Markov Chains

Course by Alexey Naumov and Sergey Samsonov, HSE, 2022

Contents

1 Lecture 1

1.1 Conditional expectation

Definition 1. Let $(\Omega, \mathcal{F}, P)$ be a probability space; $G \subseteq \mathcal{F}$ is a σ-algebra; ξ is a random variable, such that $\mathbb{E}|\xi| < \infty$. Then **conditional expectation** $\mathbb{E}(\xi|G)$ is a random variable, such that:

1. $\mathbb{E}(\xi|G)$ is G-measurable.

2. $\forall A \in G \int\limits_A \mathbb{E}(\xi|G)(w)P(dw) = \int\limits_A \xi(w)P(dw) = \mathbb{E}(\xi I_A)$

Definition 2. Let $(\Omega, \mathcal{F}, P)$ be a probability space; ξ, η are random variables. Then **conditional expectation of ξ with respect to random variable** η is $\mathbb{E}(\xi|\eta) = \mathbb{E}(\xi|\sigma_\eta)$, where $\sigma_\eta = \eta^{-1}(B), B \in \mathcal{B}(\mathbb{R})$.

We know that any function, which is measurable with respect to σ_η, can be represented as a Borel function from η, i.e there exists a Borel function $g : \mathbb{R} \to \mathbb{R}$, such that $\mathbb{E}(\xi|\eta) = g(\eta)$ P-a.s..

Definition 3. Let $G \subseteq \mathcal{F}$ be a σ-algebra. Then $\forall A \in \mathcal{F}$ **conditional probabiliy of event** A **with respect to** G $P(A|G)(w) = \mathbb{E}(I_A|G)(w)$.

Let us substitute Def. 3 into Def. 1:

$$\forall A \in \mathcal{F}, \forall B \in G \int\limits_B P(A|G)(w)P(dw) = \int\limits_B I_A P(dw) = P(A \cap B).$$

If $P(A|G)$ is constant on B, then for $G = \{\varnothing, \Omega, B, \overline{B}\}$ we receive $P(A|B) \cdot P(B) = P(A \cap B)$.

Let us derive the definition for $\mathbb{E}(\xi|\eta = y)$, where $y \in \mathbb{R}$, ξ, η are random variables on $(\Omega, \mathcal{F}, P)$. By Def. 2:

$$\forall A \in \sigma_\eta : \int\limits_A \xi(w)P(dw) = \int\limits_A \mathbb{E}(\xi|\eta)P(dw) = \int\limits_{\{w:\eta(w)\in B\}} \mathbb{E}(\xi|\eta)P(dw),$$

where B is a Borel set, which equals to $\eta(A)$. Then

$$\int\limits_{\{w:\eta(w)\in B\}} \mathbb{E}(\xi|\eta)P(dw) = \int\limits_{\{w:\eta(w)\in B\}} g(\eta(w))P(dw) = \int\limits_B g(x)P_\eta(dx),$$

where P_η is the distribution of η. On the last step we have changed the variable in the Lebesgue integral.

Definition 4. $\mathbb{E}(\xi|\eta = y)$ is a Borel function from y $g(y) : \mathbb{R} \to \mathbb{R}$:

$$\forall B \in \mathcal{B}(\mathbb{R}) \int\limits_{\{w:\eta(w)\in B\}} \xi(w)P(dw) = \int\limits_B g(x)P_\eta(dx).$$

Note that this function is P_η-a.s. unique by Radon-Nikodym theorem.

Definition 5. $P(A|\eta = y) = \mathbb{E}(I_A|\eta = y)$ $\forall A \in \mathcal{F}$.

Substituting Def. 5 into Def. 4: $P(A \cap \{\eta(w) \in B\}) = \int\limits_B P(A|\eta = y) P_\eta(dy)$.

If we fix y in the last definition, we will receive a probability distribution. If it is defined only on B with zero measure, its value is not fixed. Let us look at some examples.

Example. Let η be a random variable with a countable number of values $(x_k)_{k=1}^\infty$, $P(\eta = x_k) = p_k > 0$, $\sum\limits_{k=1}^\infty p_k = 1$. Then

$$P(A|\eta = x_k) = \frac{P(A \cap \{w|\eta(w) = x_k\})}{p_k},$$

because

$$\int\limits_B P(A|\eta = y) P_\eta(dy) = \sum_{x_k \in B} P(A|\eta = x_k) p_k.$$

Note that when $y \neq x_k$ $P(A|\eta = y)$ can be defined in any way, because it is defined only on the set of measure zero.

Example. Let $\forall B \in \mathcal{B}(\mathbb{R}^2)$ $P((\xi, \eta) \in B) = \int\limits_B f_{\xi,\eta}(x, y) dxdy$, $f_\eta(y) = \int\limits_{\mathbb{R}} f_{\xi,\eta}(x, y) dx$. Note that 2 absolutely continuous random variables always have joint distribution. However, it is not always absolutely continuous (e.g. distribution of (ξ, ξ) if ξ is a normal random variable). We can say that

$$f_{\xi|\eta}(x|y) = \begin{cases} \frac{f_{\xi,\eta}(x,y)}{f_\eta(y)} & \text{if } f_\eta(y) > 0 \\ 0 & \text{if } f_\eta(y) = 0 \end{cases}$$

Note that it is not important what we put in the second case. To prove that formula we have to check that

$$\forall A \in \mathcal{B}(\mathbb{R}) \ P(\{w|\xi(w) \in A\}|\eta = y) = \int\limits_A f_{\xi|\eta}(x|y) dx,$$

i.e. we have to check that

$$\int\limits_B P(\{w|\xi(w) \in A\}|\eta = y) P_\eta(dy) = \int\limits_{A \times B} f_{\xi,\eta}(x, y) dxdy.$$

Using Fubini's theorem and the fact that $P_\eta(dy) = f_\eta(y)dy$, since η is absolutely continuous, we receive that the left part equals to

$$\int\limits_B \left(\int\limits_A \frac{f_{\xi,\eta}(x, y)}{f_\eta(y)} dx \right) f_\eta(y) dy,$$

which, in turn, equals to the right part.

1.2 Markov kernels

Let η be a random variable defined on probability space $(\Omega, \mathcal{F}, P)$. Then $P(A|\eta = y)$ is define $\forall y \in \mathbb{R}$. Is it a measure? We know that $P(A|\eta = y) \geqslant 0$. From the linearity of conditional expectation we also have finite additivity. However, we have to check σ-additivity, i.e. we have to check if $P(\bigcup\limits_{i=1}^\infty A_i|\eta = y) = \sum\limits_{i=1}^\infty P(A_i|\eta = y)$ for $A_i \cap A_j = \varnothing$. It turns out that this equality holds true, but only P_η-a.s. To prove this, we have to represent the left part as

$\mathbb{E}(I_{\bigcup\limits_{i=1}^{\infty} A_i}|\eta = y)$. Then we can introduce $B_n = \bigcup\limits_{i=1}^{n} A_i, B_n \uparrow \bigcup\limits^{\infty} A_i$. Then $I_{B_n} \uparrow I_{\bigcup\limits^{\infty} A_k}$. After that, using Lebesgue dominated theorem, we can show that $\mathbb{E}(I_{B_n}|\eta = y) \to \mathbb{E}(I_{\bigcup\limits^{\infty} A_k}|\eta = y)$.

As the equality holds only P_η-a.s., there is a problem: for any sequence of sets A_i there will be its own set of measure 0, where the equality does not hold. As the number of sequences of A_i has the cardinality of the continuum, we can not just remove all these sets of measure zero. Therefore, $P(A|\eta = y)$ is not a measure. Hence we have to define regular conditional probabilities.

Definition 6. Let $(\Omega, \mathcal{F}, P)$ be a probability space; (E, Σ) and (G, J) are measurable spaces. Then the function $N : E \times J \to [0, 1]$ is a **probability kernel** (or **Markov kernel**), if:

- $N(x, \cdot)$ is a probability measure on (G, J) for any fixed $x \in E$;
- $N(\cdot, B)$ is Σ-measurable for any $B \in J$ (i.e. it is a measurable map $(E, \Sigma) \to ([0, 1], \mathcal{B}([0, 1]))$.

If $(E, \Sigma) = (G, J)$, then N is a probability kernel on (E, Σ).

Definition 7. Let $(\Omega, \mathcal{F}, P)$ be a probability space; $\xi \to (E, \Sigma), \eta \to (G, J)$. Then probability kernel N is a **regular conditional probability of η given ξ**, if

$$\forall A \in \Sigma, B \in J \ P(\xi \in A, \eta \in B) = \int\limits_A N(x, B) P_\xi(dx).$$

The idea of introducing this entity is to define something similar to $P(A|\eta = y)$, so that it would be a measure. Now we have to understand when this regular conditional probability exists and if it is unique. But first let us look at an example.

Example. Let ξ, η be independent random variables. Then $N(x, B) = P_\eta(B)$, because $P(\xi \in A, \eta \in B) = P(\xi \in A)P(\eta \in B)$. It is clear that N satisfies both properties from Def. 6. Note that if we have a Markov kernel, it coincides with regular conditional probability P_η-a.s..

Now let us think about **uniqueness** of N. Let there be two probability kernels $N, N' : E \times J \to [0, 1]$. Then

$$\forall B \in J \int\limits_A N(x, B) P_\xi(dx) = \int\limits_A N'(x, B) P_\xi(dx),$$

because by definition both parts equal to $P(\xi \in A, \eta \in B)$. Recall that if ξ, η are G-measurable, then

$$\forall A \in G \ \int\limits_A \xi P(dw) \leqslant \int\limits_A \eta P(dw) \Rightarrow \xi \leqslant \eta \ P\text{-a.s.}.$$

Using this fact and previously derived equality of integrals, we get

$$N(x, B) \geqslant 0 \Rightarrow N(x, B) - N'(x, B) = 0 \ P_\xi\text{-a.s.}.$$

Note that, again, for a set with zero measure P_ξ, the kernel can be defined arbitrarily, but this time it has to satisfy the first property of a probability kernel, i.e. be a measure. To conclude, all that we know about the uniqueness is

$$\forall B \in J \ N(x, B) = N'(x, B) \ P\text{-a.s.}.$$

The next point in question is **existence** of a probability kernel. Here we can say that if (G, J) is a complete separable (there exists a dense countable subset) metric space and J is its Borel σ-algebra, then there exists regular conditional distribution $N : E \times J \to [0, 1]$.

1.3 Markov chains

Definition 8. Let $(\Omega, \mathcal{F}, P)$ be a probability space. Then $(x_t)_{t\in T}$ is a **random process**, if all x_t are random variables.

Definition 9. Let $T = \mathbb{N}$, i.e. consider time to be discrete. Then let us define $\mathcal{F}_k^x = \sigma(x_0, \dots x_k)$ as a σ-algebra, where all $x_i, i \leqslant k$ are measurable. Then $(\mathcal{F}_k^x)_{k=0}^{\infty}$ is a **natural filtration**, associated with $(x_t)_{t\in T}$.

Definition 10. Let $\mathcal{F}_k \subseteq \mathcal{F}$ be σ-algebras. Then $\{\mathcal{F}_k\}_{k=0}^{\infty}$ is a **filtration**, if $\mathcal{F}_k \subseteq \mathcal{F}_{k+1}$ $\forall k \in \mathbb{N}$.

Definition 11. $(x_k)_{k=0}^{\infty}$ is **Markov chain**, if

$$P(x_{k+1} \in A|\mathcal{F}_k^x) = P(x_{k+1} \in A|x_k) \ P\text{-a.s.}.$$

An equivalent definition: for any bounded measurable function φ

$$\mathbb{E}(\varphi(x_{k+1})|\mathcal{F}_k^x) = \mathbb{E}(\varphi(x_{k+1}|x_k)) \ P\text{-a.s.}.$$

Theorem 1. *The following statements are equivalent:*

1. *(x_k) is a Markov chain*
2. *For any y, which is $\sigma(x_j, j \geqslant k+1)$-measurable, such that $\mathbb{E}|y| < \infty$ it is true that $\forall k$ $\mathbb{E}(y|\mathcal{F}_k^x) = \mathbb{E}(y|x_k)$, where $\mathcal{F}_k^x$ is a natural filtration. Note that here nothing will change if we include x_k into σ-algebra.*
3. *$\forall y$, which is $\sigma(x_j, j \geqslant k+1)$-measurable, $\forall z$, which is $\mathcal{F}_k^x$-measurable, $\forall k$ $\mathbb{E}(yz|x_k) = \mathbb{E}(y|x_k)|_{\mathbb{k}})$ P-a.s., where $\mathcal{F}_k^x$ is a natural filtration.*

Definition 12. Let Q be a Markov kernel on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, $f(x)$ is a bounded measurable functions. Let us define

$$Qf(x) = \int\limits_{\mathbb{R}} f(y)Q(x, dy).$$

Note that by definition $Q(x, dy)$ is a measure.

Remark. If a Markov chain is homogeneous, $\mathbb{E}(f(x_{k+1})|x_k = x) = Qf(x)$.

