Omega)$ on Ω that contains M), $\sigma(M)$ is

$$\sigma(M) = \{M \subseteq \Omega : M \in \mathcal{F} \ \forall\ \sigma\text{-algebras } \mathcal{F} \text{ that contain } M\}.$$

↳ eg: $\Omega = \{a,b,c,d\}$, $M = \{\{a\},\{b\}\} \rightarrow \sigma(M) = \{\phi, \Omega, \{a\}, \{b\}, \{a,b\}, \{b,c,d\}, \{a,c,d\}, \{c,d\}\}$.

eg: $\Omega = \mathbb{Z}$, $M = \{\{x\} : x \in \mathbb{Z}\} \rightarrow \sigma(M) = \mathcal{P}(\Omega)$

- Let $\Omega = \mathbb{R}$, $M = \{U \subseteq \mathbb{R} : U \text{ is open}\}$, then $\sigma(M)$ is the Borel σ-algebra $\mathcal{B}(\mathbb{R})$, i.e. it is the smallest σ-algebra that contains all open sets.

Measure and measure space

- A measure on a measurable space $(\Omega, \mathcal{F})$ is a function $M : \mathcal{F} \rightarrow [0, \infty]$ that satisfies

(i) $M(\phi) = 0$

(ii) For any disjoint seq. (A_n) in $\mathcal{F}$, we have that $M\left(\bigcup_{n\in\mathbb{N}} A_n\right) = \sum_{n\in\mathbb{N}} M(A_n)$

⇒ We call $(\Omega, \mathcal{F}, M)$ a measure space. The measure space is finite if $M(\Omega) < \infty$.

- Consider the Borel σ-algebra $\mathcal{B}(\mathbb{R})$. We can define the Lebesgue measure $M_L(A)$, $\lambda(A)$ of set A

$$M_L(A) = \lambda(A) = \inf\left\{\sum_{n\in\mathbb{N}} b_n - a_n : a_n < b_n \text{ are s.t. } A \subseteq \bigcup_{n\in\mathbb{N}} [a_n, b_n)\right\}$$

i.e. for the simple case $A = [a,b]$, $M_L(A) = b - a$.

- Consider the Borel σ-algebra $\mathcal{B}(\mathbb{R})$. We can define the counting measure $M_C(A)$ of set A

$$M_C(A) = \begin{cases} |A \cap \mathbb{N}| & \text{if } A \cap \mathbb{N} \text{ is finite} \\ \infty & \text{o/w} \end{cases}$$

i.e. the no. of elements in A. It is difficult to show that $(\mathbb{R}, \mathcal{B}(\mathbb{R}), M_C)$ is a measure space

Measurable functions

- A function $f: X \to Y$ btwn two measurable spaces (X, Σ_X) and (Y, Σ_Y) is (Σ_X, Σ_Y)-measurable if $\boxed{\forall A \in \Sigma_Y, f^{-1}(A) \in \Sigma_X, \text{ where } f^{-1}(A) = \{x \in X : f(x) \in A\}}$.
  (For general functions f, $f^{-1}(A)$ is not guaranteed to be in Σ_X).
  ⇒ If $X = Y = \mathbb{R}$, and $\Sigma_X = \Sigma_Y = \mathcal{B}(\mathbb{R})$, then such functions f are Borel measurable functions.
- Given a measure space (X, Σ_X, M), a measure space (Y, Σ_Y) and a measurable function $f: X \to Y$. The pushforward measure $f_{\#}M : \Sigma_Y \to [0, \infty]$ is defined as
  $$\boxed{f_{\#}M(A) = M(f^{-1}(A))}$$
  This means a measure space (X, Σ_X, M) and measurable function $f: X \to Y$ together induce a measure on (Y, Σ_Y).

Probability spaces

- We call a measure space (Ω, F, M) a probability space if $M(\Omega) = 1$. We usually use the notation $(\Omega, F, \mathbb{P})$.
- A probability space $(\Omega, F, \mathbb{P})$ represents an experiment.
  ↳ Ω is the sample space. Any element $\omega \in \Omega$ is an outcome of the experiment.
  ↳ F is a family of events (each event describes whether some phenomenon happened or not)
  ↳ $\mathbb{P}$ is the measure assigning probabilities to each event.
  (Intuitively, for a measurable set $A \in F$, $\mathbb{P}(A)$ represents the probability that a randomly selected outcome $\omega \in \Omega$ belongs to A.)
- For a Borel measurable function f, if (X, Σ_X, M) is a probability space, f is called a random variable (RV), the pushforward measure $f_{\#}M$ is the law of random variable f.

Almost everywhere and almost surely.

- It turns out that whatever happens on a set of measure 0 simply does not matter.
- If a property of a measurable function holds everywhere, except on some set of measure, then we say the property holds almost everywhere (a.e.).
- When we work w/ probability spaces $(\Omega, F, \mathbb{P})$, we say the property holds almost surely (a.s.).

Riemann and Lebesgue integration

Convergence of functions

- A sequence of functions $f_1, f_2 \dots$ can converge to a function f in either in a pointwise manner (each pt. could have diff. rates of convergence) or a uniform manner
- For pointwise convergence, we have $\boxed{\lim_{n\to\infty} f_n(x) = f(x) \ \forall x \quad [\lim_{n\to\infty} f_n = f]}$

  Formally, $\forall x, \varepsilon > 0$, $\exists$ a natural no. $N = N(\varepsilon, x)$ s.t. $n \geq N \Rightarrow |f(x) - f_n(x)| < \varepsilon$

  ↳ e.g. Consider $f_n(x) = x^n$, $x \in [0,1]$. $f_n(x)$ converges pointwise to $f(x)$

$$f(x) = \lim_{n\to\infty} f_n(x) = \begin{cases} 0 & x \in [0,1) \\ 1 & x = 1 \end{cases}$$

  Note $f_n(x)$ is continuous, $f(x)$ is not → pointwise convergence doesn't preserve continuity.
- For uniform convergence, we have $\boxed{\sup_x |f(x) - f_n(x)| \to 0 \quad [f_n \rightrightarrows f]}$

  Formally, $\forall \varepsilon > 0$, $\exists$ a natural no. $N = N(\varepsilon)$ s.t. $n \geq N \Rightarrow |f(x) - f_n(x)| < \varepsilon$

  ↳ e.g. Consider $f_n(x) = \frac{x}{n}$, $x \in [0,1]$. $f_n(x)$ converges uniformly to $f(x) = 0$.

Riemann integration

- In Riemann integration, we approx. the integrand w/ step functions comprised of rectangles.
- We say that $s : [a,b] \to \mathbb{R}$ is a step function, if for some partition $a = t_0 < t_1 < \dots < t_N = b$

$$s(x) = \sum_{i=1}^{N} c_i \, \mathbb{I}_{[t_{i-1}, t_i)}(x)$$

  where $c_i \in \mathbb{R}$ are constants.
- We define the integral of a step function as

$$\int_a^b s(x)\,dx = \sum_{i=1}^{N} c_i (t_i - t_{i-1})$$

- Let $f : [a,b] \to \mathbb{R}$ be a general function, define:

$$I_- = \sup_{s \leq f} \int_a^b s(x)\,dx \qquad I_+ = \inf_{s \geq f} \int_a^b s(x)\,dx$$

  If the bound from above / below are equal, i.e. $I = I_- = I_+$, then f is integrable and its integral is defined as $I = \int_a^b f(x)\,dx$
- There are a few fundamental problems w/ the Riemann integral:

  ↳ It is difficult to integrate over infinite-dim. spaces — hard to define an interval

  ↳ The exchange of limit operations req. uniform convergence of f_n (which is stringent)

$$\boxed{\lim_{n\to\infty} \int_a^b f_n(x)\,dx = \int_a^b \lim_{n\to\infty} f_n(x)\,dx}$$

  Proof: $\left| \lim_{n\to\infty} \int_a^b f_n(x)\,dx - \int_a^b f(x)\,dx \right| = \left| \lim_{n\to\infty} \int_a^b f_n(x) - f(x)\,dx \right|$

$$\leq \lim_{n\to\infty} \int_a^b |f_n(x) - f(x)|\,dx \leq \lim_{n\to\infty} (b-a) \sup_x |f_n(x) - f(x)| = 0$$

  * This is not true for pointwise convergence

  uniform convergence means $\lim_{n\to\infty} \sup_x |f_n(x) - f(x)| = 0$.

Lebesgue integration

- In Lebesgue integration, rectangles of the same height can be joined together into one shape. The area covered by the shape is the height times the measure of the set of intervals.
- Let $(\Omega, \mathcal{F})$ be a measurable space. $s: \Omega \to \mathbb{R}$ is a simple function if for some measurable sets $E_1, \dots, E_N \in \mathcal{F}$,

$$s(x) = \sum_{i=1}^{N} c_i \mathbb{I}_{E_i}(x)$$

where $c_i \in \mathbb{R}$ are constants

- If μ is a measure on $(\Omega, \mathcal{F})$, the integral of a simple function wrt μ is defined as

$$\int_\Omega s(x)\mu(dx) = \sum_{i=1}^{N} c_i \mu(E_i)$$

- Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and $f: \Omega \to [0, \infty]$ a non-negative measurable function

The integral of f wrt μ is defined as

$$\int_\Omega f(x)\mu(dx) = \inf_{s \geq f} \int_\Omega s(x)\mu(dx)$$

where the infimum is over simple functions. If the integral is finite, f is integrable.

- Extending to a real measurable function $f: \Omega \to \mathbb{R}$, we define $f(x) = f_+(x) - f_-(x)$, where

$$f_+(x) = \max\{0, f(x)\} \qquad f_-(x) = -\min\{0, f(x)\}.$$

So the integral of f is defined as:

$$\int_\Omega f(x)\mu(dx) = \int_\Omega f_+(x)\mu(dx) - \int_\Omega f_-(x)\mu(dx)$$

and f is integrable if both f_+ and f_- are integrable.

- The Lebesgue integral satisfies the following properties:

↳ $f(x) = g(x)$ a.e. $\Rightarrow \int f d\mu = \int g d\mu$.