Lemma. *$Qf(x)$ is measurable and bounded: $\|Qf\|_\infty \leqslant \|f\|_\infty$.*

Sketch of proof. Let $f \geqslant 0, f_n \uparrow f$, f_n are simple functions. Then we can use Lebesgue dominated theorem and get that $Qf(x) = \lim\limits_{n\to\infty} Qf_n(x)$, which is measurable.

$$\forall x \in \mathbb{R} \ |\int\limits_{\mathbb{R}} f(y)Q(x, dy)| \leqslant \int\limits_{\mathbb{R}} |f(y)|Q(x, dy) \leqslant \|f\|_\infty,$$

because $|f(y)| \leqslant \|f\|_\infty$. □

Definition 13. Let ν be a probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Then

$$\nu Q(B) = \int\limits_{\mathbb{R}} \nu(dy)Q(y, B).$$

Lemma. *νQ is a probability measure on* $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

Definition 14. Let Q be a Markov kernel on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Then $Q^{\otimes n}$ is a Markov kernel on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$, such that $\forall$ bounded measurable $f : \mathbb{R}^n \to \mathbb{R}$

$$Q^{\otimes n} f(y) = \int\limits_{\mathbb{R}^n} f(x_1, \ldots, x_n) Q(y, dx_1) Q(x_1, dx_2) \ldots Q(x_{n-1}, dx_n).$$

Definition 15. Let ν be a probability measure, Q is a Markov kernel. Then

$$\nu \otimes Q(A \times B) = \int\limits_A \nu(dy) Q(y, B).$$

Remark. $\nu \otimes Q$ is a measure on $(\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2))$.

Remark.

$$Q^{\otimes n}(y, A_1 \times \ldots \times A_n) = \int\limits_{A_1 \times \ldots \times A_n} Q(y, dx_1) Q(x_1, dx_2) \ldots Q(x_{n-1} dx_n).$$

Definition 16. Let $(\Omega, \mathcal{F}, P)$ be a probability space; $(x_k)_{k=0}^{\infty}, x_k : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Then $(x_k)_{k=0}^{\infty}$ is a **time-homogeneous Markov chain** with Markov kernel Q, if

$$P(x_{k+1} \in A | x_k) = Q(x_k | A) \ P\text{-a.s.}.$$

Note that $P(x_{k+1} \in A | x_k) = P(x_{k+1} \in A | \mathcal{F}_k^x)$.

Theorem 2. *(x_k) is a time-homogeneous Markov chain with Markov kernel Q and initial distribution ν $\Leftrightarrow$ $P(x_0 \in A_0, \ldots x_n \in A_n) = \nu \times Q^{\otimes n}(A_0 \times A_1 \times \ldots \times A_n)$ $\forall A_0, \ldots, A_n \in \mathbb{R}$.*

2 Lecture 2

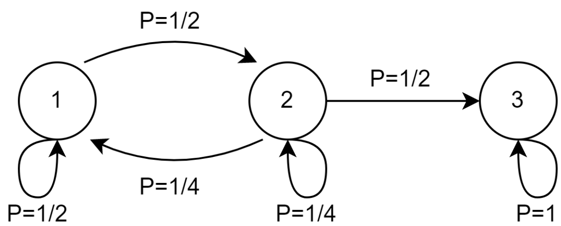
2.1 Examples of Markov chains

At the previous lecture we introduced a concept of a Markov chain, let's now consider several examples:

2.1.1 Example 1. Finite-state

Let $X = [1, 2, \ldots, r]$ be a final state Markov chain, then Markov kernel $P(x, A) = \sum_{y \in A} P_{xy}$.

Notice, that in final case, kernel can be represented as a transition matrix. For example, let the chain be:



then, the transition matrix will be:

$$\begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \\ 0 & 0 & 1 \end{bmatrix}$$

2.1.2 Example 2. Random walk

Consider $x_{k+1} = x_k + \xi_{k+1}$, where

- ξ_{k+1} - i.i.d. random variables, and $\xi \perp\!\!\!\perp \mathcal{F}_k$ - independent
- $\xi \overset{i.i.d.}{\sim} Q$ - some probability space

$Q = N(0, \sigma^2)$, $x_{k+1}|x_k \sim N(x_k, \sigma^2)$. Then, if we fix $x = x_k$, the probability of transition to the set A is:
$P(x, A) = \frac{1}{\sqrt{(2\pi)\sigma}} \int_A exp(-\frac{(y-x)^2}{2\sigma^2})dy$

2.1.3 Example 3. Langevin dynamics (LD, ULA)

Consider $x_{k+1} = x_k - \gamma U(x_k) + \sqrt{2\gamma}\xi_{k+1}$
ξ_{k+1} - i.i.d., $\xi_{k+1} \perp\!\!\!\perp x_k$ (meaning $\xi_{k+1} \perp\!\!\!\perp \sigma(x_k)$) $\xi_{k+1} \sim N(0, I)$ γ - some positive constant.

It represents discretization of continious Langevin dynamics: $dX_+ = -\nabla U(x_+)dt + \sqrt{2}dW_+$
Without the last term it is very similar to SGD. By discretization of dt when $\gamma \to 0$, it converges to continious case.

$$\pi(x) = e^{-U(x)}/z_d \Rightarrow \int \pi(x)dx = 1$$

where d is a dimension.

Example. Let $U(x) = x^2/2$ then the kernel is just a normal distribution. Therefore we know z_d. Usually, z_d is unknown, but in Langevin dynamics it is not necessary, we need only the gradient:

$$e^{-U(x)}/z_d = e^{-U(x)-\log(z_d)}$$

With some conditions, for example if U and gradient are Lipschitz functions:

$$Law(X_k) \to \pi_\gamma \approx \pi \text{ when } \gamma \approx 0$$

When X_k is fixed, the kernel:

$$P(x, A) = \frac{1}{\sqrt{4\pi\gamma}} \int_A \exp\left(-\frac{(y - x + \gamma\nabla U(x))^2}{4\gamma}\right) dy, \text{ when } d = 1$$

2.1.4 Example 4. Reinforcement learning

Let's consider an extension of Markov chain - Markov decision process. It is a Markov chain, with added so called actions and rewards.

In this case Ω is (S, A), where S - state space, A - action space. Consider only finite case.

$(S_k)_{k\geqslant 0}$ - sequence of states. At each state the action is taken with probability $\pi(\cdot|S)$ called **policy** or **strategy**.

$$S_{k+1} \sim P(\cdot|s_k = s, A_k = a) = P(s,a|\cdot)$$

Example from finance:

Let s_k - current amount of money, $a_k \in [0,1]$ - share of invested money, p - probability of winning, $(1-p)$ - probability of losing. $P(\xi_{x+1} = 1) = p = 1 - P(\xi_{k+1} = -1)$

$$S_{k+1} = s_k(1 + \xi_{k+1}A_k)$$

$$P^\pi(s_{k+1} \in \{s\}|s_k = s) = \sum_{a \in A} P(s_k,a|\{s'\})\pi(a|s_k)$$

It can be shown, that it is a Markov kernel on $(S, \sigma(S))$.

Definition 17. Markov decision process (MDP) is a tuple of 4 elements (S, A, P_a, R_a), where:

1. S - **state space** - is a set of states
2. A - **action space** - is a set of states
3. $P_a(s,s')$ - probability that action a in state s at time t will lead to state s' at time $t + 1$
4. $R_a(s,s')$ - is the immediate reward (or expected immediate reward) received after transitioning from state s to state s', due to action a

2.2 Action on measures

$\mu P(A) = \int_x \mu(dx)P(x,A)$ - action on measures.
Where $P(A)$ - is distribution of Markov chain at 1st step. $x_0 \sim \mu$ $P_\mu(X_1 \in A)$
To simulate Markov chain, we need to fix initial distribution, from which X_0 and the kernel were picked.

$$Pf(x) = \int f(y)P(x,dy), \quad X = [1,2,...,r]$$

In discrete case: $P_\mu(x_1 = j) = \sum_{k \in X} \mu(k)P(k)$
$Pf(j) = \sum_{k \in X} f(k)P(j,k)$ - expectation with respect to measure P.

$$\mu(f) = \int f(y)\mu(dy)$$

$$P^n(x,A) = \int_X P(x,dy)P^{n-1}(y,A)$$

$$P^2(x,A) = \int_X P(x,dy)P(y,A)$$

$$P^2(i,j) = \sum_{k \in X} P(i,k)P(k,j) = [P^2]_{ij}$$

$$A = \{j\}$$

Choose $\mu = \delta_x P_{\delta_x}(x_n \in A) = P^n(x, A)$

$$P_\mu(x_n \in A) = \mu P^n(A)$$

Definition 18. Markov chain is called **homogeneous Markov chain** in the case if the kernel remains constant.

Remark. In principal, the kernel can be changed, for example: decreasing step of Langevin.

2.3 Tensor product of kernels

$$P \otimes Pf(x) = \int_{x \times x} f(y,z) P(x,dy) P(y, dz)$$

$$f : X^2 \to \mathbb{R}$$

$$f(y,z) = I(y \in A, z \in B)$$

$$P \otimes Pf(x) = P_{\delta_x}(X_1 \in A, X_2 \in B)$$

$$\int_{A \times B} P(x, dy) P(y, dz)$$

Or just $\mu P \otimes Pf(x)$

3 Seminar 1

3.1 Discrete state-space Markov Chains

Let S - finite or countable state space; $(X_k)_{k=0}^{\infty}$; $X_k \in S$ (if $|S| < \infty$, $S = \{1,2,...,n\}$); where $(X_k)_{k=0}^{\infty}$ defined on $(\Omega, \mathcal{F}, P)$

$P(X_{k+1} \in A | \mathcal{F}_k) = P(X_k, A)$

It is enough to define $p_{ij} = P(X_{k+1} = j \mid X_k = i)$. In this case $P(i, A) = \sum_{j \in A} p_{ij}$; $P(i, S) = 1 \Rightarrow \sum_j p_{ij} = 1$

Let $P = (p_{ij}) \in \mathbb{R}^{|S| \times |S|}$, $p_{ij} \geqslant 0$

$\sum_{j \in S} p_{ij} = 1 \; \forall i$ - row-stochastic matrix.

Any measure μ on S: $\mu = (\mu_i)_{i \in S}$ - vector of length $|S|$.

$$\mu P(A) = \sum_{j \in A} \sum_{i \in S} \mu_i p_{ij} = \sum_{j \in A} \mu P \quad \forall A \subseteq S$$

$$P^n(i, j) = \sum_{j_1, \ldots, j_{n-1} \in S} p_{ij_1} p_{j1} p_{j2} ... p_{j_{n-1}j} = (P^n)_{ij}$$

$$f : S \to \mathbb{R}, \; f = (f_i)_{i \in S} = (f(i))_{i \in S}$$

$$Pf(i) = \int_S P(i, dy) f(y) = \sum_{j \in S} p_{ij} f_j = (Pf)_i$$

3.2 Tensor product

Remark. Act on measures is left multiplication, and act on functions - is right multiplication.

Define:

$$\mu \otimes P^{\otimes n}(A_0, ...,) := \sum_{i_a \in A} \mu \otimes P^{\otimes n}(i_0, ..., i_n)$$

$$P_\mu(X_0 \in A_0, X_1 \in A_1, ..., X_n \in A_n)$$

3.3 Classification of the states

Definition 19. State i is **connected** with j, if $\exists n_1, n_2 : \; P_{ij}^{(n_1)} > 0, P_{ij}^{(n_1)} > 0; \; n_1{,}n_2 \geqslant 0$

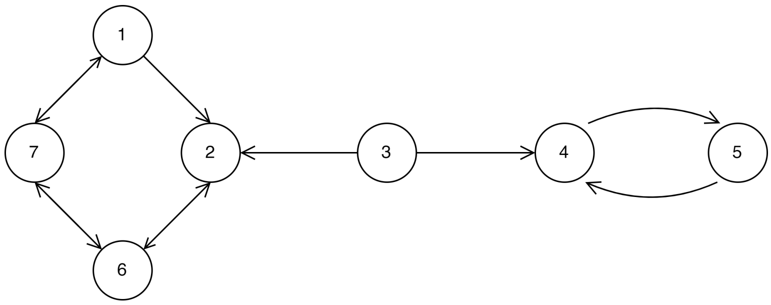
Remark. Note, if $i \longleftrightarrow j; j \longleftrightarrow k \Rightarrow i \longleftrightarrow k$

Lemma. *S can be divided into non-intersecting communicating classes w.r.t. the relation $\longleftrightarrow$. Then communicating class $C_i = \{i{,}j \in \delta : i \longleftrightarrow j\}$.*

Definition 20. Communicating class C_i is **closed**, if any $j \in S : \; i \to j$ belongs to C
$i \to j$, if $\exists n_1 : P_{ij}^{(n_1)} > 0$, $n_1 \in \mathbb{N}$.

Definition 21. Transition matrix P is called **irreducible**, if it has only one communicating class S (and it has to be closed).

Example. Consider the following chain:



In this case, communicating classes are: {1,2,6,7} (closed); {3} (not closed); {4,5} (closed).
$\pi P^n \to ?, \quad n \to \infty$
$\pi P = \pi$, $\pi = \delta_3 = (0{,}0{,}1{,}0{,}0{,}0{,}0)$
So, starting from the state 3 and then going to the left, we forever stuck in the 1st communicating class.