↳ Applying a convergence theorem on a seq. of integrable functions $f_n(x)$, we find the pointwise limit $f(x) = \lim_{n\to\infty} f_n(x)$ is also an integrable function w/

$$\int_\Omega f(x)\mu(dx) = \int_\Omega \lim_{n\to\infty} f_n(x)\mu(dx) = \lim_{n\to\infty} \int_\Omega f_n(x)\mu(dx)$$

(Monotone convergence theorem: $f_n(x) \leq f_{n+1}(x)$ a.e. and $\sup_n \int_\Omega f_n(x)\mu(dx) < \infty$.

Dominated convergence theorem: $\exists$ integrable function $g(x)$ s.t. $|f_n(x)| \leq g(x)$ a.e.)

↳ The Lebesgue integral w/ counting measure μ_c is simply a sum

$$\int_\Omega f(x)\mu_c(dx) = \sum_{n \in \mathbb{N}} f(n)$$

↳ When $(\Omega, \mathcal{F}, \mathbb{P})$ is a prob. space, and $X: \Omega \to \mathbb{R}$ is a RV, we write the Lebesgue integral

$$E[X] = \int_\Omega X(\omega)\mathbb{P}(d\omega)$$

Radon-Nikodym derivatives

- Let μ and ν be two measures on a measurable space $(\Omega, \mathcal{F})$. ν is absolutely continuous wrt μ (denoted as $\nu \ll \mu$) if for any measurable set $A \in \mathcal{F}$

$$\boxed{\mu(A) = 0 \Rightarrow \nu(A) = 0}$$

- If $\mu \ll \nu$ and $\nu \ll \mu$, then μ and ν are said to be equivalent.
- Let μ and ν be two finite measures on a common measurable space $(\Omega, \mathcal{F})$. If $\nu \ll \mu$, then there exists a measurable function $f(x)$ s.t. for all measurable sets $A \in \mathcal{F}$,

$$\boxed{\nu(A) = \int_A f(x)\, \mu(dx)}$$

f is unique a.e. and is the Radon-Nikodym derivative of ν wrt μ, denoted as

$$f = \frac{d\nu}{d\mu} \qquad \hookrightarrow \qquad \boxed{\nu(dx) = f(x)\, \mu(dx)}.$$

- Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a prob space, and $X: \Omega \to \mathbb{R}$ a RV w/ the law $\mu_X(A) = \mathbb{P}(X^{-1}(A))$. If $\mu_c \ll \mu_X$ / $\mu_c \ll \mu_X$, then X is a continuous RV, and its Radon-Nikodym derivative $\frac{d\mu_X}{d\mu_c}$ / $\frac{d\nu_X}{d\nu_c}$ is the RV's probability density function / probability mass function.

Product measure spaces.

- The product space $(\Omega, \mathcal{F}, \mu)$ of $(\Omega_1, \mathcal{F}_1, \mu_1)$ and $(\Omega_2, \mathcal{F}_2, \mu_2)$ is defined as follows:
  - ↳ The sample space $\Omega = \Omega_1 \times \Omega_2$ is all pairs of Ω_1 and Ω_2.
  - ↳ The σ-algebra $\mathcal{F} = \mathcal{F}_1 \otimes \mathcal{F}_2$ is the smallest σ-algebra s.t. $\forall A \in \mathcal{F}_1, B \in \mathcal{F}_2$, $A \times B \in \mathcal{F}$.
  - ↳ The measure $\mu = \mu_1 \otimes \mu_2$ is a measure s.t. $\forall A \in \mathcal{F}_1, B \in \mathcal{F}_2$, $\mu(A \times B) = \mu_1(A)\mu_2(B)$
- The Fubini-Tonelli thm. states for measure spaces (X, Σ_X, μ_X) and (Y, Σ_Y, μ_Y), where μ_X and μ_Y are σ-finite measures, and f is a measurable function on the product space, we have

$$\int_X \int_Y |f(x,y)|\, \mu_Y(dy)\, \mu_X(dx) = \int_Y \int_X |f(x,y)|\, \mu_X(dx)\, \mu_Y(dy) = \int_{X \times Y} |f(x,y)|\, \mu_X \otimes \mu_Y(d(x,y))$$

If the integral is finite, f is integrable and we have

$$\int_X \int_Y f(x,y)\, \mu_Y(dy)\, \mu_X(dx) = \int_Y \int_X f(x,y)\, \mu_X(dx)\, \mu_Y(dy) = \int_{X \times Y} f(x,y)\, \mu_X \otimes \mu_Y(d(x,y))$$

Markov Chain Monte Carlo

Markov Chains

Discrete Markov Chains

- Let $\{X_n\}_{n\in\mathbb{N}}$ be a random process taking values in a finite set $S=\{1,\dots,K\}$. $\{X_n\}_{n\in\mathbb{N}}$ is a Markov chain if for any values $x_0,\dots x_n\in S$ we have

$$\mathbb{P}(X_n=x_n \mid X_0=x_0,\dots, X_{n-1}=x_{n-1}) = \mathbb{P}(X_n=x_n \mid X_{n-1}=x_{n-1})$$

i.e. the future is independent of the past, given the present

- If, in addition, the distribution $\mathbb{P}(X_n=x_n \mid X_{n-1}=x_{n-1})$ is independent of n, it is a time-homogeneous Markov chain. We define the transition matrix P as

$$P_{ij} = \mathbb{P}(X_n=j \mid X_{n-1}=i) \qquad \forall i,j\in S$$

i.e. Markov chains generate random sequences by looking only at the previously generated value

- The dist. of the first value is the initial dist. and is often written as a vector μ, where its i-th component μ_i is $\mu_i=\mathbb{P}(X_0=i)$. It then follows that

$$\mathbb{P}(X_n=i) = (\mu P^n)_i$$

- If $\{X_n\}_{n\in\mathbb{N}}$ is a time-homogeneous Markov chain w/ transition matrix P, then π is a stationary dist. of $\{X_n\}_{n\in\mathbb{N}}$ if

$$\pi P = \pi$$

- A Markov chain is irreducible if for any $i,j\in S$, there is an $n\in\mathbb{N}$ s.t.

$$\mathbb{P}(X_n=j \mid X_0=i) > 0$$

i.e. any state is reachable from any state.

- A Markov chain is aperiodic if for any $i\in S$, there is an $n\in\mathbb{N}$ s.t.

$$\mathbb{P}(X_n=i \mid X_0=i) > 0 \qquad \text{and} \qquad \mathbb{P}(X_{n+1}=i \mid X_0=i) > 0$$

- If a Markov chain is irreducible and aperiodic, then it is regular ergodic. A regular ergodic Markov chain has a limiting dist. (unique stationary dist.) π which puts the mass on every element of the state space S.

- For a regular ergodic Markov chain $\{X_n\}_{n\in\mathbb{N}}$ w/ limiting dist. π, let $V_i(n)=\sum_{j=0}^{n-1}\mathbb{I}(X_j=i)$ be the no. of visits to state i up until time n $\forall i\in S$. The ergodic thm. states that

$$\frac{V_i(n)}{n} \to \pi_i \quad \text{a.s.}$$

i.e. the prop. of time spent in state i converges to its stationary prob. π_i as $n\to\infty$.

- We can thus evaluate the expected value $E_{X\sim\pi}[f(x)]$ for a function $f: S\to\mathbb{R}$ on the state space S

$$E\left[\frac{1}{n}\sum_{j=0}^{n-1} f(X_j)\right] = \sum_{i\in S} f(i)\, E\left[\frac{V_i(n)}{n}\right] \to \sum_{i\in S} f(i)\,\pi_i = E_{X\sim\pi}[f(x)]$$

General state space Markov chains

- A Markov transition kernel on a measurable space (X, Σ_X) is a function $P: X \times \Sigma_X \to [0,1]$ s.t.
  - ↳ For each fixed $x \in X$, $P(x, \cdot)$ is a prob. measure on (X, Σ_X)
  - ↳ For each fixed $A \in \Sigma_X$, $P(\cdot, A)$ is a non-negative measurable function on X.
- For any measurable sets $A_i \in \Sigma_X$, fix a starting pt. $x \in X$. We have:

$$\mathbb{P}(X_1 \in A_1 \mid X_0 = x) = P(x, A_1)$$

$$\mathbb{P}(X_1 \in A_1, X_2 \in A_2 \mid X_0 = x) = \int_{x \in A_1} P(x, dx_1) P(x_1, A_2)$$

$$\mathbb{P}(X_1 \in A_1, X_2 \in A_2, \ldots, X_n \in A_n \mid X_0 = x) = \int_{x_1 \in A_1} P(x, dx_1) \int_{x_2 \in A_2} P(x_1, dx_2) \cdots \int_{x_n \in A_n} P(x_{n-2}, dx_{n-1}) P(x_{n-1}, A_n)$$

- Given an initial dist. μ for X_0, then we have

$$\mathbb{P}(X_0 \in A_0, X_1 \in A_1, \ldots X_n \in A_n) = \int_{x_0 \in A_0} \mu(dx_0) \int_{x_1 \in A_1} P(x_0, dx_1) \cdots \int_{x_{n-1} \in A_{n-1}} P(x_{n-2}, dx_{n-1}) P(x_{n-1}, A_n)$$

* The integrals are well-defined by the measurability of the function $P(\cdot, A): X \to [0,1]$

- A random process $\{X_n\}_{n \in \mathbb{N}}$ is a time-homogeneous Markov chain w/ transition kernel P if for any values $x_0, \ldots, x_n \in X$ and $A \in \Sigma_X$ we have

$$\mathbb{P}(X_n \in A \mid X_0 = x_0, \ldots, X_{n-1} = x_{n-1}) = \mathbb{P}(X_n \in A \mid X_{n-1} = x_{n-1}) = P(x_{n-1}, A)$$

- There are a few operations involving transition kernels. Let $f: X \to \mathbb{R}$ be a measurable function, $\mu: \Sigma_X \to \mathbb{R}^+$ a measure and P, Q transition kernels.
  - ① PQ is a transition kernel, corr. to first transitioning w/ P, and then Q.

$$PQ(x, A) = \int P(x, dy) Q(y, A)$$

  * We denote n transitions of P as $P^n = P \cdots P$.
  - ② $Pf: X \to \mathbb{R}$ is a function corr. to the expected value of f after a transition w/ P starting at $x \in X$.

$$Pf(x) = \int P(x, dy) f(y)$$

  - ③ μP is a measure after one transition, defined as

$$\mu P(A) = \int P(x, A) \mu(dx)$$

  * μ is stationary for P if $\mu P = \mu$

- When the transition kernel P has a Radon-Nikodym derivative (density) w/ r.t. the Lebesgue measure, we denote it as $p(x,y)$, where we have

$$P(x, dy) = p(x, y)\, dy \quad \leftrightarrow \quad P(x, A) = \int_A p(x, y)\, dy$$

* Discrete Markov chains are just a special case of general state space Markov chains. The transition matrix P corr. to the Radon-Nikodym derivative of P w.r.t. μ_c.

Markov chain Monte Carlo

Detailed balance

- A Markov chain w/ transition kernel P is reversible/satisfies detailed balance wrt the measure π^* if for all $A, B \in \Sigma_X$.

$$\int_A P(x, B)\, \pi^*(dx) = \int_B P(y, A)\, \pi^*(dy) \quad \leftrightarrow \quad p(x,y)\,\pi(x) = p(y,x)\,\pi(y)$$

where $\pi(x)$ is the R-N derivative (density) of measure π^*, so $\pi^*(dx) = \pi(x)dx$.

$$\left(\int_A \left(\int_B p(x,y)dy\right)\pi(x)dx = \int_B \left(\int_A p(y,x)dx\right)\pi(y)dy \rightarrow p(x,y)\,\pi(x) = p(y,x)\,\pi(y)\right).$$

i.e. the prob. of observing transition $A \to B$ is equal to that of observing transition $B \to A$.

- If the transition kernel P satisfies the detailed balance eqns for the density π, then

$$\pi P = \pi$$

i.e. π is the invariant density of the Markov chain w/ transition kernel P.

Metropolis–Hastings (MH) algorithm

- Suppose there is a target prob. measure π^* from which we want to draw samples. We consider simulating a random process from a Markov chain w/ an invariant density π.
- Consider a transition operator $P(x, dy)$ of the form

$$P(x, dy) = p(x,y)\,dy + r(x)\,\delta_x(dy)$$

where δ_x is the dirac measure (1 if the argument is in x, o/w 0)

- If we have $p(x,x) = 0$ and $\int_{\mathbb{R}^n} P(x, dy) = 1$, then we have

$$1 = \int_{\mathbb{R}^n} P(x, dy) = \int_{\mathbb{R}^n} p(x,y)\,dy + \int_{\mathbb{R}^n} r(x)\,\delta_x(dy) \rightarrow r(x) = 1 - \int_{\mathbb{R}^n} p(x,y)\,dy$$

which is the prob. that the chain remains at x after a transition.

- If $p(x,y)$ satisfies reversibility, then π is the invariant density of the transition kernel P

$$\begin{aligned}
\int_{\mathbb{R}^n} P(x, A)\,\pi^*(dx) &= \int_{\mathbb{R}^n} \left(\int_A P(x, dy)\right)\pi(x)dx \\
&= \int_{\mathbb{R}^n} \left(\int_A p(x,y)\,dy + \int_A r(x)\,\delta_x(dy)\right)\pi(x)dx \\
&= \int_A \int_{\mathbb{R}^n} p(x,y)\,\pi(x)dx\,dy + \int_{\mathbb{R}^n} r(x)\,\delta_x(A)\,\pi(x)dx \\
&\overset{\text{detailed balance}}{=} \int_A \int_{\mathbb{R}^n} p(y,x)\,dx\,\pi(y)dy + \int_A r(x)\pi(x)dx \\
&= \int_A (1 - r(y))\,\pi(y)dy + \int_A r(x)\pi(x)dx = \int_A \pi(y)dy = \pi^*(A).
\end{aligned}$$

- Consider a proposal density $q(x,y)$ s.t. $\int q(x,y)dy = 1$.

If $x \to y$ transitions are more freq. than $y \to x$ transitions, we have $q(x,y)\pi(x) > q(y,x)\pi(y)$

We rebalance both sides by introducing $\alpha(x,y) < 1$, $\alpha(x,y)\,q(x,y)\,\pi(x) = q(y,x)\,\pi(y)$.

$$\alpha(x,y) = \frac{q(y,x)\,\pi(y)}{q(x,y)\,\pi(x)}.$$

On the contrary, if we have $q(x,y)\,\pi(x) < q(y,x)\,\pi(y)$, set $\alpha(x,y) = 1$, so for $x \neq y$,

$$p(x,y) = \alpha(x,y)\,q(x,y), \qquad \text{where } \alpha(x,y) = \min\left\{\frac{\pi(y)\,q(y,x)}{\pi(x)\,q(x,y)}, 1\right\}.$$

- The overall transition operator P is thus given by

$$P(x, dy) = \alpha(x,y)\, q(x,y)\, dy + \left[1 - \int \alpha(x,y)\, q(x,y)\, dy\right] \delta_x(dy) \quad \text{or}$$

$$P(x, A) = \int_A \alpha(x,y)\, q(x,y)\, dy + \left[1 - \int \alpha(x,y)\, q(x,y)\, dy\right] \delta_x(A)$$

which is reversible by construction, therefore has Π as its invariant density.

* Note that the normalising const. of the target density $\Pi(x) = \frac{\phi(x)}{\int \phi(x)\, dx}$ is not req.

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

If the proposal density is symmetric, i.e. $q(x,y) = q(y,x)$,

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

- Most commonly, $q(x,y)$ is chosen as a Normal dist, i.e. $q(x,y) \propto \exp\left(-\frac{\|x-y\|}{2\sigma^2}\right)$.

However, it is not trivial to optimally choose σ^2.

↳ σ^2 too small: Markov chain move very slowly across state space

↳ σ^2 too big: A proposal is rarely accepted.

* Both cases yield high correlation samples → bad mixing.

- The MH algorithm has a few drawbacks:

↳ It is ineffective for sampling multimodal dist. — especially if the target dist. Π has two highly conc. and distant modes (chain stuck in one of them).

↳ Reversible Markov chains are likely to make steps back, and therefore explore the state space in a slow, diffusive fashion.

Pseudo-code for the MH algorithm.

- Given the target density Π, a proposal density q, the first state X_0, and the no. of iterations N,

```
for i = 1, ..., N do                    N iterations
    Y ~ Q(X_{i-1}, ·)                   Make a proposal using Q(X_{i-1}, ·)
    u ~ U[0,1]
    if u < α(X_{i-1}, Y) then
        X_i = Y                         Accept w/ prob. α(X_{i-1}, Y), set X_i = Y
    else
        X_i = X_{i-1}                   Rej. w/ prob. 1-α(X_{i-1}, Y), set X_i = X_{i-1}
    end if
end for
return X_0, ..., X_n
```

Mixture transition kernels

- The mixture transition kernel is composed of the mixture

$$\gamma P_1(x, dy) + (1-\gamma) P_2(x, dy) \quad , \quad \gamma \in [0,1]$$

where transition kernels $P_1(x, dy)$ and $P_2(x, dy)$ both have invariant density $\pi(x)$.

- The mixture transition kernel has invariant density $\pi(x)$.

$$\int_{\mathbb{R}^d} [\gamma P_1(x, dy) + (1-\gamma) P_2(x, dy)] \pi(dx) = \pi(dy)$$

(LHS $= \gamma \int_{\mathbb{R}^d} P_1(x, dy)\, \pi(x) dx + (1-\gamma) \int_{\mathbb{R}^d} P_2(x, dy)\, \pi(x) dx = \gamma \pi(dy) + (1-\gamma)\pi(dy) = \pi(dy)$) .

- The mixture transition kernel is useful for drawing samples from multimodal target densities.

Product transition kernels.

- Consider breaking up a vector up into sub-blocks, $x = (x_1, x_2)$ where $x \in \mathbb{R}^{d_1+d_2}$, $x_1 \in \mathbb{R}^{d_1}$, $x_2 \in \mathbb{R}^{d_2}$
- The product transition kernel is given by

$$P_1(x_1, dy_1 | x_2) P_2(x_2, dy_2 | y_1)$$

where conditional transition kernels $P_1(x_1, dy_1 | x_2)$ and $P_2(x_2, dy_2 | y_1)$ have invariant densities

$\pi_{1|2}(\cdot | x_2)$ for fixed x_2, $\pi_{2|1}(\cdot | y_1)$ for fixed y_1 respectively

(i.e. $\pi^*_{1|2}(dy_1 | x_2) = \int P_1(x_1, dy_1 | x_2)\, \pi_{1|2}(x_1 | x_2) dx_1$; $\pi^*_{2|1}(dy_2 | y_1) = \int P_2(x_2, dy_2 | y_1)\, \pi_{2|1}(x_2 | y_1) dx_2$

- The product transition kernel has invariant density $\pi(x_1, x_2) = \pi_{1|2}(x_1 | x_2)\, \pi_2(x_2) = \pi_{2|1}(x_2 | x_1)\, \pi_1(x_1)$.

$$\int_{\mathbb{R}^{d_2}} \int_{\mathbb{R}^{d_1}} P_1(x_1, dy_1 | x_2) P_2(x_2, dy_2 | y_1) \pi^*(dx_1, dx_2) = \pi^*(dy_1, dy_2)$$

(LHS $= \int_{\mathbb{R}^{d_2}} \int_{\mathbb{R}^{d_1}} P_1(x_1, dy_1 | x_2) P_2(x_2, dy_2 | y_1)\, \pi(x_1, x_2) dx_1 dx_2$

$= \int_{\mathbb{R}^{d_2}} P_2(x_2, dy_2 | y_1) \left[\int_{\mathbb{R}^{d_1}} P_1(x_1, dy_1 | x_2)\, \pi_{1|2}(x_1 | x_2) dx_1 \right] \pi_2(x_2) dx_2$

$= \int_{\mathbb{R}^{d_2}} P_2(x_2, dy_2 | y_1)\, \pi^*_{1|2}(dy_1 | x_2)\, \pi_2(x_2) dx_2 = \int_{\mathbb{R}^{d_2}} P_2(x_2, dy_2 | y_1) \frac{\pi_{2|1}(x_2 | x_1)\, \pi^*(dy_1)}{\pi_2(x_2)} \pi_2(x_2) dx_2$

$= \pi^*_1(dy_1) \int_{\mathbb{R}^{d_2}} P_2(x_2, dy_2 | y_1)\, \pi_{2|1}(x_2 | y_1) dx_1 = \pi^*_1(dy_1)\, \pi^*_{2|1}(dy_2 | y_1) = \pi^*(dy_1, dy_2)$

Gibbs sampling (GS) algorithm

- In GS, for each component j of x, x_j, in turn, we sample a new value from the conditional distribution of x_j given all other variables x_{-j}.

$$\hat{x}_j \sim p(x_j | x_1 \dots x_{j-1}, x_{j+1} \dots x_D)$$

This is equivalent to using the product transition kernel $\prod_{j=1}^{D} P_j(x_j, dy_j | x_{-j})$ and using MH w/ exact conditionals as proposals.