Definition 22. State $i \in S$ is **recurrent**, if $P_{S_i}(i \text{ is visited } \infty \text{ many times}) = 1$

Definition 23. State $i \in S$ is **non-recurrent**, if $P_{S_i}(i \text{ is visited finitely many times}) = 1$

Lemma. *Every state $i \in S$ is either recurrent or non-recurrent.*

Proof. Define:

- $V := \sum \mathbb{I}\{X_n = i\}$
- Also let random variable $\tau_i := \inf\{n \geqslant 1 : X_n = i\}$
- $p_i := P_{s_i}(\tau_i < 0) = P(\tau_i < \infty | X_0 = i)$

Then

$$P_{\delta_i}(V_i = \infty) = P_{\delta_i}\left(\bigcap_{k=1}^{\infty}\{V_i \geqslant k\}\right) = \lim_{k\to\infty} P_{\delta_i}(V_i \geqslant k)$$

$$P(V_i > 1) = p_i = P_{\delta_i}(\tau_i < \infty)$$

We need to prove $P_{\delta_i}(V_i > k) = p_i^k$

$$
\begin{aligned}
P_{\delta_i}(V_i > 2) = \sum_{k\geqslant 1}\sum_{m\geqslant 1} P_{\delta_i}(X_0 = i, X_1 \neq i, ..., X_k = i, X_{k+1} \neq i, ..., X_{k+m} = i) = \\
\sum_{k\geqslant 1}\sum_{m\geqslant 1} P_{\delta_i}(X_{k+1} \neq i, ..., X_{k+m} = i \mid X_0 = i, ..., X_1 \neq i, ..., X_k = i) P_{\delta_i}(x_0 = i, x_1 \neq 0, ..., x_k = i) \\
\sum_{k\geqslant 1}\sum_{m\geqslant 1} P_{\delta_i}(X_{k+1} \neq i, X_{k+m} = i \mid X_k = i) P_{\delta_i}(\tau_i = k) = \\
\sum_{k\geqslant 1}\sum_{m\geqslant 1} P_{\delta_i}(\tau_i = m) P_{\delta_i}(\tau_i = k) = \\
\left(\sum_{m\geqslant 1} P_{\delta_i}(\tau_i = m)\right)\left(\sum_{k\geqslant 1} P_{\delta_i}(\tau_i = k)\right) = p_i^2
\end{aligned}
$$

$$
P_{\delta_i}(V_i = \infty) = \lim_{k\to\infty} p_i^k = \begin{cases} 1 & \text{if } p_i = 1 \\ 0 & \text{otherwise} \end{cases}
$$

□

4 Lecture 3

4.1 Kolmogorov's strong law of large numbers

Suppose $x_j, j \geqslant 0$ are independent and identically distributed with π as an invariant probability. The law of large numbers will hold, i.e.

$$
Law(x_j) = \pi, \forall f : \pi(f) < \infty
$$

$$
\frac{1}{n}\sum_{j=0}^{n-1} f(x_j) \xrightarrow[n\to\infty]{a.s} \pi(f) = \int_X f(x)\pi(dx)
$$

Example: We will need numerical method to solve the integral since $X = \mathbb{R}^5$ or any higher dimensions. The points are in a grid and we should take the points with high probability (classic Monte-Carlo method). It takes a d-dimensional ball B, then its volume can be estimated as the following:

$$
B^d(r) \approx r^d
$$

$$
B^d(1) \approx 1
$$

$$
B^d(0.99) \approx (0.99)^d \approx 0(d >> 1)
$$

Considering the d-dimensional sphere S:

$$
S^{d-1}
$$

$$
x_j \sim \mathbb{N}(0,1); j = 1,...,d
$$

$$
\theta = \frac{(x_1,...,x_d)}{\sqrt{{x_1}^2 + ... + {x_d}^2}} \sim U(S^{d-1}) \text{ (check!)}
$$

$$
X \sim \mathbb{N}(0,\Sigma)
$$

$$
Y \sim \mathbb{N}(0,I), Y_1,...,Y_d \overset{i.i.d.}{\sim} \mathbb{N}(0,1)
$$

$$X = \Sigma^{\frac{1}{2}} Y$$

Check how to sample from normal distribution.
Aim: Sampling from distribution π, we need to find Markov kernel P such that $(X,x)-$ state space

$$\forall \xi \in P_1(X) : Law_{\xi}(X_N) = \xi P^N \approx \pi$$

or

$$\xi P^N \xrightarrow{N \to \infty} \pi$$

where P_1 - probability distribution on (X,x), N - burn-in period.

$$d(\xi P^N, \pi) \leqslant \Sigma(X,\xi)$$

4.2 Invariant

π is invariant with respect to (w.r.t) P if

$$\pi P = \pi$$

$$\pi P = Law_{\pi}(X_1) = Law_{\pi}(X_0)$$
$$\pi P^2 = (\pi P)P = \pi P = P$$

where $\pi P^2 = Law_{\pi}(x_2), \pi = Law_{\pi}(x_0)$.

Assume that we can start from $\pi(\xi = \pi)$ and π is invariant w.r.t. P

$$Law_{\pi}(x_j) = \pi$$

x_0,x_1,...,$x_N \sim \pi$, but x_j are all dependent.

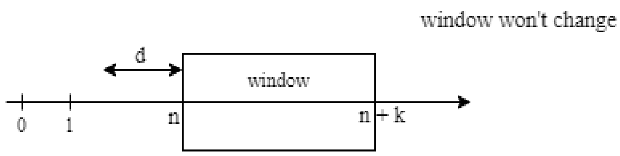
$$x_0 \sim \xi(\cdot)$$
$$x_1 \sim P(x_0, \cdot)$$
$$x_2 \sim P(x_1, \cdot)$$
$$Law_{\xi}(x_N) \approx \pi$$

How to construct kernel P?

Proposition: Let $(x_k)_{k \geqslant 0}$ is a Markov chain with kernel P and initial distribution π. $(x_k)_{k \geqslant 0}$ is stationary iff $\pi P = \pi$

Stationary

$$Law(x_n, x_{n+1}, ..., x_{n+k}) = (x_0, ..., x_k)$$



Proof
1) $(x_k)_{k \geqslant 0}$ stationary.

$$\pi = Law_{\pi}(x_0) = Law_{\pi}(x_1) = \pi P$$

2)

$$\pi P = \pi$$
$$Law_{\pi}(x_n, x_{n+1}, ..., x_{n+k}) = \pi P^n \otimes P^{\otimes k} = \pi \otimes P^{\otimes k} (\text{independent of n})$$

4.3 Total variation distance

$$d(\xi P^n, \pi) \leqslant ?$$
$$X = [1,...,r]$$
$$\sum_{j:\xi(j)\geqslant\mu(j)} (\xi(j) - \mu(\xi)) = \sum_{j=1}^{r} |\xi(j) - \mu(j)| \text{ (check!)}$$
$$d_{TV}(\xi,\mu) = \frac{1}{2}\sum_{j=1}^{r} |\xi(j) - \mu(j)| = \frac{1}{2}\sup_{A\subseteq X} |\xi(A) - \mu(A)| \text{ (check!)}$$
$$= \frac{1}{2}\sup_{f:X\to[-1,1]} \left| \int_X f d\xi - \int_f d\mu \right|$$

We can take test function f and test the distance.

4.4 Kantorovich Wasserstein distance

$$W_{d,p}(\xi,\mu) = \left\{ \inf_{\zeta\in\pi(\xi,\mu)} \int_{X\times X} d^p(x,y)\zeta(dx,dy) \right\}^{\frac{1}{p}}$$

The quatity $W_{d,p}(\xi,\mu)$ is called the *Kantorovich Wasserstein distance* between two probability measures ξ and μ.
where d - metric; $\pi(\xi,\mu)$ coupling of ξ and μ :

$$\zeta(X,A) = \mu(A)$$
$$\zeta(A,X) = \xi(A)$$
$$d(x,y) = 1_{\{x\neq y\}} \Rightarrow \text{Total variation} \to \text{Kantorovich distance 1}$$
$$d(x,y) = \|x - y\|_2 \to \text{Kantorovich distance 2}$$

4.5 Exponential convergence in total variation for ergodic transition matrices

Take arbitrary kernel Q :

$$d_{TV}(\xi Q, \mu Q) = \sum_{j\in J} (\xi Q(j) - \mu Q(j)) = \Big[J := \{j : \xi Q(j) \geqslant \mu Q(j)\} \Big] =$$
$$= \sum_{j\in J}\sum k \in X(\xi(k)Q(k,j) - \mu(k)Q(k,j)) \leqslant \sum_{k:\xi(k)\geqslant\mu(k)} (\xi(k) - \mu(k)) \sum_{j\subset X} Q(k,j) = d_{TV}(\xi,\mu)$$
$$\exists a > 0 : Q(k,j) \geqslant a > 0 \forall k,j \in X$$

Take $Q = P^s \Rightarrow \exists s \in \mathbb{N} : P^s(i,j) \geqslant a > 0 \ \forall i,j \in X$

$$\xi_n := \xi_0 P^n$$
$$d_{TV}(\xi_n,\xi_{n+k}) = d_{TV}(\xi P^n,\xi P^{n+k}) \leqslant (1-a)d_{TV}(\xi P^{n-s},\xi P^{n+k-s}) \leqslant$$
$$\leqslant (1-a)^m d_{TV}(\xi P^{n-sm},\xi P^{n+k-sm}) \leqslant (1-a)^m$$

$$m : 0 < n - sm < s$$

$$\text{If } n \to \infty, k \to \infty \Rightarrow m \to \infty$$

$$\Rightarrow \pi := \lim_{n \to \infty} \xi P^n \to \text{Limiting point of Cauchy sequence}$$

$$\pi P = \lim_{n \to \infty} \xi P^n P = \lim_{n \to \infty} \xi P^{n+1} = \pi$$

$$\exists \ \pi_1 \neq \pi_2$$

$$\pi_1 P = \pi_1, \pi_2 P = \pi_2$$

$$d_{TV}(\pi_1,\pi_2) = d_{TV}(\pi_1 P,\pi_2 P) \leqslant (1-a) d_{TV}(\pi_1,\pi_2) \Rightarrow \pi_1 = \pi_2$$

$$d_{TV}(\xi P^n,\pi) = d_{TV}(\xi P^n,\pi P^n) \leqslant (1-a)^m d_{TV}(\xi P^{n-ms},\pi P^{n-ms}) \leqslant (1-a)^m \leqslant (1-a)^{\frac{n}{s}-1} = (1-a)^{-1}(\beta)^n$$

$$\beta = (1-a)^{\frac{1}{s}} < 1$$

If $s >> 1 \Rightarrow \beta \to 1 \Rightarrow$ convergence can be very slow

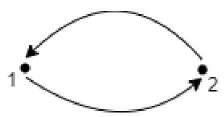
5 Seminar 2

5.1 Recurrent and non-recurrent

Example

$$\pi P = \pi \not\Rightarrow \lambda P^n \to \pi, n \to \infty, \text{where } \lambda \in \mathbb{R}^{|s|} - \text{initial distribution}$$

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



$$\pi P = \pi \Rightarrow \pi = \left(\frac{1}{2}, \frac{1}{2}\right) - \text{invariant}$$

$$\lambda P^n = \begin{cases} (\lambda_1,\lambda_2), n = 2k \\ (\lambda_2,\lambda_1), n = 2k+1 \end{cases}$$

P - irreducible, not ergodic

$$\lambda = (\lambda_1, \lambda_2)$$

$$P_{i,j}^{(m)} > 0 \Rightarrow (\xi P^n,\pi) \leqslant c.\rho^{\lceil \frac{n}{m} \rceil}$$

Definition 24. State $i \in S$ is recurrent, if $P(V_i = \infty | x_0 = i) = 1$

$$V_i = \sum_{n=0}^{\infty} I\{x_n = i\}$$

$i \in S$ is non-recurrent, if $P(V_i = \infty | x_0 = i) = 0$

Each state $i \in S$ is either recurrent, or non-recurrent

$$\tau_i := \inf n \geqslant 1, x_n = i; g_i = P(\tau_i < \infty | x_0 = i)$$

$$P(V_i > k | x_0 = i) = g_i^k$$

Corollary. State $i \in S$ is recurrent, if $\sum_{n \geqslant 0} P_{ii}^{(n)} = \infty$
State $i \in S$ is non recurrent, if $\sum_{n \geqslant 0} P_{ii}^{(n)} < \infty$ where $P_{ii}^{(n)} := P(x_n = i | x_0 = i)$
Proof

$$\underbrace{E[V_i | x_0 = i]}_{=\sum_{k=1}^{\infty} kP(V_i = k)} = \underbrace{\sum_{k=0}^{\infty} P(V_i \geqslant k | x_0 = i)}_{=\sum_{l=k}^{\infty} P(V_i = l)} = \sum_{k=0}^{\infty} g_i^k = \begin{cases} \infty; g_i = 1 \\ \frac{1}{1-g_i}; g_i < 1 \end{cases}$$

$$E[V_i | x_0 = i] = E\Big[\sum_{n=0}^{\infty} I\{x_n = i\} | x_0 = i\Big] = \sum_{n=0}^{\infty} p_{ii}^{(n)}$$

Corollary. If i is recurrent/non-recurrent, $j \leftrightarrow i \Rightarrow j$ is also recurrent/non-recurrent.
Proof

$$\exists r : p_{ij}^{(r)} > 0; s : p_{ji}^{(s)} > 0$$

$$p_{ii}^{(n+r+s)} \geqslant p_{ij}^{(r)} p_{jj}^{(n)} p_{ji}^{(s)}$$

$$p_{jj}^{(n+r+s)} \geqslant \underbrace{p_{ji}^{(s)}}_{>0} p_{ii}^{(n)} \underbrace{p_{ij}^{(r)}}_{>0}$$

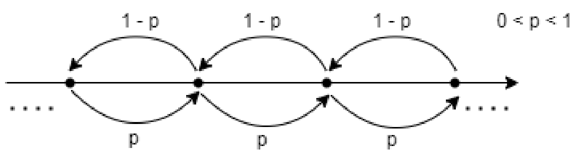
Hence, $p_{jj}^{(n)}$ and $p_{ii}^{(n)}$ converge or do not converge simultaneously.