- For $D=2$, the transition $(x_1, x_2) \to (y_1, y_2)$ consists of $(x_1, x_2) \xrightarrow{[1]} (y_1, x_2) \xrightarrow{[2]} (y_1, y_2)$

↳ step [1]: $q([x_1, x_2], [y_1, x_2]) = \pi(y_1 | x_2)$; $q([y_1, x_2], [x_1, x_2]) = \pi(x_1 | x_2)$

$$\therefore \frac{\pi(y_1, x_2)\, q([y_1, x_2], [x_1, x_2])}{\pi(x_1, x_2)\, q([x_1, x_2], [y_1, x_2])} = \frac{\pi(y_1 | x_2)\pi(x_2)\, \pi(x_1 | x_2)}{\pi(x_1 | x_2)\pi(x_2)\pi(y_1 | x_2)} = 1 \to \alpha([x_1, x_2], [y_1, x_2]) = 1$$

↳ step [2]: $q([y_1, x_2], [y_1, y_2]) = \pi(y_2 | y_1)$; $q([y_1, y_2], [y_1, x_2]) = \pi(x_2 | y_1)$

$$\therefore \frac{\pi(y_1, y_2)\, q([y_1, y_2], [y_1, x_2])}{\pi(y_1, x_2)\, q([y_1, x_2], [y_1, y_2])} = \frac{\pi(y_1 | y_2)\pi(y_2)\, \pi(x_2 | y_1)}{\pi(x_2 | y_1)\pi(y_1)\, \pi(y_2 | y_1)} = 1 \to \alpha([y_1, x_2], [y_1, y_2]) = 1$$

Pseudo-code for the GS algorithm.

- Given the target density π, the first state X_0 and no. of iterations N,

for $i = 1, \ldots, N$ do

for $j = 1, \ldots, D$ do

$X_i^j \sim \pi(\cdot \mid X_i^1 \ldots X_i^{j-1}, X_i^{j+1} \ldots X_i^D)$

end for

end for

return $X_0, \ldots X_N$

Data augmentation

- Consider a target density $\pi(\theta)$ where $\theta \in \mathbb{R}^p$. Now augment the model w/ $\phi \in \mathbb{R}^p$, so the augmented target density is $\pi(\theta, \phi)$.
- The desired density can be recovered as

$$\pi(\theta) = \int \pi(\theta, \phi) d\phi$$

→ We consider a Markov chain w/ invariant density $\pi(\theta, \phi)$ and draw samples $\theta^{(n)}, \phi^{(n)} \sim \pi(\theta, \phi)$

For each $\theta^{(n)}$, it follows that $\pi(\theta^{(n)}) = \int \pi(\theta^{(n)}, \phi) d\phi$ – each $\theta^{(n)}$ marginally dist. w/ $\pi(\cdot)$.

- We select ϕ s.t. the exact conditionals $\pi(\phi \mid \theta)$ and $\pi(\theta \mid \phi)$ can be sampled directly.

Hilbert spaces and Sobolev spaces

Vector, metric, Banach and Hilbert spaces

Vector space

- A vector space V is a collection of vectors $\underline{x}, \underline{y}, \underline{z}$ that satisfy the following axioms.
  - ↳ Addition commutative: $\underline{x}+\underline{y} = \underline{y}+\underline{x} \in V$
  - ↳ Addition associative: $\underline{x}+(\underline{y}+\underline{z}) = (\underline{x}+\underline{y})+\underline{z} \in V$
  - ↳ Zero vector exists: $\underline{x}+\underline{0} = \underline{x}$
  - ↳ Additive inverse exists: $\underline{x}+(-\underline{x}) = \underline{0}$
  - ↳ Scalar multiplication closed: $\alpha\underline{x} \in V$, $\alpha \in \mathbb{F}$
  - ↳ Scalar multiplication distributive: $\alpha(\underline{x}+\underline{y}) = \alpha\underline{x}+\alpha\underline{y} \in V$, $\alpha \in \mathbb{F}$
  - ↳ Scalar multiplication associative: $\alpha(\beta\underline{x}) = \beta(\alpha\underline{x})$, $\alpha, \beta \in \mathbb{F}$.
  - ↳ Zero and unit scalar: $0\underline{x} = \underline{0} \in V$, $1\underline{x} = \underline{x}$.
- Vector space is a Group under addition.

Inner product space

- An inner product space is a vector space V over the field $\mathbb{F}$ together w/ an inner product, which is the mapping $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{F}$ that satisfies $\forall \underline{x}, \underline{y}, \underline{z} \in V$, $\forall \alpha, \beta \in \mathbb{F}$.
  - ↳ Conjugate symmetry: $\langle \underline{x}, \underline{y} \rangle = \overline{\langle \underline{y}, \underline{x} \rangle} \to \langle \underline{x}, \underline{x} \rangle \in \mathbb{R}$
  - ↳ Linearity: $\langle \alpha x + \beta y, z \rangle = \alpha \langle \underline{x}, \underline{z} \rangle + \beta \langle \underline{y}, \underline{z} \rangle$
  - ↳ Positive-definiteness: $\langle \underline{x}, \underline{x} \rangle > 0$ if $\underline{x} \neq 0$ ($\langle \underline{x}, \underline{x} \rangle = 0$ iff $\underline{x} = \underline{0}$)
- For vectors $\underline{a}, \underline{b} \in \mathbb{C}^n$, the inner product is defined as

$$\langle \underline{a}, \underline{b} \rangle = \sum_{i=1}^{n} a_i^* b_i$$

For functions f, g of variable on interval $x \in [a,b] \subset \mathbb{R}$, the inner product is defined as

$$\langle f, g \rangle = \int_a^b f(x) g(x)\, dx$$

Metric space

- A metric space is a vector space V that has a metric function, which is a mapping $\rho : V \times V \to \mathbb{R}$ that satisfies $\forall \underline{x}, \underline{y}, \underline{z} \in V$
  - ↳ Zero self-distance: $\rho(\underline{x}, \underline{y}) = 0$ iff $\underline{x} = \underline{y}$
  - ↳ Positivity: $\rho(\underline{x}, \underline{y}) > 0$ if $\underline{x} \neq \underline{y}$
  - ↳ Symmetry: $\rho(\underline{x}, \underline{y}) = \rho(\underline{y}, \underline{x})$
  - ↳ Triangle inequality: $\rho(\underline{x}, \underline{z}) \leq \rho(\underline{x}, \underline{y}) + \rho(\underline{y}, \underline{z})$

Normed spaces

- A normed space is a vector space V over the field $\mathbb{F}$ together w/ a norm, which is a mapping $\|\cdot\| : V \to \mathbb{R}$ that satisfies $\forall \underline{u}, \underline{v} \in V$, $\forall \lambda \in \mathbb{F}$
  - ↳ Non-negativity: $\|\underline{v}\| \geq 0$
  - ↳ Positive-definiteness: $\|\underline{v}\| = 0$ iff $\underline{v} = \underline{0}$
  - ↳ Absolute homogeneity: $\|\lambda \underline{v}\| = |\lambda| \|\underline{v}\|$
  - ↳ Triangle inequality: $\|\underline{u}+\underline{v}\| \leq \|\underline{u}\| + \|\underline{v}\|$
- For inner product spaces V, the norm can be induced by the inner product. $\forall \underline{v} \in V$,

$$\|\underline{v}\| = \sqrt{\langle \underline{v}, \underline{v} \rangle}$$

For any two elements $\underline{x}, \underline{y} \in V$, we have the following (in)equalities:

- ↳ Parallelogram law: $\|\underline{x}+\underline{y}\|^2 + \|\underline{x}-\underline{y}\|^2 = 2(\|\underline{x}\|^2 + \|\underline{y}\|^2)$

$(\langle \underline{x}+\underline{y}, \underline{x}+\underline{y} \rangle + \langle \underline{x}-\underline{y}, \underline{x}-\underline{y} \rangle = \langle \underline{x}, \underline{x} \rangle + 2\langle \underline{x}, \underline{y} \rangle + \langle \underline{y}, \underline{y} \rangle + \langle \underline{x}, \underline{x} \rangle - 2\langle \underline{x}, \underline{y} \rangle + \langle \underline{y}, \underline{y} \rangle)$

- ↳ Cauchy-Schwarz inequality: $|\langle \underline{x}, \underline{y} \rangle| \leq \|\underline{x}\| \|\underline{y}\|$

(Let $\underline{z} = \underline{x} - \lambda \underline{y}$ for some scalar λ.

$\|\underline{z}\|^2 = \|\underline{x} - \lambda \underline{y}\|^2 = \|\underline{x}\|^2 - 2\lambda \langle \underline{x}, \underline{y} \rangle + \lambda^2 \|\underline{y}\|^2 \geq 0$

→ we have $A\lambda^2 + B\lambda + C \geq 0$. where $A = \|\underline{y}\|^2$, $B = -2\langle \underline{x}, \underline{y} \rangle$, $C = \|\underline{x}\|^2$

since $A\lambda^2 + B\lambda + C$ is non-negative, we have $\Delta = B^2 - 4AC \leq 0$,

$(-2\langle x, y \rangle)^2 - 4\|\underline{x}\|^2 \|\underline{y}\|^2 \geq 0 \quad \rightarrow \quad |\langle x, y \rangle| \geq \|\underline{x}\|^2 \|\underline{y}\|^2$)

* l_p, L_p norms are defined w/o reference to an inner product → norm is more general.
* A normed space is also a metric space. w/ $\rho(\underline{x}, \underline{y}) = \|\underline{x} - \underline{y}\|$

orthogonality and orthonormality

- For a set of vectors $\{v_1, v_2 \ldots v_n\}$ in an inner product space to be orthogonal, each pair of distinct vectors are orthogonal,

$$\langle v_i, v_j \rangle = 0 \qquad \forall i \neq j$$

If each vector is a unit vector, the set of vectors is orthonormal.

$$\|v_i\| = 1 \quad \forall i$$

- A set of vectors $v_1, v_2 \ldots, v_n$ is linearly independent iff

$$\sum_{i=1}^{n} c_i v_i = 0 \quad \Rightarrow \quad c_i = 0 \ \forall i$$

Any set of orthonormal vectors $e_1, e_2 \ldots e_n$ is linearly independent, thus can serve as a basis for a finite dim. inner product space V. Every vector $v \in V$ can be written

$$v = \sum_{i=1}^{n} \alpha_i e_i$$

where $\alpha_1, \alpha_2 \ldots \alpha_n$ are the coordinates of v.

Completeness

- A sequence of vectors $\{x_1, x_2, x_3, \ldots\}$ is a Cauchy sequence of vectors if for any positive $\varepsilon > 0$, there exists a no. N s.t.

$$\|x_m - x_n\| < \varepsilon \quad \forall\, m, n > N$$

i.e. the seq. of terms x_m, x_n get closer and closer in an unlimited manner as $m, n \to \infty$.

- An infinite sequence of vectors $\{x_1, x_2 \ldots\}$ is said to be convergent if there exists an element $x \in V$ s.t.

$$\|x_n - x\| \to 0$$

- If every Cauchy sequence in a vector space V converges to a limit that is also in V, then the space V is complete.

* All finite dim. normed spaces are complete (not necessarily for infinite dim.).

Banach spaces

- A Banach space is a normed space that is complete
- All finite dim. normed spaces are complete → Banach spaces.

Not all infinite dim normed spaces are complete (some are, eg l^p, L^p spaces).

- The l^p space consists of all elements $x = (x_1, x_2 \ldots)$ s.t. its l-p norm is finite,

$$\|x\|_{l_p} = \left(\sum_{i=1}^{\infty} |x_i|^p\right)^{1/p} < \infty \quad \text{for } 1 \le p < \infty$$

For $p = \infty$, $\|x\|_{l_\infty} = \sup_i |x_i| < \infty$

- The L^p space consists of measurable functions f defined on the measure space (X, Σ_x, M) s.t. its L-p norm is finite,

$$\|f\|_{L_p} = \left(\int_X |f(x)|^p \, dM\right)^{1/p} < \infty \quad \text{for } 1 \le p < \infty$$

For $p = \infty$, $\|f\|_{L_\infty} = \operatorname{ess\,sup}_{x \in X} |f(x)| < \infty$

- The l^p, L^p spaces are normed and complete → they are Banach spaces
* Only for $p = 2$, l^p, L^p spaces have an inner product (norm induced by inner product)

Pre-Hilbert space

- A pre-Hilbert space is an inner product space that is incomplete.
- eg: Consider the seq. of functions $\{f_n\}$ defined on the unit line.

$$f_n(x) = \begin{cases} 1 & x \in [0, 1/2] \\ 1 - 2n(x - 1/2) & x \in [1/2, \frac{1}{2}(1 + \frac{1}{n})] \\ 0 & x \in [\frac{1}{2}(1 + \frac{1}{n}), 1] \end{cases}$$

Note that $\|f_n - f_m\|_{L_2} = \sqrt{\int_0^1 (f_n(x) - f_m(x))^2 dx} = (1 + \frac{n}{m})\sqrt{\frac{1}{6n}}$, $\lim_{m,n \to \infty} \|f_n - f_m\|_{L_2} = 0$

so $\{f_n\}$ is a Cauchy seq. that converges to $f(x) = \begin{cases} 1 & x \in [0, 1/2] \\ 0 & x \in [1/2, 1] \end{cases}$

$\{f_n\}$ is continuous, f is not continuous → space is incomplete → pre-Hilbert space.

Hilbert space

- A Hilbert space is a Banach space endowed w/ an inner product.
- Let $\{e_i\}$ be an infinite set of orthonormal vectors in Hilbert space $\mathcal{H}$.

  For any function f (a pt in infinite dim. space $\mathcal{H}$) denote the coefficients $c_i = \langle e_i, f\rangle$.

  ↳ Bessel's inequality: $\boxed{\sum_{i=1}^{\infty} c_i^2 \le \|f\|^2}$

  (Consider finite sum from first n terms in the infinite seq. $f_n = \sum_{i=1}^{n} c_i e_i$.

  $\|f_n\|^2 = \langle f_n, f_n\rangle = \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \langle e_i, e_j\rangle = \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \delta_{ij} = \sum_{i=1}^{n} c_i^2$

  $\langle f, f_n\rangle = \sum_{i=1}^{\infty}\sum_{j=1}^{n} c_i c_j \langle e_i, e_j\rangle = \sum_{i=1}^{\infty}\sum_{j=1}^{n} c_i c_j \delta_{ij} = \sum_{i=1}^{n} c_i^2$

  Cauchy-Schwarz: $|\langle f, f_n\rangle|^2 \le \|f\|^2 \|f_n\|^2 \rightarrow \sum_{i=1}^{n} c_i^2 \le \|f\|^2$. Holds for $n=\infty$)

  * In ℓ^2/L^2 space, $\|f\| < \infty$. by Bessel's inequality. $\sum_{i=1}^{\infty} c_i^2 < \infty$.

  ↳ Parseval's identity: $\boxed{\{e_i\} \text{ is complete} \iff \sum_{i=1}^{\infty} c_i^2 = \|f\|^2}$

  * We can use Parseval's identity to check that an infinite basis $\{e_i\}$ is complete.

- ℓ^p, L^p spaces have an inner product for $p=2$, $\rightarrow \ell^2, L^2$ spaces are Hilbert spaces.
- A sequence of functions in L^2 space converge if for any small $\varepsilon > 0$, there exists a no. N s.t. $\|f_n - f\|_{L_2} < \varepsilon \ \forall n > N$. As $\|f\|_{L_2} = \sqrt{\int_a^b |f(x)|^2 dx}$, we have

$$\|f_n - f\|_{L_2} = \sqrt{\int_a^b |f_n(x) - f(x)|^2 dx} \rightarrow 0$$

i.e. convergence in the L_2 norm means convergence a.e. $\rightarrow f, f_n$ are equivalence classes.

* We only req. equivalence a.e. as sets w/ zero measure don't contribute to the integral

$\rightarrow L^2$ space is a space of equivalence functions. i.e. each equivalence class is grouped as the same pt. in the L^2 space.

- Consider a set of orthonormal functions $\{\phi_k\}$ that are square integrable ($\int |\phi_k(x)|^2 dx < \infty$)

  Any function $f \in L^2$ can be written as a Fourier series wrt $\{\phi_k\}$.

$$f = \sum_{k=1}^{\infty} c_k \phi_k$$

where $c_k = \langle f, \phi_k\rangle$ are the Fourier coefficients of f.

By Bessel's inequality, $\sum_{k=1}^{\infty} c_k^2 \le \|f\|_{L_2}^2$, and since $f \in L^2$, $\|f\|_{L_2} < \infty$, therefore,

$$\sum_{k=1}^{\infty} c_k^2 < \infty.$$

∴ The seq. of Fourier coeff $\{c_k\}$ is an element of the ℓ^2 space.

$\rightarrow$ The elements $f \in L^2$ and $c = (c_1, c_2 \ldots) \in \ell^2$ are connected via the Fourier coeff $c_k = \langle f, \phi_k\rangle$.

The ℓ^2, L^2 spaces are isomorphic, and ℓ^2 is a coordinate system for L^2.

Reproducing Kernel Hilbert Space (RKHS)

- A RKHS is a Hilbert space of functions in which point evaluation is a continuous linear functional – if functions are close in norm, they are also close pointwise, i.e.

$$\|f-g\|_{L_2} \to 0 \quad \Rightarrow \quad |f(x)-g(x)| \to 0, \ \forall x.$$

- A RKHS defines a reproducing kernel K_x, which is symmetric and positive definite, s.t.

$$f(x) = \langle f, K_x \rangle$$

Conversely, Moore-Aronszajn thm. states that a reproducing kernel K_x uniquely defines a RKHS

- consider a dataset $(x_1,y_1)\dots(x_N,y_N) \in X \times \mathbb{R}$ representing $y_i = f(x_i)$.

By calculus of variations, the regularised empirical error $\frac{1}{N}\sum_{n=1}^{N} |f(x_n)-y_n|^2 + \lambda \|f\|^2$ is minimised by functions of the form

$$\hat{f}(\cdot) = \sum_{n=1}^{N} \alpha_n K(\cdot, x_n)$$

where $\alpha = (\alpha_1, \dots \alpha_N)$ are given by the soln of the linear system

$$(K + (N\lambda) I_N)\,\alpha = y.$$

where λ is a +ve scalar, K is a $N \times N$ matrix w/ elements $K(x_i, x_j)$,

and y is a $N \times 1$ vector w/ elements y_i

→ This enables function approximation via the form $\hat{f}(\cdot) = \sum_{n=1}^{N} \alpha_n K(\cdot, x_n)$

↳ Proof: $\|f\|^2 = \langle f, f \rangle = \sum_i \sum_j \alpha_i \alpha_j \langle K(\cdot, x_i), K(\cdot, x_j) \rangle = \sum_i \sum_j \alpha_i \alpha_j K(x_i, x_j) = \alpha^T K \alpha$

$$\therefore \quad \frac{\partial}{\partial \alpha} \frac{1}{N} (K\alpha - y)^T (K\alpha - y) + \lambda \alpha^T K \alpha = 0 \to \frac{1}{N} K(K\alpha - y) + \lambda K \alpha = 0$$

$$\Rightarrow \alpha^* = \underset{\alpha \in \mathbb{R}^N}{\operatorname{argmin}} \frac{1}{N} (K\alpha - y)^T (K\alpha - y) + \lambda \|f\|^2 = (K + (N\lambda) I_N)^{-1} y$$

Sobolev spaces and function approximation

Weak derivative

- Consider the function $f: [0,1] \to \mathbb{R}$. The function $h: [0,1] \to \mathbb{R}$ is a weak derivative of f if for every differentiable function $g: [0,1] \to \mathbb{R}$ w/ $g(0) = g(1) = 0$ it holds that

$$\int_0^1 f(x) g'(x)\,dx = -\int_0^1 h(x) g(x)\,dx$$

- The notion of weak derivatives can be generalised to multiple dim over more general domains, and higher derivatives follow straightforwardly (not important – just know they exist + are well defined)

- e.g.: Consider the function $f: [0,2] \to [0,1]$, $f(x) = \begin{cases} x & x \in [0,1) \\ 1 & x \in [1,2] \end{cases}$

For any differentiable test function $\phi: [0,2] \to \mathbb{R}$, w/ $\phi(0) = \phi(2) = 0$, then

$$-\int_0^1 x \phi'(x)\,dx = -[x\phi(x)]_0^1 + \int_0^1 \phi(x)\,dx = -\phi(1) + \int_0^1 \phi(x)\,dx \; ; \; -\int_1^2 \phi(x)\,dx = -[\phi(x)]_1^2 = \phi(1)$$

$$\therefore \; -\int_0^2 f(x)\phi(x)\,dx = \int_0^1 \phi(x)\,dx = \int_0^2 \mathbb{1}_{x \in [0,1)}\,dx \to Df(x) = \mathbb{1}_{x \in [0,1]}$$

Sobolev spaces

- Denote a space of smooth continuous functions of degree k (has k derivatives), defined on a domain which is an open set on the real line $\Omega \subset \mathbb{R}$ as $C^k(\Omega)$.
The space $C^k(\Omega)$ is incomplete → enlarge to a Lebesgue space. L^p.