5.2 Invariant measure

Theorem 3. *Let the transition matrix P be irreducible and recurrent (it has 1 communicating class, which is irreducible and recurrent). Then there exists invariant measure $(\mu(i))_{i \in S}$ where $0 \leqslant \mu(i) < \infty; \mu P = \mu; \mu-$ invariant measure; μ is unique up to proportionality constant.*

Corollary. Either $\underbrace{\sum_{i \in S}^{\infty} \mu(i) = \infty}_{\text{P is null recurrent}}$ for any invariant μ or $\underbrace{\sum_{i \in S}^{\infty} \mu(i) < \infty}_{\text{positive recurrent}}$

Proof J.K. Norris. lecture 8
Example. Random walk on $\mathbb{Z}$



i) irreducibility: $\exists\, j > i : p_{ij}^{(j-i)} = p^{(j-i)} > 0$
ii) recurrent
If state 0 is recurrent?
Proof

$$\sum_{n=0}^{\infty} P_{00}^{(n)} = \sum_{k=0}^{\infty} \frac{(2k)!}{(k!)^2} p^k (1-p)^k \sim c \sum_{k=0}^{\infty} \frac{1}{\sqrt{k}} \underbrace{(4p(1-p))^k}_{<1}$$

$$\sim \begin{cases} c\sum_{k=0}^{\infty}\frac{1}{\sqrt{k}}g^k, g<1, \text{for } p\neq\frac{1}{2} \\ c\sum_{k=0}^{\infty}\frac{1}{\sqrt{k}}=\infty, \text{for } p=\frac{1}{2}\end{cases}$$

$$P_{00}^{(n)}=\begin{cases}0, n=2k+1\\ C_{2k}^k p^k(1-p)^k, n=2k\end{cases}$$

$p=\frac{1}{2}\Leftrightarrow$ recurrent; $p\neq\frac{1}{2}\Leftrightarrow$ non-recurrent;
iii) invariant distributions: $(\pi(i))_{i\in\mathbb{Z}}$ is invariant $\Leftrightarrow \pi p=\pi$

$$(\pi p)_j=\pi(j-1)p+\pi(j+1)(1-p) \qquad \pi(j)=\pi_j$$

$$\pi_j=\pi_{j-1}p+\pi_{j+1}(1-p); j\in\mathbb{Z}$$

$$\pi_{j+1}(1-p)-\pi_j+\pi_{j-1}p=0$$

Characteristic polynomial:

$$\lambda^2(1-p)-\lambda+p=0$$

$$\lambda_1=1$$

$$\lambda_2=\frac{p}{1-p}, p\neq\frac{1}{2}$$

Hence $\pi_j=c_1\lambda_1^j+c_2\lambda_2^j=c_1+c_2\left(\frac{p}{1-p}\right)^j; c_1,c_2-$ constants

$$\pi_j\geqslant 0\Rightarrow c_1\geqslant 0; c_2\geqslant 0$$

$$\pi_j=1;\forall j\in\mathbb{Z}$$

$$\text{or } \pi_j=\left(\frac{p}{1-p}\right)^j;\forall j\in\mathbb{Z}$$

$p=\frac{1}{2}:\pi_j=c_1+c_2 j; c_1,c_2-$ constants; $\pi_j\geqslant 0\Rightarrow c_2=0, c_1>0$
So, in this case $\pi_j=c>0-$ unique invariant measure (up to proportionality).
Null recurrent Markov Chain.

5.3 Detailed balance condition

Definition 25. $P\in\mathbb{R}^{|s|\times|s|};\pi\in\mathbb{R}^{|s|};\pi\geqslant 0$ is in detailed balance with P, if

$$\underbrace{\pi_i P_{ij}}_{=P_\pi(x_0=i,x_1=j)} = \underbrace{\pi_j P_{ji}}_{=P_\pi(x_0=j,x_1=i)}, \forall(i,j)\in S\times S$$

Lemma. *If π is in detailed balance with $P\Rightarrow\pi$ is invariant, that is $\pi P=\pi$*

Proof

$$(\pi P)_j=\sum_{i\in S}\pi_i P_{ij}=\sum_{i\in S}\pi_j P_{ji}=\pi_j\underbrace{\left(\sum_{i\in S}P_{ji}\right)}_{=1}=\pi_j$$

Example. (Random walk)

$$\pi_i P_{i,i+1}=\pi_{i+1}P_{i+1,i}$$

$$\pi_i P=\pi_{i+1}(1-P)$$

$$\pi_{i+1}=\left(\frac{P}{1-P}\right)\pi_i\Rightarrow\pi_i=\left(\frac{P}{1-P}\right)^i - \text{in detailed balance with P}$$

5.4 Invariant distribution

$$G = (V,E); |V| < \infty, A = (a_{ij}); a_{ij} \geqslant 0; A = A^\top; a_{ij} = a_{ji}$$
$$P_{ij} = \frac{a_{ij}}{\sum_{j\in v} a_{ij}}, \forall (i,j) \in V \times V; \sum_{j\in V} a_{ij} > 0, \forall i \in V$$

Find invariant distribution.

$$\pi_i P_{ij} = \pi_j P_{ji} \Rightarrow \pi_i \frac{a_{ij}}{\sum_{j\in V} a_{ij}} = \pi_j \frac{a_{ji}}{\sum_{i\in V} a_{ji}}$$
$$\Rightarrow \pi_i (\sum_{i\in V} a_{ji}) = \pi_j (\sum_{j\in V} a_{ij})$$
$$\pi_i = \frac{\sum_{j\in V} a_{ij}}{\sum_i \sum_j a_{ij}} - \text{invariant distribution}$$

6 Lecture 4

6.1 Reversibility property

Definition 26. A kernel P is reversible w.r.t. ξ if $\xi \otimes P(A \times B) = \xi \otimes P(B \times A)$, $\forall A, B \in X$

If X is finite:

$$\xi(i)P(i,j) = \xi(j)P(j,i)$$
$$\mathbb{E}_\xi[f(x_0,x_1)] = \int\limits_{X\times X} f(x_0,x_1)\xi(dx_0)P(x_0,dx_1) =$$
$$= \int\limits_{X\times X} f(x_0,x_1)\xi(dx_1)P(x_1,dx_0) = \int\limits_{X\times X} f(x_1,x_0)\xi(dx_0)P(x_0,dx_1) = \mathbb{E}_\xi[f(x_1,x_0)]$$

Proposition. Let $\xi \in P_1(x)$ and P is reversible w.r.t. ξ. Then ξ is invariant.

$$\xi P(A) = \xi(A)$$

Proof.
$$\xi P(A) = P_\xi(x_1 \in A) = P_\xi(x_0 \in X, x_1 \in A) = \xi \otimes P(X \times A) =$$
$$= \xi \otimes P(A \times X) = P_\xi(x_0 \in A, x_1 \in X) = P_\xi(x_0 \in A) = \xi(A)$$

□

If sampling from π is difficult then it is possible to find an approximation such that:

$$\pi(dx) = \frac{\tilde{dx}}{Z_d}; Z_d = \int\limits_X \tilde{\pi}(dx)$$

From Bayesian statisticas $(Y_1, \dots Y_N)$,$\pi_0(\theta)$

$$P(\theta|Y) = \frac{\prod_{j=1}^r P(Y|\theta)\pi_0(\theta)}{\int \prod_{j=1}^r P(Y|\theta)\pi_0(\theta)\theta}$$

6.2 Metropolis-Hastings algorithm.

In this case is necessary to find a kernel P such that:

i) $\pi P = \pi$

ii) $d_{\pi x}(\mu P^n, \pi) \to 0$ as$n \to \infty$

iii) $X_0 \sim \mu \to$ it is easy, e.g. take $\mu = \delta_x$ or $\mu \sim \mathcal{N}(0,I)$, $x|j \sim P(x|j-1,\cdot)$

Definition 27. (Metropolis-Hastings algorithm) Take an easy to sample kernel $Q(x,dy) = q(x,y)dy$, for a fixed X_0 and assume that $X_0, X_1, \dots X_k$ were already sampled and we need to sample X_{k+1}, then we sample:

$$Y_{k+1} \sim Q(x_k,\cdot)$$

$$Y_{k+1} = \begin{cases} Y_{k+1}; \text{ with probability } \alpha(X_k, Y_{k+1})(\text{Accept proposal}) \\ X_k; \text{ with probability } 1-\alpha(X_k, Y_{k+1})(\text{Reject proposal}) \end{cases}$$

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

6.2.1 Example 1

Let's consider the kernel $q(x,y) = q(y,x)$, e.g. $q(x,y) = \bar{q}(|x-y|)$ and $Y_{k+1} = X_k + \xi_{k+1}$, $\xi_{k+1} \sim \bar{q}$, and the bimodal distribution in Figure 1.

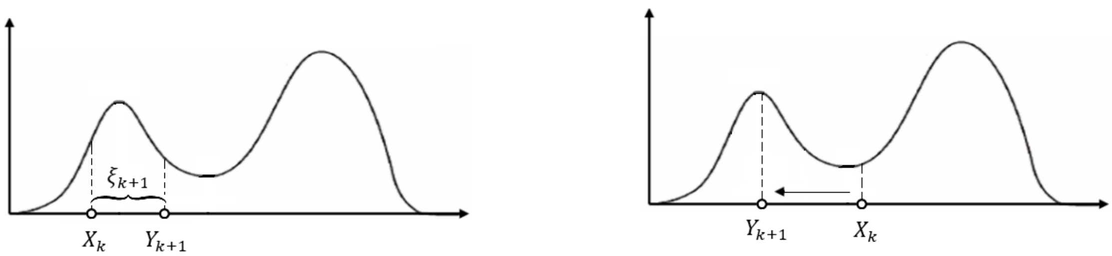


Figure 1.a: High probability state to lower.

Figure 1.b: Low probability state to higher.

Figure 1: Bimodal distributions

Considering the example of a bimodal distribution, the probability of moving from a high probability state to another with a lower one is (Figure 1.a):

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

While for the opposite case (Figure 1.b), the probability is $\alpha = 1$.

It is also possible to get the following scenario where it is not possible to move from one mode to the other one (a common problem in GANs).

In Figure 2 all the samples will be taken from a single mode, and the model will get stuck, to avoid this problem is necessary to pick a good initial point and a good kernel.

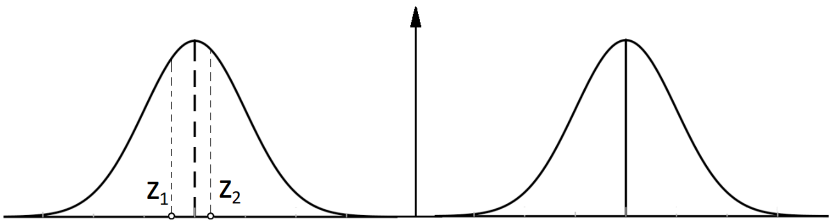


Figure 2: Mode Collapse

Definition 28. (Langevin algorithm)

$$dY_t = -\nabla U(Y_t)dt + \sqrt{2}dW_t$$

$$\pi(dx) = e^{-U(x)}$$

$$\text{Law}(Y_t) \to \pi$$

The Unaltered Langevin Algorithm (ULA) is:

$$Y_t = Y_t - \gamma \nabla U(Y_t) + \sqrt{2\gamma}\xi_{t+1}$$

where $\xi_{t+1} \overset{i.i.d.}{\sim} \mathcal{N}$

$$\text{Law}_{y_0}(Y_t) \xrightarrow{\text{u.s.c.}} \tilde{\pi}_\gamma \approx \Pi_{\gamma \to 0}$$

$$Y_{t+1} = X_t - \gamma \nabla U(X_t) + \sqrt{2}\xi_{t+1}$$

$$\text{Law}_{X_0} X_t \to \pi$$

Considering a kernel of M.H.

$$P(x,A) = \int\limits_A q(x,y)\alpha(x,y)dy + \int\limits_X (1 - \alpha(x,y))q(x,y)dy$$

for a initial point $\delta_x(A)$, considering the terms of the sum as (1) and (2), respectively.

Checking reversibility:

$$q(x,y)\alpha(x,y) = \min\{(\pi(x)q(x,y); \pi(y)q(x,y)\} = q(y,x)\alpha(y,x)$$

$$\int\limits_{X \times X} \pi(dx)P(x,y)f(x,y) = \int\limits_{X \times X} P(y,dx)f(x,y)$$

(1)

$$\int\limits_{X \times X} \pi(x)q(x,y)\alpha(x,y)f(x,y)dxdy = \int\limits_{X \times X} \pi(y)q(y,x)\alpha(y,x)f(x,y)dxdy$$

$$= \int\limits_{X \times X} \pi(x)q(x,y)\alpha(x,y)f(x,y)dxdy$$

(2)

$$\int\limits_{X\times X} \pi(x)q(x,y)(1-\alpha(x,y))q(x,y)\tau_x(dy)f(x,y)dx = \int\limits_{X} \pi(x)q(x,y)(1-\alpha(x,x))q(x,x)f(x,x)$$

$$= \int\limits_{X\times X} \pi(y)q(x,y)(1-\alpha(y,x))q(y,x)\delta_y(dx)f(y,x)dxdy$$

7 Lecture 5

7.1 φ-irreducibility. Aperiodicity. Ergodicity of φ-irreducible and aperiodic chain

Example. Then, the transition matrix will be:

$$\begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Where $\pi_0 = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, $\pi_0 P = \pi_0$.
$P_{\delta_{x_1}}(x_n = 3) \nrightarrow \pi_0(3)$, as $n \to \infty$.
But if we take $\pi_1 = (\frac{1}{2}, \frac{1}{2}, 0)$, still $\pi_1 P = \pi_1$, but it convergences to π_1.

Definition 29. Markov chain is φ**-irreducible** if $\exists$ δ-finite measure φ on (X, Ω) such, that for $\forall A \in \Omega$ with $\varphi(A) > 0$ and $\forall x \in X$ and $\exists n = n(x, A)$ such that $P^n(x, A) > 0$.

Example. $q \in C(X^2)$, $q > 0$
$\pi(A) = \int\limits_A \pi(x)\lambda(dx)$, $\lambda(dx) = dx$.
We want to find φ, which will prove φ-irreducibility.
Let's try $\varphi = \pi$ and fix $A : \pi(A) > 0$. There $\exists B_R(0)$: $A_R = A \cap B_R(0)$ and $\pi(A_R) > 0$.
Now A_R is limited and $\forall x \in \Omega$:

$$\inf_{y \in A_R} \{\min\{q(x,y), q(y,x)\} \geq \varepsilon\} > 0$$

$P(x, A) \geq P(x, A_R) \geq \int_{A_R} q(x,y)\alpha(x,y)dy$
Middle of formula
$\geq \varepsilon \int_{A_R^1} 1 \cdot \lambda(dy) + \varepsilon \int_{A_R^2} \frac{\pi(y)}{\pi(x)}\lambda(dy) = \varepsilon \cdot \lambda(A_R^1) + \frac{\varepsilon}{\pi(x)}\pi(A_R^2) > 0$

Example. Then, the transition matrix will be:

$$\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$$

Where $\varphi = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$.
Let's take $A : \varphi(A) > 0$ and $\forall x \exists n : P^n(x, A) > 0$.
If $A = 3, x = 1, \Longrightarrow$ for $n = 2 P^2(1{,}3) = 1$.
$\pi = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$
$\pi P = \pi$
$P_{\delta_{x_1}}(x_n = 3) \nrightarrow \pi(3)$

Definition 30. Markov kernel P with invariant distribution π is **aperiodic** if $\nexists d \geq 2$ and $\nexists x_1, ..., x_d$ such that $X_j \cap X_k = 0$, $X = X_1 \coprod ... \coprod X_d$,
$\forall x \in X_i, P(x, X_{i+1}) = 1$ and
$\forall x \in X_d, P(x, X_1) = 1$ and $\pi(x_i) > 0$

Example. Let $X = X_1 \coprod X_2$: $\pi(x_i) > 0$,
$X_{i_R} = X_i \cap B_{R_i}(0)$, and $\pi(x_{i_{R_i}}) > 0$
$\forall x \in X_i$:

$$\inf_{y \in x_{i_{R_i}}} \{\min\{q(x,y), q(y,x)\} \geq \varepsilon\} > 0$$

$P(x, X_1) \geq P(x, x_{1_{R_1}}) \geq \int_{x_{1_{R_1}}} q(x,y)\alpha(x,y)\lambda(dy) > 0$
then $P(x, X_2) \neq 1 \Rightarrow$ Chain is aperiodic.

$$d_{TV}(\mu, \nu) = \|\mu - \nu\|_{TV} = \sup_{f: X \to [0,1]} \left| \int_x f d\mu - \int_x f d\nu \right|$$

Theorem 4. *If*

- (X, Ω) *be such set that* Ω *is countable generated* σ*-algebra (Borel* σ*-algebra,* $X \subseteq R^d$*).*

- $\exists \pi : \pi P = \pi$

- φ*-irreducible and aperiodic*

Then $\lim_{n\to\infty} \|\sigma_x P^n - \pi\|_{TV} = 0$

Example.

$$\widetilde{\pi}(x) = \frac{1}{1+x^2}, Q(x,\cdot) = U[x-1, x+1]$$

UGE(uniformly geometrically ergodic):

$$\exists C, \rho : \|\delta P^n - \pi\|_{TV} \leq C\rho^n$$

GE(geometrically ergodic):

$$\exists M : X \to [0,\infty] : |\delta_x P^n - \pi\|_{TV} \leq M(x)^n,$$
$$M(x) < \infty, \pi - \text{a.s.}$$

A subset $C \subseteq X$ is small ((n_0,ε,ν)-small) if $\exists n_0 \in N, \varepsilon > 0$, and probability measure ν on (X, Ω) such that

$$P^{n_0}(x,A) \geq \varepsilon\nu(A), \forall x \in C, \forall A \in \Omega$$

(minorisation condition)

Example. Example of ε, ν construction:

$$\varepsilon_{n_0} = \sum_{y \in X} \inf_{x \in C} P^{n_0}(x,y) > 0$$

$$\nu(y) = \frac{1}{\varepsilon_{n_0}} \inf_{x \in C} P^{n_0}(x,y)$$

$$\|\mu - \nu\|_{TV} = \sup_A |\mu(A) - \nu(A)| \leq P(X \neq Y)$$

where $X \sim \mu, Y \sim \nu$

7.2 Coupling construction

Let's say $x_0 = x, x_0' \sim \pi$ and we have pair (x_n, x_n').

1. if $x_n = x_n'$, then $x_{n+1} = x_{n+1}' \sim P(x_n, \cdot)$ (i.e. further they will be equal)
2. else if $(x_n{,}x_n') \in C \times C$, with probability $\varepsilon : x_{n+n_0} = x_{n+n_0}' \sim \nu(\cdot)$,
   else with probability $1-\varepsilon$:
   $x_{n+n_0} \sim \frac{1}{1-\varepsilon}[P(x_n{,}\cdot) - \varepsilon\nu(\cdot)]$, $x_{n+n_0}' \sim \frac{1}{1-\varepsilon}[P(x_{n+1}'{,}\cdot) - \varepsilon\nu(\cdot)]$
3. else $x_{n+1} \sim P(x_n, \cdot)$, $x_{n+1}' \sim P(x_n'{,}\cdot)$

Example. $P_x(X_n \in A) = \delta_x P^n(A)$
$P_\pi(X_n' \in A) = \pi P^n(A) = \pi(A)$
$\|P^n(x{,}\cdot) - \pi(\cdot)\|_{TV} \leq P(X_n \neq X_n') \leq (1-\varepsilon)^{\frac{n}{n_0}}$

7.3 Drift condition

Definition 31. P satisfies drift condition if $\exists b > 0, \lambda \in (0{,}1)$, and $V : X \to [1{,}\infty)$ such that
$PV(x) \leq \lambda V(x) + b\mathbb{1}_C(x)$,
$PV(x) = \int\limits_X V(y)P(x{,}dy)$.

Then: $\pi(PV) \leq \lambda\pi(V) + b\pi(C)$.
If we integrate and take into account that $P\pi = \pi$:
$\pi(V) \leq \lambda\pi(V) + b \Rightarrow \pi(V) \leq \frac{b}{1-\lambda}$

7.4 Small set and drift condition

$\exists \rho \in (0{,}1) and C > 0$ such, that:

$$\sup_{|f|\leq V} |\int f(y)P^n(x{,}dy) - \int f(y)\pi(dy)| \leq CV(x)\rho^n$$

Definition 32. Let P be Markov kernel on $(X{,}F)$, ξ, ξ' — probability measures on $(X{,}F)$. Then **Dobrushin coefficient** is

$$\Delta(P) := \sup_{\xi\neq\xi'} \frac{\|\xi P - \xi' P\|_{TV}}{\|\xi - \xi'\|_{TV}}$$

Definition 33. P is called **Uniformly Geometrically ergodic (UGE)** if $\exists m \in N$: $\Delta(P^m) < 1$, where $\Delta(P^m) := \sup\limits_{\xi\neq\xi'} \frac{\|\xi P^m - \xi' P^m\|_{TV}}{\|\xi - \xi'\|_{TV}}$

Lemma. *P is UGE* $\Rightarrow \forall \xi \|\xi P^n - \pi\|_{TV} \leq \zeta\{\Delta(P^m)\}^{[\frac{n}{m}]}$, *where*
$\zeta = \max\limits_{0\leq k\leq m-1} \|\xi P^k - \pi\|_{TV} \leq 1$

Proof. $\|\xi P^n - \pi\|_{TV} = \{\text{as } \pi \text{ is invariant}\} = \|\xi P^n - \pi P^n\|_{TV} = \|\xi P^{n-m}P^m - \pi P^{n-m}P^m\|_{TV} \leq \Delta(P^m)\|\xi P^{n-m} - \pi P^{n-m}\|_{TV} \leq \{\Delta(P^m)\}^{[\frac{n}{m}]} \cdot \|\xi P^k - \pi\|_{TV}, k < m$ ☐

Definition 34. Space X is $(m{,}\varepsilon)$**-small**, if $\exists$ probability measure ν such that $\forall A \in \mathcal{F}$:

$$P^m(x{,}A) \geq \varepsilon \cdot \nu(A), \forall x \in X$$

Lemma. *If X is $(m{,}\varepsilon)$-**small**, $\Delta(P^m) \leq 1-\varepsilon$.*

Lemma. *In (1) it is enough to take $\xi = \delta_x, \xi' = \delta'_x, x \neq x'$ (Moulines)*

$\|S_x - S_{x'}\|_{TV} = 1, x = x'$
$P^m(x,A) \geq \varepsilon \cdot \nu(A)$
$P^m(x,A) = \varepsilon \cdot \nu(A) + \mu(A)$; $\mu(A) \geq 0$, $mu(\cdot)$ - non-negative measure
$\widetilde{\mu}(A) := \frac{1}{1-\varepsilon}\mu(A)$ – probability measure
$P^m(x,A) = \varepsilon \cdot \nu(A) + \widetilde{\mu}'(A) \cdot (1-\varepsilon)$; $\widetilde{\mu}, \widetilde{\mu}'$ — probability measures.
$\|P^m(x,\cdot) - P^m(x',\cdot)\|_{TV} = \|S_x P^m - S_x P^m\|_{TV} = \sup\limits_{A\in\mathcal{F}} |P^n(x,A) - P^n(x',A)| = (1-\varepsilon) \cdot \sup\limits_{A\in\mathcal{F}} |\widetilde{\mu}(A) - \widetilde{\mu}'(A)| \leq 1-\varepsilon$

Theorem 5. *(Metropolis-Hastings)*
$\pi(x) \sim \mathcal{N}(0{,}1)$ — target distribution.
$\lambda(x|y) \sim \mathcal{N}(y{,}\sigma^2)$, $\sigma^2 \ll$ proposal
$x_0 = z$ – very large
$\|S_z P^n - \pi\|_{TV} \leq c \cdot \rho^n \cdot v(z)$

7.5 i-SIR algorithm

(iterated sequential importance resampling)
We want to generate from π, have access to samples from λ, where $\pi(x), \lambda(x) > 0, \forall x \in R^d$ — densities with respect to Lebesgue measure.
$\pi(x) = \frac{\widetilde{\pi}(x)}{\int \widetilde{\pi}(y)dy}$, $\widetilde{\pi}$ - known, $\int \widetilde{\pi}(y)dy$ - unknown.
On step k:

- X_k — current observation
- Generate $N-1$ i.i.d. observations from λ: $y_1^k = x_k, y_2^k, ..., y_N^k \sim$ i.i.d. from λ.
- Compute $w_i^k := \frac{\frac{\widetilde{\pi}(y_i^k)}{\lambda(y_i^k)}}{\sum\limits_{j=1}^{N} \frac{\pi(y_j^k)}{\lambda(y_j^k)}} = \frac{\omega(y_i^k)}{\sum\limits_{j=1}^{N}(y_j^k)}$; $\omega(x) := \frac{\pi(x)}{\lambda(x)}$
- Choose $I_k \leftarrow Catw_i^k$; $X_{k+1} := y_{I_k}^k$
- Re-iterate

8 Lecture 6

8.1 Ergodicity

Definition 35. Markov kernel P is **uniformly geometrically ergodic** if it admits unique invariant distribution π and $||\xi P^n - \pi||_{TV} \leqslant \zeta\rho^n$ for some constant ζ (independent of ξ), $0 < \rho < 1$ and any probability measure ξ.