- Consider a function $f \in L^p(\Omega)$ defined on $\Omega \subset \mathbb{R}$. Let f have weak derivatives up to kth degree which also belong to $L^p(\Omega)$. A Sobolev space $W_p^k(\Omega)$ is defined by the set of functions

→ $$W_p^k(\Omega) = \{ f \in L^p(\Omega) : D^s f \in L^p(\Omega),\ 0 \le s \le k \}$$

w/ the norm defined as follows:

$$\|f\|_{W_p^k}^p = \int_\Omega |f(x)|^p dx + \int_\Omega |Df(x)|^p dx + \dots + \int_\Omega |D^k f(x)|^p dx. \rightarrow \|f\|_{W_p^k} = \left(\sum_{s=0}^{k} \int_\Omega |D^s f(x)|^p dx \right)^{1/p}$$

- For the case $p=2$, $W_2^k(\Omega)$ is a Hilbert space w/ the inner product

$$\langle f, g \rangle_{W_2^k(\Omega)} = \sum_{s=0}^{k} \langle D^s f, D^s g \rangle$$

* Both the norm and inner product encode aspects of the smoothness of the function classes.

clarify for how case, use which def'n(?)

Function approximation by polynomial expansion

- Consider functions from the set $C_2[-\pi, \pi] = C[-\pi, \pi] \cap L^2[-\pi, \pi]$ (continuous + square integrable)
We can represent functions from this set using Fourier expansions

$$f(x) = \sum_{k=0}^{\infty} c_k \exp(ikx), \quad c_k \propto \int_{-\pi}^{\pi} f(x) \exp(-ikx) dx$$

Note the L_2 norm is given by $\|f\|_{L_2}^2 = \sum_{k=0}^{\infty} c_k^2 < \infty$.

$$\left(\|f(x)\|_{L_2}^2 = \int_{-\pi}^{\pi} \left(\sum_i c_i \phi_i(x) \right) \left(\sum_j c_j \phi_j(x) \right) dx = \sum_i \sum_j c_i c_j \int_{-\pi}^{\pi} \phi_i(x) \phi_j(x) dx = \sum_i \sum_j c_i c_j \delta_{ij} = \sum_{k=0}^{\infty} c_k^2 \right)$$

- Now consider the Sobolev space W_2^s, defined here as $W_2^s = \{ f \in C_2[-\pi, \pi] : \|D^s f\|_{L_2}^2 < \infty \}$.

→ Here, we define the norm as $\|f\|_{W_2^s}^2 = \|D^s f\|_{L_2}^2$, which is given by $\|f\|_{W_2^s}^2 = \sum_{k=0}^{\infty} k^{2s} c_k^2 < \infty$

$$\left(\|f(x)\|_{W_2^s}^2 = \|D^s f\|_{L_2}^2 = \left\| \sum_k (ik)^s c_k e^{ikx} \right\|_{L_2}^2 = \sum_k |(ik)^s| c_k^2 = \sum_k k^{2s} c_k^2 \right)$$

↑ s → high order terms in Fourier series have less contribution.

* For the norm to bounded, c_k need to decay at a rate which increases w/ s.

- Consider an increasing space H_n as the set of trigonometric polynomials of degree n. The optimal approx. is

$$f_n(x) = \sum_{k=1}^{n} c_k \exp(ikx)$$

For $f \in W_2^s$, the approximation error is given by $\mathcal{E}_n[f] < \frac{1}{n^{2s}} \|f\|_{W_2^s}^2$

$$\left(\mathcal{E}_n[f] = \|f - f_n\|_{L_2}^2 = \left\| \sum_{k=n+1}^{\infty} c_k \exp(ikx) \right\|_{L_2}^2 = \sum_{k=n+1}^{\infty} c_k^2 = \sum_{k=n+1}^{\infty} c_k^2 \frac{k^{2s}}{k^{2s}} < \frac{1}{n^{2s}} \sum_{k=n+1}^{\infty} k^{2s} c_k^2 < \frac{1}{n^{2s}} \sum_{k=1}^{\infty} k^{2s} c_k^2 = \frac{1}{n^{2s}} \|f\|_{W_2^s}^2 \right)$$

($k > n$)

* The rate of convergence is faster and impact of increasing model complexity n is greater as the smoothness of the function being approx. s increases

- For a function in a $d=2$ dim. domain, we have $f_n(x) = \sum_{k=1}^{n} \sum_{m=1}^{n} c_{km} \exp(ikx + imy)$

$$\|f\|_{W_2^s}^2 = \|D^s f\|_{L_2}^2 = \|D^s f_x\|_{L_2}^2 + \|D^s f_y\|_{L_2}^2 = \sum_{k=1}^{\infty} \sum_{m=1}^{\infty} k^{2s} c_{km}^2 + \sum_{k=1}^{\infty} \sum_{m=1}^{\infty} m^{2s} c_{km}^2 = \sum_{k=1}^{\infty} \sum_{m=1}^{\infty} (k^{2s} + m^{2s}) c_{km}^2 \rightarrow \mathcal{E}_n[f] < \frac{1}{n^{2s}} \|f\|_{W_2^s}^2$$

For a function in a d dim. domain, we have $f_n(x) = \sum_{i_1 \dots i_d=1}^{n} c_{i_1 \dots i_d} \exp(i [i_1 x_1 + \dots i_d x_d])$

$$\|f\|_{W_2^s}^2 = \sum_{i_1 \dots i_d=1}^{\infty} (i_1^{2s} + \dots + i_d^{2s}) c_{i_1 \dots i_d}^2 \rightarrow \mathcal{E}[f] < \frac{1}{d n^{2s/d}} \|f\|_{W_2^s}^2$$

Gaussian measures in Hilbert space

Lebesgue measure in infinite dimensions

- The Lebesgue measure m_L is a generalization of length. In $\mathbb{R}$, we have $m_L([a,b]) = b-a$.
- The Lebesgue measure m_L has the following properties:
  - ↳ Finiteness and monotonicity: If $A \subset B \subset \mathbb{R}$, then $0 \le m_L(A) \le m_L(B) \le \infty$
  - ↳ Translation invariance: $\forall A \subset \mathbb{R}, x_0 \in \mathbb{R}$, $m_L(A + x_0) = m_L(A)$, where $A + x_0 = \{x + x_0 : x \in A\}$.
- In $\mathbb{R}^D$, we have $m_L([a_1,b_1] \times \dots \times [a_D,b_D]) = \prod_{d=1}^{D} (b_d - a_d)$.

Consider the case $D = \infty$. For a Hilbert space H, there is a countable orthonormal set (acts as a basis in H, i.e. $\{e_i \mid i = 1,2,\dots\}$ w/ $\langle e_i, e_j \rangle = \delta_{ij}$.

We have $\|e_i - e_j\|^2 = \|e_i\|^2 - 2\langle e_i, e_j \rangle + \|e_j\|^2 = 1 + 1 = 2$ if $i \neq j$.

Denoting a ball centered at c w/ radius r as $B(c,r)$, place $\{B(e_i, \frac{1}{2})\}_{i \in \mathbb{N}}$ in $B(0,2)$.

→ we have $B(e_i, \frac{1}{2}) \cap B(e_j, \frac{1}{2}) = \emptyset$ and $\bigcup_{i \in \mathbb{N}} B(e_i, \frac{1}{2}) \subset B(0,2)$.

Monotonicity + countable additivity of m_L: $\sum_{i \in \mathbb{N}} m_L(B(e_i, \frac{1}{2})) \le m_L(B(0,2)) < \infty$. [1]

(m_L is σ-finite)

Translation invariance of m_L: $B(e_i, \frac{1}{2}) = B(e_j, \frac{1}{2}) \to \sum_{i \in \mathbb{N}} m_L(B(e_i, \frac{1}{2})) = \text{const} \times \sum_{i \in \mathbb{N}} = \infty$. [2]

$\Rightarrow$ [1], [2] contradict, so there is no Lebesgue measure m_L in infinite dim. Hilbert spaces).

Gaussian measure in infinite dimensions

- Define the infinite dimensional product measure on $\mathbb{R}^\infty$ as

$$\mu = \prod_{n=1}^{\infty} \mu_n$$

where each μ_n is the standard Gaussian in $\mathbb{R}$, $\mu_n(dx) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}x^2)\,dx$

- Consider the function f_λ

$$f_\lambda(x) = \exp\left(-\frac{\lambda}{2} \sum_{k=1}^{\infty} a_k x_k^2\right)$$

which is only well-defined (finite) if $\sum_{k=1}^{\infty} a_k x_k^2 < \infty$.

We define the Hilbert space $\ell_{2,a}$ w/ inner product $\langle x, y \rangle = \sum_{k=1}^{\infty} a_k x_k y_k$.

$$x \in \ell_{2,a} = \left\{ x \in \mathbb{R}^\infty : \sum_{k=1}^{\infty} a_k x_k^2 = \|x\|_{\ell_{2,a}} < \infty \right\}$$

Taking the Lebesgue integral wrt infinite dim. standard Gaussian product measure,

$$\int_{\mathbb{R}^\infty} f_\lambda \, d\mu = \int_{\mathbb{R}^\infty} f_\lambda \prod_{k=1}^{\infty} d\mu_k = \prod_{k=1}^{\infty} \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\lambda}{2}[1 + \lambda a_k] x_k^2\right) dx_k = \prod_{k=1}^{\infty} \sigma_k \int_{-\infty}^{\infty} N(x_k \mid 0, \sigma_k^2)\,dx_k = \prod_{k=1}^{\infty} \sigma_k$$

where $\sigma_k = \frac{1}{\sqrt{1 + \lambda a_k}}$, so we have $\int_{\mathbb{R}^\infty} f_\lambda \, d\mu = \frac{1}{\sqrt{\prod_{k=1}^{\infty} (1 + \lambda a_k)}}$

* If $\sum_{k=1}^{\infty} a_k < \infty$, then $\prod_{k=1}^{\infty} (1 + \lambda a_k) < \infty$.

As $\lambda \to 0$, $f_\lambda(x) = 1$ if $x \in \ell_{2,a}$, i.e. $f_\lambda(x) = \mathbb{1}_{x \in \ell_{2,a}} \to \lim_{\lambda \to 0} \int_{\mathbb{R}^\infty} f_\lambda \, d\mu = \int_{\ell_{2,a}} d\mu = \mu(\ell_{2,a}) = 1$.

i.e. we have a prob. measure μ that gives full measure to whole seq. space $\ell_{2,a}$ (isomorphic to L_2).

- It can be shown that in ℓ^2 w/ Gaussian measure $\mu = N(0, S_\lambda = (a_n \lambda_n)_{n=1}^{\infty})$.

$$x \in \ell^2 \quad \text{if} \quad \sum_{k=1}^{\infty} a_k < \infty.$$

i.e. infinite dim. seq. $x \in \ell^2$ has finite Gaussian measure

Full covariance operators in infinite dimensions

matrix but infinite dim.

- In a Hilbert space H, the Gaussian measure μ w/ mean M and covariance operator C $\mu = N(m, C)$ has the rep. of linear operator C being trace class / nuclear, i.e. $\text{trace}(C) < \infty$.
- The trace operator in Hilbert space w/ an orthonormal basis is defined by

$$\text{trace}(C) = \sum_{n \geq 1} \langle e_n, Ce_n \rangle$$

Note for zero mean functions f, g on domain D s.t. $f, g \in H$, we define for such,

$$\begin{aligned} \langle f, Cg \rangle &= E[\langle f, u \rangle \langle u, g \rangle] = E\left[\int_D \int_D f(x)(u(x)u(y))g(y)\,dy\,dx\right] \\ &= E\left[\int_D f(x)\left(\int_D u(x)u(y)g(y)dy\right)dx\right] = \int_D f(x)\left[\int_D E[u(x)u(y)]g(y)dy\right]dx \\ &= \int_D f(x)\left(\int_D c(x,y)g(y)dy\right)dx. \end{aligned}$$

so we have $Ce_n = \int_D c(x,y)e_n(y)\,dy$, $\langle e_n, Ce_n \rangle = \int_D e_n(x) \int_D c(x,y) e_n(y)\,dy\,dx$.

* $c(x,y)$ is the covariance function defining a Gaussian process.

Equivalence of Gaussian measures

- In finite dim, two Gaussian measures are absolutely continuous (equivalent) wrt each other.
  e.g. In $\mathbb{R}$, $\mu = N(0, \sigma^2)$, $\mu_m = N(m, \sigma^2)$, both have the densities wrt Lebesgue measure
  → for $\sigma \neq 0$, any $A \in \mathbb{R}$, $\mu(A) = 0 \Leftrightarrow \mu_m(A) = 0 \Leftrightarrow m_L(A) = 0$.
- However, in ℓ^2, $\mu = N(0, C)$ and $\mu_v = N(v, C)$ are either equivalent or singular wrt each other, depending on v. (Gaussian measure is quasi invariant)
- The subspace containing all translations v for which the measures are absolutely continuous wrt each other is the image of the space ℓ^2 under the operator $C^{1/2}$. (This space is the Cameron-Martin space / RKHS).

Bayes rule in Hilbert space

- In Hilbert space, the Radon-Nikodym derivative btwn a posterior measure μ^y and a prior measure μ^0 can define the corresponding likelihood function

$$\frac{d\mu^y}{d\mu^0}(x) = L(y|x)$$

- The reference measure for a Hilbert space is the Gaussian measure, so the prior measure over functions is $\mu^0 = N(m, C)$. (formal defn of Gaussian process prior)

Advanced MCMC

High dimensional MCMC

Gaussian random walk Metropolis Hastings (GRW-MH) algorithm

- For GRW-MH, the candidate sample v is generated by adding Gaussian noise $\xi \sim N(0,C)$ to the current sample u

$$\boxed{v = u + \beta\xi}$$

where β is the step size parameter

- The corresponding proposal function q is symmetric

$$q(u,v) = N(v \mid u, \beta^2 C) = N(u \mid v, \beta^2 C) = q(v,u)$$

The acceptance probability α therefore can be simplified to

$$\alpha(u,v) = \min\left\{\frac{\pi(v)\,q(v,u)}{\pi(u)\,q(u,v)}, 1\right\} = \min\left\{\frac{\pi(v)}{\pi(u)}, 1\right\}.$$

- In an infinite dim Hilbert space H, the proposal is well-defined and random draws will have finite norm and converge if C is trace class

There is no Lebesgue measure in infinite dim → replace $\pi(x)$ w/ $\mu^Y(x)$.

- For the case $\mu^0 = N(0,C)$ and $\frac{d\mu^Y}{d\mu^0}(x) \propto L(y|x) = \exp(-\Phi(x))$, the acceptance probability α becomes

$$\boxed{\alpha(u,v) = \min\left\{\frac{\mu^Y(v)}{\mu^Y(u)}, 1\right\} = \min\left\{\frac{L(y|v)\mu^0(v)}{L(y|u)\mu^0(u)}, 1\right\} = \min\left\{\frac{\exp(-\Phi(v))\exp(-\frac{1}{2}\langle v, C^{-1}v\rangle)}{\exp(-\Phi(u))\exp(-\frac{1}{2}\langle u, C^{-1}u\rangle)}, 1\right\} = \min\left\{\exp(J(v)-J(u)), 1\right\}}$$

where $J(x) = -\frac{1}{2}\langle x, C^{-1}x\rangle - \Phi(x) = -\frac{1}{2}\|C^{-1/2}x\|^2 - \Phi(x)$. Note $\langle x, C^{-1}x\rangle = \|x\|_{C^{-1}}$

- Let the eigenvalues of C be λ_i^{-2}. $u \sim N(0,C) \to u_i \sim N(0,\lambda_i^{-2}) \to \frac{1}{\lambda_i}u_i \sim N(0,1)$.

The eigenvalues of $C^{-1/2}$ are $\frac{1}{\lambda_i}$, so $E[\|C^{-1/2}u\|^2] = E[\sum_i (\frac{1}{\lambda_i}u_i)^2] = \sum_i \underbrace{\mathrm{Var}[\frac{1}{\lambda_i}u_i]}_{\to 1} = \infty$.

→ GRW-MH does not work for infinite dim.

Preconditioned Crank-Nicholson (pCN) algorithm

- For pCN, the candidate sample v is a linear combination of the current sample u and Gaussian noise $\xi \sim N(0,C)$

$$\boxed{v = \sqrt{1-\beta^2}\,u + \beta\xi}$$

where β is the step size parameter

- The proposal function is given by $q(x,y) = N(y \mid \sqrt{1-\beta^2}x, \beta^2 C) \propto \exp\left(-\frac{1}{2\beta^2}\langle y - \sqrt{1-\beta^2}x, C^{-1}(y-\sqrt{1-\beta^2}x)\rangle\right)$, so

$$-2\beta^2 \log\frac{q(v,u)}{q(u,v)} = \cancel{\langle u, C^{-1}u\rangle} - \cancel{\langle v, C^{-1}v\rangle} - 2\cancel{\sqrt{1-\beta^2}}\langle v, C^{-1}u\rangle + 2\cancel{\sqrt{1-\beta^2}}\langle v, C^{-1}u\rangle - \cancel{(1-\beta^2)}\langle u, C^{-1}u\rangle + \cancel{(1-\beta^2)}\langle v, C^{-1}v\rangle$$

$$\therefore \frac{q(v,u)}{q(u,v)} = \exp\left(-\frac{1}{2}\langle u, C^{-1}u\rangle + \frac{1}{2}\langle v, C^{-1}v\rangle\right)$$

The acceptance probability α therefore can be simplified to

$$\alpha(u,v) = \min\left\{\frac{\mu^Y(v)\,q(v,u)}{\mu^Y(u)\,q(u,v)}, 1\right\} = \min\left\{\frac{L(y|v)\,\mu^0(v)\exp(-\frac{1}{2}\langle u, C^{-1}u\rangle)}{L(y|u)\,\mu^0(u)\exp(-\frac{1}{2}\langle v, C^{-1}v\rangle)}, 1\right\}$$

- For the case $\mu^0 = N(0,C)$ and $\frac{d\mu^Y}{d\mu^0}(x) \propto L(y|x) = \exp(\Phi(x))$, the acceptance probability α becomes

$$\boxed{\alpha(u,v) = \min\left\{\frac{L(y|v)}{L(y|u)}, 1\right\} = \min\left\{\exp(\Phi(u)-\Phi(v)), 1\right\}}$$

→ well defined for infinite dim, provided Φ is well defined.

Derivation of pCN algorithm.

- consider the stochastic DE w/ $\mu = N(0,C)$ as an invariant measure

$$du = -u\,ds + \sqrt{2C}\,db$$

where b is standard brownian motion.