Definition 36. Markov kernel P is **V-geometrically ergodic** if $\exists V : X \to [1, +\infty)$, such that $\forall x \in X ||\delta_x P^n - \pi||_V \leqslant c\rho^n V(x)$ where c is a constant and $0 < \rho < 1$.

Definition 37. Let μ is signed measure, then $||\mu||_V = \frac{1}{2}\sup_{||f||_V \leqslant 1}\left[\int f(x)\mu(dx)\right]$.

Example. $V(X) \equiv 1 \Rightarrow ||\mu||_V = ||\mu||_{TV} = \frac{1}{2}\sup_{|f|<1}\left[\int f(x)\mu(dx)\right]$.

If $\mu = \xi - \xi'$, ξ, ξ' are probability measures, then $||\mu||_{TV} = \frac{1}{2}\sup\left[\int f(x)\xi(dx) - \int f(x)\xi'(dx)\right] = \sup_{A\in\mathcal{F}} |\xi(A) - \xi'(A)|$.

Example. For $f: X \to \mathbb{R}$ define $||f||_V = \sup_{x \in X} \frac{|f(x)|}{V(x)}$.

Let $f : ||f||_V < \infty, (X_k)_{k=0}^{\infty}$, $\mathrm{Law}(X_0) = \xi, \xi P^n(f) = E_\xi[f(X_n)]$, then

$$|E_\xi[f(X_n)] - \pi(f)| = |\xi P^n(f) - \pi(f)| = \left| \left[\int \xi(dx) \left[\int P^n(x, dx) f(y) \right] \right] - \pi(f) \right| \leqslant$$

$$\leqslant |\text{Jensen}| \leqslant \int_X | \int_X P^n(x{,}dx) f(y) - pi(f)| \xi(dx) \leqslant 2||f||_V c \rho^n \xi(V) \xrightarrow[n\to\infty]{} 0.$$

Remark. $\xi(V)$ can be large.

Definition 38. Drift condition: C is small set (w.r.t. P), then $\int P(x, dx)V(y) = PV(x) \leqslant \lambda V(x) + b\mathbb{I}\{x \in C\}, 0 < \lambda < 1, b$ is a constant.

Let π be invariant distribution of P, then $\pi(V) \leqslant \lambda\pi(V) + b$ and $\pi(V) \leqslant \frac{b}{1-\lambda}$.

8.2 Central Limit Theorem

Let $X_1, \dots, X_n$ are i.i.d., $(\Omega, \mathcal{F}, P)$ is probability space, $0 < \mathrm{Var}X_i < \infty$, then

- $\frac{X_1 + \dots + X_n}{n} \xrightarrow[n\to\infty]{P\text{-a.s.}} E[X_1]$
- **CLT:** $\frac{X_1 + \dots + X_n - nE[X_1]}{\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, \sigma^2)$, $\sigma^2 = \mathrm{Var}X_1$

$\frac{\sum_{i=1}^n f(x_i)}{n} \to \hat{\pi}_n(f)$ to estimate $\pi(f)$, then asymptotic confidence interval from CLT. So we can do MCMC.

$(X_k)_{k=0}^\infty$ is MC with kernel P, π is invariant distribution of P.

$||\xi P^k - \pi||_{TV} \xrightarrow[k\to\infty]{} 0$ ($\mathrm{Law}(X_k) \xrightarrow[k\to\infty]{} \pi$ in TV norm).

$\frac{1}{\sqrt{n}} \left(\sum_{k=0}^{n-1} \{f(X_k) - \pi(f)\} \right) \xrightarrow{d} \mathcal{N}(0, \sigma^2_{\text{asympt}})$ where $\sigma^2_{\text{asympt}} \neq \mathrm{Var}[f(X_1)]$.

σ^2_{asympt} should not depend on ξ!

Example. $(X_k)_{k=0}^\infty$ is MC with kernel P, π is invariant distribution of P, f is bounded function, then

$$S_n = \frac{1}{\sqrt{n}} \left(\sum_{k=0}^{n-1} \{f(X_k) - \pi(f)\} \right)$$

$$\mathrm{Var}_\pi[S_n] = \frac{1}{n} \left[\mathrm{Var}_\pi \left(\sum_{k=0}^{n-1} \{f(X_k) - \pi(f)\} \right) \right] =$$

$$= \frac{1}{n} \left[n\mathrm{Var}_\pi[f] + \sum_{l=1}^{n-1} 2(n-l)\rho^{(l)}(f) \right] = \mathrm{Var}_\pi[f] + 2\sum_{l=1}^{n-1} \frac{n-l}{n} \rho^{(l)}(f)$$

due to $\mathrm{Cov}_\pi(f(X_k), f(X_{k+l})) = E_\pi[(f(X_k) - \pi(f))(f(X_{k+l}) - \pi(f)) = E_\pi[(f(X_0) - \pi(f))(f(X_l) - \pi(f))] = \rho^{(l)}(f)$.

(1) Suppose that $\sum_{l=1}^\infty |\rho^{(l)}(f)| < \infty$.

Under **(1)**:

$$\mathrm{Var}_\pi(S_n) \xrightarrow[n\to\infty]{} \mathrm{Var}_\pi(f) + 2\sum_{l=1}^\infty \rho^{(l)}(f)$$

$$\sigma^2_{\text{asympt}} = \mathrm{Var}_\pi(f) + 2\sum_{l=1}^\infty \rho^{(l)}(f)$$

$$\left|\sum_{l=1}^{n-1}\left(1-\frac{l}{n}\right)\rho^{(l)}(f)-\sum_{l=1}^{\infty}\rho^{(l)}(f)\right|\leqslant\left|\sum_{l=1}^{\infty}\rho^{(l)}(f)\right|+\left|\sum_{l=1}^{n}\frac{l}{n}\rho^{(l)}(f)\right|\leqslant$$

$$\leqslant\left|\sum_{l=1}^{\infty}\rho^{(l)}(f)\right|+\left|\sum_{l=1}^{\lceil 4\sqrt{n}\rceil}\frac{l}{n}\rho^{(l)}(f)\right|+\left|\sum_{l=\lceil 4\sqrt{n}\rceil+1}^{n}\frac{l}{n}\rho^{(l)}(f)\right|\xrightarrow[n\to\infty]{}\frac{n^{1/4}}{n^{3/4}}||f||_{\infty}^{2}$$

due to $\left|\sum_{l=1}^{\infty}\rho^{(l)}(f)\right|\xrightarrow[n\to\infty]{}0$, $\left|\sum_{l=\lceil 4\sqrt{n}\rceil+1}^{n}\frac{l}{n}\rho^{(l)}(f)\right|\xrightarrow[n\to\infty]{}0$.

How to verify that $\sum_{l=1}^{\infty}|\rho^{(l)}(f)|<\infty$?

Example. Let $(X_k)_{k=0}^{\infty}$ be a UGE chain, f is bounded ($\forall\xi: ||\xi P^n-\pi||_{TV}\leqslant\zeta\rho^n$).

$$|\rho^{(l)}(f)|=\left|E_\pi\Big[\{f(X_0)-\pi(f)\}\{f(X_l)-\pi(f)\}\Big]\right|=\left|E_\pi\Big[f(X_0)\{f(X_l)-\pi(f)\}\Big]\right|=$$

$$=\left|\int f(y)\left\{\int P^l(y,dz)f(z)-\pi(f)\right\}\pi(dy)\right|$$

due to $E_\pi[\pi(f)\{f(X_l)-\pi(f)\}]=0$.

$$E_\pi\Big[f(X_0)\{f(X_l)-\pi(f)\}\Big]=E_\pi\Big[E\Big[f(X_0)\{f(X_l)-\pi(f)\}|X_0\Big]\Big]=E_\pi\Big[f(X_0)E\Big[f(X_l)-\pi(f)|X_0\Big]\Big]=$$

$$=\int f(y)E\Big[f(X_l)-\pi(f)|X_0=y\Big]\pi(dy)=E\Big[f(X_l)|X_0=y\Big]=P^lf(y)=\int P^l(y,dz)f(dz).$$

Therefore, we have

$$|\rho^{(l)}(f)|\leqslant\int|f(y)|\left|\int P^l(y,dz)f(z)-\pi(f)\right|\pi(dy)\leqslant 2\int|f(y)|\cdot||\delta_yP^l-\pi||_{TV}\cdot||f||_\infty\pi(dy)\leqslant$$

$$\leqslant 2\int|f(y)|\zeta\rho^l||f||_\infty\pi(dy)\leqslant 2\zeta\rho^l||f||_\infty^2.$$

Remember that $||\mu||_{TV}=\frac{1}{2}\sup_{|f|\leqslant 1}\int f(x)\mu(dx)$.

Hence, under UGE we have

$$\sum_{l=1}^{\infty}|\rho^{(l)}(f)|<\infty.$$

8.3 Martingales

Definition 39. $(X_k)_{k=0}^{\infty}$ on $(\Omega,\mathcal{F},P)$, $(\mathcal{F}_k)_{k=1}^{\infty}$ is a filtration, $\forall i:\mathcal{F}_i\subseteq\mathcal{F}_{i+1}$, $\mathcal{F}_i\subseteq\mathcal{F}$ and $\forall k: E|X_k|<\infty$. Then $\{X_k\}$ is called **a martingale** if $E[X_n|\mathcal{F}_{n-1}]=X_{n-1}$, P-a.s.

Remark. Usually $\mathcal{F}_k=\sigma(X_0,\dots,X_k)$.

Example. Let $X_0,\dots,X_n$ be i.i.d., $E|X_i|<\infty$, $S_n=\sum_{i=0}^{n}\{X_i-\mu\}$, $\mathcal{F}_k=\sigma(X_0,\dots,X_k)$ where $\mu=EX_1$, then $\{S_n\}_{n=0}^{\infty}$ is a matringale. Why?

$$E[S_n|\mathcal{F}_{n-1}]=E[\sum_{i=0}^{n-1}\{X_i-\mu\}+X_n-\mu|\mathcal{F}_{n-1}]=S_{n-1}+E[X_n-\mu|\mathcal{F}_{n-1}]=S_{n-1}$$

due to $E[X_n-\mu|\mathcal{F}_{n-1}]=E[X_n-\mu]=0$.

Definition 40. Given a Markov kernel P with invariant distribution π and a bounded function f, a function $\hat{h}$ is called **a solution to Poisson equation** if:

(2) $\forall x \in X : \hat{h}(x) - P\hat{h}(x) = f(x) - \pi(f)$ where $P\hat{h}(x) = \int P(x, dy)\hat{h}(y)$.

Theorem 6. *Let P satisfy UGE and $||f||_\infty < \infty$. Then:*

$$\hat{h}(x) = \sum_{k=0}^{\infty} \left\{P^k f(x) - \pi(f)\right\}$$

*will be solution to Poisson equation **(2)**.*

Proof.

$$\forall x : |\hat{h}(x)| \leqslant \sum_{k=0}^{\infty} \left|P^k f(x) - \pi(f)\right| \leqslant 2||f||_\infty \sum_{k=0}^{\infty} \left||S_k P^k - \pi\right||_{TV} \leqslant \frac{2||f||_\infty \zeta}{1-\rho}.$$

$$\hat{h}(x) - P\hat{h}(x) = \sum_{k=0}^{\infty} \left\{P^k f(x) - \pi(f)\right\} - P\left\{\sum_{k=0}^{\infty} P^k f - \pi(f)\right\}(x) =$$

$$= \sum_{k=0}^{\infty} \left\{P^k f(x) - \pi(f)\right\} - \sum_{k=1}^{\infty} \left\{P^k f(x) - \pi(f)\right\} = f(x) - \pi(f).$$

□

Theorem 7. *Let P satisfy UGE and let f be bounded $||f||_\infty < \infty$. Then:*

$$\frac{1}{\sqrt{n}} \sum_{k=0}^{n-1} \{f(X_k) - \pi(f)\} \xrightarrow{P_\pi} \mathcal{N}\left(0, \sigma^2(f)\right).$$

Remark. Convergence in P_π means convergence in distribution under $\mathrm{Law}(X_0) = \pi$.

Proof. Let $\hat{h}(x)$ be a solution of **(2)**:

$$\hat{h}(x) = \sum_{k=0}^{\infty} \left\{P^k f(x) - \pi(f)\right\}.$$

Then:

$$\frac{1}{\sqrt{n}} \sum_{k=0}^{n-1} \{f(X_k) - \pi(f)\} = \frac{1}{\sqrt{n}} \sum_{k=0}^{n-1} \left\{\hat{h}(X_k) - P\hat{h}(X_k)\right\} =$$

$$= \frac{1}{\sqrt{n}} \sum_{k=1}^{n-1} \left\{\hat{h}(X_k) - P\hat{h}(X_{k-1})\right\} + \frac{1}{\sqrt{n}}\left(\hat{h}(X_0) - P\hat{h}(X_{n-1})\right)$$

due to $E[\hat{h}(X_k) - P\hat{h}(X_{k-1}) | X_{k-1}] = P\hat{h}(X_{k-1}) - P\hat{h}(X_{k-1}) = 0$.