- Applying the Theta discretisation scheme yields a discrete-time Markov chain

$$v - u = -\delta\left((1-\theta)u + \theta v\right) + \sqrt{2C\delta}\,\xi_0, \quad \theta \in [0,1]$$

where u is the current position, v is the next position, δ is the discrete time diff, and $\xi_0 \sim N(0,1)$

Rearranging for v: $\quad v = (1+\delta\theta)^{-1}\left(u - \delta(1-\theta)u + \sqrt{2C\delta}\,\xi_0\right)$

Setting $\theta = 1/2$: $\quad v = \left(1+\frac{\delta}{2}\right)^{-1}\left(\frac{2-\delta}{2}u + \sqrt{2C\delta}\,\xi_0\right)$

Let $\beta = \sqrt{2\delta}\left(1+\frac{\delta}{2}\right)^{-1}$: $\quad v = \sqrt{1-\beta^2}\,u + \beta\xi_0$

- It turns out the two measures defining the acceptance prob. are equivalent only for $\theta = \frac{1}{2}$

Taking the i-th coordinate of $v = (1+\delta\theta)^{-1}\left(u - \delta(1-\theta)u + \sqrt{2C\delta}\,\xi_0\right)$, we have

$$v_i = \frac{1-\delta(1-\theta)}{1+\delta\theta}u_i + \frac{\sqrt{2\delta\lambda_i^2}}{1+\delta\theta}g_i$$

where λ_i^2 are the eigenvalues of C and g_i is the standard Normal

The corresponding expectation and variance are given by

$$E[v_i] = 0 \qquad \mathrm{Var}[v_i] = \left(\frac{1-\delta(1-\theta)}{1+\delta\theta}\right)^2\lambda_i^2 + \frac{2\delta\lambda_i^2}{(1+\delta\theta)^2}$$

For two product Gaussian measures to be equivalent, we have the convergence criterion

$$\sum_{i=1}^{\infty}\left(\frac{\mathrm{Var}[v_i]}{\mathrm{Var}[u_i]} - 1\right)^2 < \infty, \quad \text{i.e. } \frac{\mathrm{Var}[v_i]}{\mathrm{Var}[u_i]} = 1$$

$$\Rightarrow \frac{\mathrm{Var}[v_i]}{\mathrm{Var}[u_i]} = \left(\frac{1-\delta(1-\theta)}{1+\delta\theta}\right)^2 + \frac{2\delta}{(1+\delta\theta)^2} = 1 \rightarrow \theta = 1/2.$$

Langevin MCMC

Log concave and m-strongly log-concave

- A probability measure w/ density π is log-concave if

$$\log\pi(\lambda x + (1-\lambda)y) \geq \lambda\log\pi(x) + (1-\lambda)\log\pi(y), \quad \lambda \in (0,1)$$

More compactly, $\pi(\lambda x + (1-\lambda)y) \geq \pi(x)^{\lambda}\,\pi(y)^{1-\lambda}$

- A probability measure w/ density π is m-strongly log-concave if

$$\log\pi(\lambda x + (1-\lambda)y) \geq \lambda\log\pi(x) + (1-\lambda)(\log\pi(y) + \frac{m\lambda(1-\lambda)}{2}\|x-y\|^2, \quad \lambda \in (0,1)$$

This is a stronger notion that req. there not be flat regions and the max. is unique

↳ e.g. Gaussians — log-density = quadratic function

The Langevin SDE

- The Langevin SDE is given by

$$dX_t = -\nabla U(X_t)\,dt + \sqrt{2}\,dB_t$$

where $U:\mathbb{R}^d \to \mathbb{R}$ is a potential and $\{B_t\}_{t\geq 0}$ is Brownian motion

- The Langevin SDE has the stationary measure

$$\boxed{\pi \propto \exp(-U(x))}$$

i.e. to sample from π, we set $U(x) = -\log \pi(x) \to \nabla U(x) = -\nabla \log \pi(x)$. so we have

$$\boxed{dX_t = \nabla \log \pi(X_t)\,dt + \sqrt{2}\,dB_t}$$

- Before reaching the stationary measure, $X_t \sim \rho_t$ satisfies the Fokker-Planck equation

$$\boxed{\frac{\partial \rho_t}{\partial t} = -\nabla \cdot (\rho_t \nabla \log \pi) + \nabla^2 \rho_t}$$

Asymptotically, we have $\partial \rho_t = 0$, $(as\ t \to \infty)$.

$$\nabla^2 \rho_t = \nabla \cdot (\rho_t \nabla \log \pi) \to \nabla \rho_t = \rho_t \nabla \log \pi \to \frac{1}{\rho_t}\nabla \rho_t = \nabla \log \pi \to \nabla \log \rho_t = \nabla \log \pi.$$

$$\therefore \boxed{\lim_{t\to\infty} \rho_t = \pi}$$

→ Simulating the SDE provided samples from π in the limit.

- The Langevin SDE has many nice properties:

↳ For a m-strongly log-concave π, the convergence rate to the stationary measure π is exponentially fast,

$$\boxed{W_2(\pi_t, \pi) \leq e^{-mt}\left\{\|x - x^*\| + (d/m)^{1/2}\right\} \quad \text{for } \pi_0 = \delta_x,\ x^* = \operatorname*{argmax}_{x\in\mathbb{R}^d} \log \pi(x)}$$

where W_2 is a distance on the space of prob. measures.

Metropolis-adjusted Langevin algorithm (MALA)

- The continuous SDE can't be simulated exactly → use discretization schemes. but it introduces bias → we correct for the bias via a Metropolis accept/reject step.

- consider first-order Euler discretisation. using $X_0 \sim \pi_0$, we have

$$X_{k+1} = X_k + \gamma \nabla \log \pi(X_k) + \sqrt{2\gamma}\, W_{k+1}$$

where $\{W_k\}_{k\geq 0}$ are standard Normal and $\gamma > 0$ is the step size.

- Given the state of the Markov chain X_k, propose candidate sample $\hat{X}_{k+1}$

$$\boxed{\hat{X}_{k+1} = X_k + \gamma \nabla \log \pi(X_k) + \sqrt{2\gamma}\, W_{k+1}}$$

and accept w/ probability

$$\boxed{\alpha(X_k, \hat{X}_{k+1}) = \min\left\{\frac{\pi(\hat{X}_{k+1})\, q(\hat{X}_{k+1}, X_k)}{\pi(X_k)\, q(X_k, \hat{X}_{k+1})},\ 1\right\}}$$

where $q(x, x') = N(x' \mid x + \gamma \nabla \log \pi(x),\ 2\gamma) \propto \exp\left(-\frac{1}{4\gamma}\|x' - x - \gamma \nabla \log \pi(x)\|^2\right)$.

- The step size γ has to be tuned s.t. $\boxed{\alpha = 0.574}$.

- MALA may suffer in high dimensions d — acceptance prob. α may get exponentially slower in d. which results in poor mixing.

Unadjusted Langevin algorithm (ULA)

- Run a Langevin SDE discretisation w/o the Metropolis step, i.e.

$$X_{k+1} = X_k + \gamma \nabla \log \pi(X_k) + \sqrt{2\gamma}\, W_{k+1}$$

the chain converges to a stationary measure π_γ which is not the target measure π, but performs well in high dim. (Metropolis step avoided)

- consider the example of a Gaussian target $\pi(x) = N(x|\mu, \Sigma)$.

The Langevin SDE in this case is the Ornstein-Uhlenbeck (OU) process.

$$dX_t = -\Sigma^{-1}(X_t - \mu)dt + \sqrt{2}\, dB_t$$

the ULA scheme is given by

$$X_{k+1} = X_k - \gamma \Sigma^{-1}(X_k - \mu) + \sqrt{2\gamma}\, W_{k+1}$$

Under appropriate conditions, the chain has the invariant measure

$$\pi_\gamma = N(\mu, \Sigma(I - \tfrac{\gamma}{2}\Sigma^{-1})^{-1})$$ ← proof in EP, X_{k+1} is Gaussian under $E[X_{k+1}]$, $\mathrm{Var}[X_{k+1}]$

→ simulates the correct mean but introduces bias to the variance

- For m-strongly log-concave π, the convergence rate of the ULA w/ $X_0 \sim \pi_0$

$$W_2(\pi_k, \pi) \le (1-m\gamma)^k W_2(\pi_0, \pi) + 1.65\,(L/m)(\gamma d)^{1/2}, \quad \gamma \le 2/m+L$$

where we have L-Lipschitz gradients and W_2 is a distance on the space of prob. measures

* the asymptotic bias is of order $O(\gamma^{1/2})$ — can be made arbitrarily small w/ small γ.

Langevin schemes for Bayesian inference

- consider target distributions of the form $\pi(x) \propto p(x) \prod_{i=1}^n L(y_i|x)$, where $p(x)$ is the prior over x and $L(y_i|x)$ is the likelihood for data vector y_i. usually $n \gg 1$.

- If we apply MALA, we run into two problems:

↳ Gradient computation for the proposal is too costly. $O(n)$ → bottleneck.

$$\nabla \log \pi(x) = \nabla \log p(x) + \sum_{i=1}^n \nabla \log L(y_i|x)$$

↳ Acceptance prob. computation too costly. $O(n)$ → avoided w/ Metropolis-free schemes.

$$\alpha = \min\left\{\frac{\pi(X_{k+1})\, q(X_{k+1}, X_k)}{\pi(X_k)\, q(X_k, X_{k+1})}, 1\right\}, \quad \pi(x) \propto p(x) \prod_{i=1}^n L(y_i|x)$$

- We can estimate the gradient cheaply if we subsample data — random mini batch.

$$\widehat{\nabla \log \pi(X_k)} = \nabla \log p(X_k) + \frac{n}{|I_k|} \sum_{i \in I_k} \nabla \log L(y_i|X_k)$$

where $I_k \subset \{1, \dots n\}$ is the index set of iteration k, and $|I_k|$ is the size of the mini batch.

- If we run ULA w/ $\widehat{\nabla \log \pi(X_k)}$, we have stochastic gradient Langevin dynamics (SGLD)

$$X_{k+1} = X_k + \gamma \widehat{\nabla \log \pi(X_k)} + \sqrt{2\gamma}\, W_{k+1}$$

where $E[\widehat{\nabla \log \pi(X_k)}] = \nabla \log \pi(X_k)$ and each iteration costs $O(|I_k|)$

- For m-strongly log-concave π, we have similar convergence guarantees as ULA

$$W_2(\pi_k, \pi) \le O(e^{-\gamma m k} + \gamma \tfrac{n}{|I|})$$

* we just need to ensure the variance of the gradient estimates are controlled.