Let $R_1 = \frac{1}{\sqrt{n}}\left(\hat{h}(X_0) - P\hat{h}(X_{n-1})\right)$. $R_1 \to 0$ in probability:

$$P(|R_1| > \varepsilon) \leqslant \frac{1}{\sqrt{n}\varepsilon} E\left[\left|\hat{h}(X_0)\right| + \left|P\hat{h}(X_{n-1})\right|\right] \leqslant \frac{4\xi ||f||_\infty^2}{\sqrt{n}\varepsilon(1-\rho)} \xrightarrow[n\to\infty]{} 0 \quad \forall \text{ fixed } \varepsilon > 0.$$

$$S_l = \sum_{k=1}^{l-1} \left\{\hat{h}(X_{k-} P\hat{h}(X_{k-1})\right\} \Rightarrow E\left[S_l \middle| \mathcal{F}_{l-1}\right] = S_{l-1}$$

where $\mathcal{F}_{l-1} = \sigma(X_0, \ldots, X_{l-1})$.

$$E\left[S_l \middle| \mathcal{F}_{l-1}\right] = S_{l-1} + E\left[\hat{h}(X_l) - P\hat{h}(X_{l-1}) \middle| \mathcal{F}_{l-1}\right] = S_{l-1} + E\left[\hat{h}(X_l) - P\hat{h}(X_{l-1}) \middle| X_{l-1}\right] = S_{l-1}.$$

□

Lemma. *Let* $(Z_n, \mathcal{F}_n)$: $E\big[Z_n \big| \mathcal{F}_{n-1}\big] = 0$, $\forall n : E|Z_n|^2 < \infty$, $Z_k = \hat{h}(X_k - P\hat{h}(X_{k-1}))$.

- $\frac{1}{n}\sum_{j=1}^{n} E\big[Z_j^2 \big| \mathcal{F}_{j-1}\big] \xrightarrow{P} \sigma^2$
- $\frac{1}{n}\sum_{k=1}^{n} E\big[Z_k^2 \mathbb{I}\{|Z_k| > \varepsilon\sqrt{n}\}\big] \xrightarrow{P} 0$

Then:

$$\frac{1}{\sqrt{n}}\sum_{k=0}^{n-1} Z_k \xrightarrow{d} \mathcal{N}(0, \sigma^2).$$

Proof. (Cor. E.4.2. in Markov Chains, Douc et. al.)

$$\textbf{(3)}\ \frac{1}{n}\sum_{k=1}^{n} E\big[(\hat{h}(X_k) - P\hat{h}(X_{k-1}))^2 \big| \mathcal{F}_{k-1}\big] \xrightarrow{P} ?$$

$$E\big[(\hat{h}(X_k) - P\hat{h}(X_{k-1}))^2 \big| \mathcal{F}_{k-1}\big] = E\big[(\hat{h}(X_k) - P\hat{h}(X_{k-1}))^2 | X_{k-1}\big] = \tilde{h}(X_{k-1}).$$

$$|\hat{h}(X_k) - P\hat{h}(X_{k-1})| \leqslant 2||\hat{h}||_\infty \Rightarrow \tilde{h} \text{ can be chosen to be bounded.}$$

$$\textbf{(3)} = \frac{1}{n}\sum_{k=1}^{n} \tilde{h}(X_{k-1}) = \pi(\tilde{h}) + \frac{1}{n}\sum_{k=1}^{n}\big\{\tilde{h}(X_{k-1}) - \pi(\tilde{h})\big\}.$$

$$\frac{1}{n}\sum_{k=1}^{n}\big\{\tilde{h}(X_{k-1}) - \pi(\tilde{h})\big\} = \mathrm{Var}_\pi\left(\frac{1}{n}\sum_{k=1}^{n}\big\{\tilde{h}(X_{k-1}) - \pi(\tilde{h})\big\}\right) \xrightarrow[n\to\infty]{} 0.$$

$$\mathrm{Var}_\pi\left[\frac{1}{n}\sum_{k=1}^{n}\big\{\tilde{h}(X_{k-1}) - \pi(1)\big\}\right] = \frac{1}{n}\big[\sigma^2_{\text{asympt}}(\tilde{h}) + o(1)\big] \xrightarrow[n\to\infty]{} 0.$$

We have CLT with variance given by $\pi(\tilde{h})$: $E_\pi\big[\tilde{h}(X_l)\big] = E_\pi\big[\big(\hat{h}(X_1) - P\hat{h}(X_0)\big)^2\big] = E_\pi\big[\hat{h}^2(X_1) + \big(P\hat{h}(X_0)\big)^2 - 2\hat{h}(X_1)Ph(X_0)\big]$.

□

Example.

$$E_\pi\big[\hat{h}(X_1)P\hat{h}(X_0)\big] = E_\pi\Big[\big(P\hat{h}(X_0)\big)^2\Big] = E_\pi\big[\hat{h}^2(X_0)\big] - E_\pi\Big[\big(P\hat{h}(X_0)\big)^2\Big] =$$

$$= E_\pi\Big[\big(f(X_0) - \pi(f)\big)\big[\hat{h}(X_0) + P\hat{h}(X_0)\big]\Big] = \mathrm{Var}_\pi[f] + 2\sum_{l=1}^{\infty} E_\pi\Big[\big(f(X_0) - \pi(f)\big)\big(f(X_l) - \pi(f)\big)\Big].$$

9 Lecture 7

9.1 CLT for arbitrary initial distribution

Theorem 8. *Assume that* $||\xi P^n - \xi' P^n||_{TV} \to 0$ *as* $n \to \infty$ *for arbitrary probability measures* ξ, ξ' *and*

$$\frac{1}{\sqrt{n}}\sum_{i=0}^{n-1} h(X_k) \xrightarrow{P_\xi} \mu.$$

Then, if h *is bounded,*

$$\frac{1}{\sqrt{n}}\sum_{i=0}^{n-1} h(X_k) \xrightarrow{P'_\xi} \mu.$$

Remark. If $\exists\pi$ — invariant distribution for P such that $\forall\xi$ $||\xi P^n - \pi||_{TV} \to 0$, then

$$||\xi P^n - \xi' P^n||_{TV} = ||\xi P^n - \pi + \pi - \xi' P^n||_{TV} \leq ||\xi P^n - \pi||_{TV} + ||\pi - \xi' P^n||_{TV} \to 0.$$

Proof. Let us prove using convergence of characteristic functions. Take

$$\varphi_{\xi,n}(t) = E_\xi \left[\exp\left(it\sum_{k=0}^{n-1} h(X_k)/\sqrt{n}\right)\right]$$

and

$$\varphi_\mu(t) = E_{x\sim\mu}[\exp(itx)].$$

Then we have $\varphi_{\xi,n}(t) \to \varphi_\mu(t)$ as $n \to \infty$. Take $\mathcal{F}_m = \sigma(X_0, \dots, X_m)$. Then

$$E\left[E\left[\exp\left(it\sum_{k=0}^{n-1} h(X_k)/\sqrt{n}\right) \Big| \mathcal{F}_m\right]\right] =$$

$$= E\left[\exp\left(it\sum_{k=0}^{m} h(X_k)/\sqrt{n}\right) E\left[\exp\left(it\sum_{k=m+1}^{n-1} h(X_k)/\sqrt{n}\right) \Big| X_m\right]\right] =$$

$$= E\left[\exp\left(it\sum_{k=0}^{m} h(X_k)/\sqrt{n}\right) g(X_m)\right] = E\left[g(X_m)\right] + E\left[\left(\exp\left(it\sum_{k=0}^{m} h(X_k)/\sqrt{n}\right) - 1\right) g(X_m)\right] =$$

$$= E\left[g(X_m)\right] + R_{m,n}(\xi)$$

Take $m = n^{1/3}$, then

$$\left|\frac{it}{\sqrt{n}}\sum_{k=0}^{m} h(X_k)\right| \leq \frac{t(m+1)||h||_\infty}{\sqrt{n}} \to 0 \text{ as } n \to \infty.$$

Hence, since $|g| \leq 1$ almost surely, $R_{m,n}(\xi)$ goes to 0 as n goes to infinity. Now we do the same for ξ' and get $\varphi_{\xi',n}(t) = E_{\xi'}[g(X_m)] + R_{m,n}(\xi')$. Finally,

$$|\varphi_{\xi,n}(t) - \varphi_{\xi',n}(t)| \leq |E_\xi[g(X_m)] - E_{\xi'}[g(X_m)]| + |R_{m,n}(\xi)| + |R_{m,n}(\xi')| \leq$$

$$\leq 2||\xi P^m - \xi' P^m||_{TV} + |R_{m,n}(\xi)| + |R_{m,n}(\xi')| \to 0 \text{ as } n \to \infty.$$

Thus $\varphi_{\xi',n}(t) \to \varphi_\mu(t)$ as $n \to \infty$. □

9.2 Diffusion process example

Take $X_{k+1} = (1-\gamma)X_k + \sqrt{2\gamma}\xi_{k+1}$, where ξ_k are i.i.d. standard normal variables and $0 < \gamma < 1$.

- It is not hard to see that the invariant distribution π for this chain will be normal with 0 mean and the following variance:

$$\text{var}X_{k+1} = (1-\gamma)^2\text{var}X_k + 2\gamma$$

$$\sigma^2 = (1-\gamma)^2\sigma^2 + 2\gamma$$

$$\sigma^2 = \frac{1}{1-\gamma/2}$$

- The kernel will be

$$P(x, A) = P(X_{k+1} \in A | X_k = x) = \frac{1}{\sqrt{4\pi\gamma}} \int_A \exp\left(-\frac{(y-(1-\gamma)x)^2}{4\gamma}\right) dy$$

- Take $V(x) = 1 + x^2$. Then we can check drift condition:

$$PV(x) = E[X_1^2 + 1 | X_0 = x] = 1 + E\left[((1-\gamma)x + \sqrt{2\gamma}\xi_1)^2\right] = 1 + (1-\gamma)^2 x^2 + 2\gamma =$$

$$= (1 - 2\gamma + \gamma^2)(1 + x^2) + 4\gamma - \gamma^2$$

Thus for $\lambda = (1 - 2\gamma + \gamma^2) < 1$ and $b = 4\gamma - \gamma^2$ we have $PV(x) \leq \lambda V(x) + b$.

- Now we check small set condition for balls of radius R.

$$P(x, A) \geq \varepsilon \nu(A) \text{ for any } x \in \{V(x) \leq R\}$$

How to pick ε and ν?

$$P(x, A) = \frac{1}{\sqrt{4\pi\gamma}} \int_A \exp\left(-\frac{(y-(1-\gamma)x)^2}{4\gamma}\right) dy =$$

$$= \frac{1}{\sqrt{4\pi\gamma}} \exp\left(\frac{-(1-\gamma)^2 x^2}{4\gamma}\right) \int_A \exp\left(-\frac{y^2 - 2(1-\gamma)xy}{4\gamma}\right) dy \geq$$

$$\geq \frac{1}{\sqrt{4\pi\gamma}} \exp\left(\frac{-(1-\gamma)^2 R^2}{4\gamma}\right) \int_A \exp\left(-\frac{y^2 + 2\sqrt{R}(1-\gamma)|y|}{4\gamma}\right) dy$$

- And now we look at the convergence. We have

$$X_{k+m} = (1-\gamma)^m X_k + \sum_{\ell=1}^{m} (1-\gamma)^{m-\ell} \sqrt{2\gamma} \xi_{k+\ell}$$

$$\mathrm{cov}_{\delta_0}(X_k, X_{k+m}) = (1-\gamma)^m \mathrm{Var}_{\delta_0}(X_k)$$

Take $X_0 = 0$. Then

$$X_k = \sum_{\ell=1}^{k} (1-\gamma)^{k-\ell} \sqrt{2\gamma} \xi_\ell \sim \mathcal{N}(0, \sigma_k^2)$$

$$\sigma_k^2 = 2\gamma \sum_{\ell=1}^{k} (1-\gamma)^{2(k-l)} = \frac{2\gamma - 2\gamma(1-\gamma)^{2k}}{1-(1-\gamma)^2}$$

$$\sigma_\infty^2 = 2\gamma \sum_{\ell=1}^{\infty} (1-\gamma)^{2(k-l)} = \frac{2\gamma}{1-(1-\gamma)^2} = \frac{1}{1-\gamma/2}$$

Then, to bound $||\delta_0 P^k - \pi||_{TV}$, we need to understand what is the total variation distance for normal distributions. Here we use google link and see that this distance between $\mathcal{N}(0, \sigma^2)$ and $\mathcal{N}(0, \sigma^2(1+\varepsilon))$ is at most $C|\varepsilon|$ where $C > 0$ is some positive constant. In our case we have $\sigma_k^2/\sigma_\infty^2 = 1 - (1-\gamma)^{2k}$, thus

$$||\delta_0 P^k - \pi||_{TV} \leq C(1-\gamma)^{2k}$$

(maybe I forgot square root somewhere, but seems right to me)

9.3 Witch hat example

For some very small ε take

$$\pi_a(x) = \begin{cases} 1, & \text{if } x \in [a, a+\varepsilon] \\ \varepsilon, & \text{if } x \in [0,1] \setminus [a, a+\varepsilon] \end{cases}$$

Then the normalized density will be $\pi(x) = \frac{1}{2\varepsilon - \varepsilon^2}\pi_a(x)$. Let us apply Metropolis Hastings algorithm with uniform proposal: given x_k, the point y_{k+1} is drawn from $U[0,1]$. Then we have

$$\alpha(x_k, y_{k+1}) = \min\left(1, \frac{\pi(y_{k+1})}{\pi(x_k)}\right)$$

Such chain turns out to be UGE. To prove this, we will prove small set condition

$$P(x, A) = \int_0^1 \alpha(x,y) I\{y \in A\} dy + I\{x \in A\} \int_0^1 (1 - \alpha(x,y)) dy \geq \int_0^1 \alpha(x,y) I\{y \in A\} dy =$$

$$= \int_0^1 \min\left(1, \frac{\pi(y_{k+1})}{\pi(x_k)}\right) I\{y \in A\} dy \geq \varepsilon \nu(A)$$

where $\nu(A)$ is $U[0,1]$, and ε is the same as in the definition of $\pi(x)$.

However, such sampling algorithm will not work well in practice, despite being UGE.

10 Lecture 28 Jan (?8?)

Consider $X_j \overset{\text{i.i.d.}}{\sim} P_{data}, j \in \{1, \ldots, n\}$. Usually P_{data} is unknown, and we have three common problems.

- estimate P_{data}, i.e., get $\hat{P_{data}}$;
- sample from P_{data}, i.e., sample from $\hat{P_{data}}$;
- for $f \in L_2(P_{data})$, $\int f^2(x) P_{data}(x) dx < \infty$, estimate $P_{data} = \int f(x) P_{data}(x) dx$

KDE: Kernel Density Estimation

$\hat{P_{data}}(x) = \frac{1}{n}\sum_{j=1}^n K_h(X_j - x)$, where $K_h(y)$ - kernel.
KDE suffers from curse of dimensionality. It usually works well for up to 3 dimensions.
A Bayesian stats case:
For $X_j \overset{\text{i.i.d.}}{\sim}, \theta \in \Theta \subseteq \mathbb{R}^d$ we have $p(X|\theta), \pi_0(\theta)$ - prior distributions of X and on Θ.
Then the posterior is

$$p(\theta | X_1, \ldots, X_n) = \frac{\prod_{j=1}^n P(X_j|\theta)\pi_0(\theta)}{\int_\Theta \prod_{j=1}^n P(X_j|\theta)\pi_0(\theta) d\theta}$$

For $d \geqslant 4$ we again have curse of dimensionality. Still, we can do this easily up to normalising constant. Hence

$$\mathbb{E}_{\theta \sim p(\cdot|X_1, \ldots, X_n)}[f(\theta)] = \int_\Theta f(y) p(y|X_1, \ldots, X_n) dy,$$

which even for $f(\theta) = \theta$ is going to be quite hard.

Generative Adversarial Nets (GANs)

Generator: $G \in \mathbb{G} : Z \to X, Z \subseteq \mathbb{R}^m, X \subseteq \mathbb{R}^n$, where Z is the latent space, and X are, e.g., images; usually $m \leqslant n$.

We sample in the latent space; some prior, e.g., $\mathcal{N}(0,1)$: $z \sim p_0, G(z)$ - "fake" images.

Discriminator: $D \in \mathbb{D} : X \to [0,1]$.

How do we train it?

$$\textbf{Vanilla GAN:} \mathcal{L}(G,D) = \mathbb{E}_{x \sim p_{data}}\left[log(D(x))\right] + \mathbb{E}_{z \sim p_0}\left[1 - log(D(G(z)))\right] \to \min_{G \in \mathbb{G}} \max_{D \in \mathbb{D}}$$

$$\textbf{Wasserstein GAN:} \mathcal{L}(G,D) = \mathbb{E}_{x \sim p_{data}}\left[log(D(x))\right] + \mathbb{E}_{z \sim p_0}\left[1 - log(D(G(z)))\right] \to \min_{G \in \mathbb{G}} \max_{D \in \mathbb{D}},$$

where $D \in \mathrm{Lip}(1)$; $W_1(\xi, \eta) = \sup_{f \in \mathrm{Lip}(1)} |\int f d\xi - \int d d\eta|$.

Denote the density of new objects as $P_G \sim G(z)$; fix $G \in \mathbb{G}$.

Then if $P_G \approx P_{data}, D^*(x) = \frac{p_{data}(x)}{p_{data}(x) + p_G(x)} \approx \frac{1}{2}$.

For $d^* = logit(D^*)$, $\frac{p_{data}(x)}{p_{data}(x) + p_G(x)} = \frac{1}{1 + P_G/P_{data}} = \frac{1}{1 + \exp - d^*(x)} \Rightarrow p_{data}(x) = p_G(x) \cdot \exp d^*(x)$ is the true density.

In reality we have $D(x) \approx D^*(x), d(x) = logit D(x) \Rightarrow \hat{p}_d ata(x) = p_G(x) e^{d(x)}/Z$, where Z is the unknown normalising constant, i.e., it is not the true density. When $D \approx D^* \Rightarrow \hat{p}_{data} \approx p_{data}$. This is called an Energy Based Model (EBM).

Monte-Carlo Methods

With Monte-Carlo (MC) is is a little bit easier as we assume we can draw from various ("simple-to-draw") distributions well.

Take $X_j \overset{\text{i.i.d.}}{\sim} \pi$, e.g., $\pi = p_{data}, \pi = \hat{p}_{data}$.

Aim: $\pi(f) = \int f \pi(x) dx = \int f \pi(dx)$.

$\hat{\pi}(f) = \frac{1}{N} \sum_{j=0}^{N-1} f(X_j) \overset{\text{a.s.}}{\to} \pi(f)$ as $N \to \infty$ if $\pi(f) < \infty$.

Kolmogorov's SLLN. For a d-dimensional case we can get too many partial sums to calculate. With MC we basically choose the points for integration smarter, where the density is the highest. However for $d >> 1$ MC will not be looking at points of highest density.

Importance Sampling (IS)

$X \subseteq \mathbb{R}^d, \lambda$ - easy to sample distribution (proposal).

$$\pi(f) = \int_X f(x) \pi(x) dx = \int \frac{\pi(x)}{\lambda(x)} \lambda(x) dx = \lambda(fw),$$

where $w(x) = \frac{\pi(x)}{\lambda(x)}$

$$\hat{\pi}(f) = \frac{1}{N} \sum_{j=1}^{N} f(Y_j) w(Y_j), Y_j \overset{\text{i.i.d.}}{\sim} \lambda \lambda(fw) < \infty \Rightarrow \hat{\pi}(f) \overset{\text{a.s.}}{\to} \pi(f) (\text{SLLN})$$

Q: how to choose optimal λ?

$$\mathrm{Var}_\lambda \left[f(Y_0, w(Y_0))\right] = \mathbb{E}_\lambda \left[f^2(Y_0) w^2(Y_0)\right] - \mathbb{E}_\lambda^2 \left[f(Y_0) w(Y_0)\right] = \pi^2(f)$$

$$\mathbb{E}_\lambda \left[f^2 w^2\right] \geqslant |\text{Jensen ineq.}| \geqslant \mathbb{E}_\lambda (f|w)^2 \overset{*}{=} \left(\int |f| \pi(x) dx\right)^2$$

$$\mathrm{Var}_\lambda \hat{\pi}(f) = \frac{1}{N} \mathrm{Var}_\lambda \left[f(Y_0) w(Y_0)\right]$$

$(*)$ we cannot really integrate $\lambda^*(x) = \frac{|f(x)| \pi(x)}{\int |f(x)| \pi(x) dx}$

11 Seminar Jan 28

Generating from π

- Rejection sampling: we want to generate from $\pi(\cdot)$, π is completely known
- Choose $p(x)$ - density, s.t. we can sample from $p(\cdot)$, and

$$\sup_x \frac{\pi(x)}{p(x)} \leqslant M; M < \infty$$

$$\exists \nu : \pi << \nu; p << \nu (\text{abs. constant.})$$
$$\pi(x) \leqslant M \cdot p(x), \forall x \in \mathbb{R}, \text{ const. } M > 1$$

Algorithm:

- sample $y \sim p$
- sample $\mathcal{U}[0,1]$ independent of y
- if $\mathcal{U} \leqslant \frac{\pi(y)}{Mp(y)} \Rightarrow$ accept y; else reject

What should we ensure for the procedure to be correct?
Theorem: Let $\mathcal{F}(x) = P_\pi(x \leqslant t)$ - cdf of density π. Then

$$P\left(y \leqslant y | \mathcal{U} \leqslant \frac{\pi(y)}{Mp(y)}\right) = \mathcal{F}(t).$$

Proof:

$$\frac{P\left(y \leqslant y, \mathcal{U} \leqslant \frac{\pi(y)}{Mp(y)}\right)}{P\left(\mathcal{U} \leqslant \frac{\pi(y)}{Mp(y), y\in\mathbb{R}}\right)} = \frac{\int_{-\infty}^{t}\int_0^{\frac{\pi(x)}{Mp(x)}} p(x)dxdu}{\int_{-\infty}^{+\infty}\int_0^{\frac{\pi(x)}{Mp(x)}} p(x)dxdu} = \frac{\int_{-\infty}^{t} \frac{\pi(x)}{Mp(x)}p(x)dx}{\int_{-\infty}^{+\infty} \frac{\pi(x)}{Mp(x)}p(x)dx}$$

$$= \frac{\int_{-\infty}^{t} \frac{\pi(x)}{M}dx}{\int_{-\infty}^{+\infty} \frac{\pi(x)}{M}dx} = F(t) \blacksquare$$

$$\bullet P(\text{Accept}) = P(\mathcal{U} \leqslant \frac{\pi(y)}{Mp(y)}) = \int_{-\infty}^{+\infty}\int_0^{\frac{\pi(x)}{Mp(x)}} p(x)dxdu = \int_{-\infty}^{+\infty} \frac{\pi(x)}{Mp(x)}p(x)dx = \frac{1}{M}$$

thus $M^* = \sup_x \frac{\pi(x)}{p(x)}$. An upper bound is fine even though you will be rejecting more often than you wish. Underestimating M will yield biased results. • if $\pi(x) = \frac{\tilde{\pi}(x)}{Z}$, Z is unknown $\Rightarrow$ you can try to estimate Z then use rejection sampling.

$$\int_{-\infty}^{+\infty} \frac{\tilde{\pi}(x)}{Z \cdot M \cdot p(x)} p(x)dx = \frac{1}{M} \Rightarrow \int_{-\infty}^{+\infty} \frac{\tilde{\pi}(x)}{p(x)}dx = Z$$

$$\hat{Z}_N = \frac{1}{N}\sum_{i=1}^{N} \frac{\tilde{\pi}(y_i)}{p(y_i)}, \text{ where } y_i \sim p(\cdot);$$

- run rejection sampling for $\pi(x) = \frac{\tilde{\pi}(x)}{\hat{Z}}$.

• if $d \in [2, 10] \Rightarrow$ rejection sampling is fine;

• if $\pi(x)$ - bounded, compact support $\Rightarrow$ sample from $\pi(x)$ using the proposal $\mathcal{U}(K)$ - uniform over K.

Example:

$$y \sim B(\alpha, \beta) \Leftrightarrow \pi_y(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)},\ x \in [0, 1].$$

Example:

$$\pi(x) = \frac{1}{\sqrt{2\pi}} \exp -\frac{x^2}{2}; p(x) = \frac{1}{4} \exp -\frac{|x|}{2}, x \in \mathbb{R}^1$$

$$M = \sup_x \frac{\pi(x)}{p(x)} = \sup_x \left(\frac{4}{\sqrt{2\pi}} \exp -\frac{x^2}{2} + \frac{|x|}{2} \right) = \{x = \frac{1}{2},\ \text{symmetric, take } x \geqslant 0\} = \frac{4}{\sqrt{2\pi} e^{\frac{1}{8}}} = \sqrt{\frac{8}{\pi}} e^{\frac{1}{8}} \approx 1.8$$

$$M \approx 0.56; x \in \mathbb{R}^d \Rightarrow M_d = M^d;\ \text{acceptance probability: } (0.56)^d.$$

Good story but unfortunately for small dimensions only.

Example: when IS is useless

You want to estimate $P(|x| \geqslant 4) = \mathbb{E}\left[\mathbf{1}\{|x| \geqslant 4\}\right], x \sim \mathcal{N}(0, 1)$.
Usual Monte-Carlo estimate: $\mathbb{E}\left[\mathbf{1}\{|x| \geqslant 4\}\right] \leftarrow \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\{|x_i| \geqslant 4\}$.
You will get 0. Great approach in terms of absolute error, bad in terms of relative error.
IS: $g(y) \sim \mathcal{N}(5, 1)$ - proposal; $\mathbb{E}\left[\mathbf{1}\{|x| \geqslant 4\}\right] \leftarrow \frac{1}{N} \sum_{i=1}^{N} \frac{\mathbf{1}\{|x_i| \geqslant 4\} p(y_i)}{g(y_i)}$, where $p(\cdot) \sim \mathcal{N}(0, 1), y_i \sim g(\cdot)$
$\frac{p(x)}{g(x)} \mathbf{1}\{|x| \geqslant 4\}$ is quite small.