Part I: Stochastic Models

Aim: The aim here is to review key results in random processes and present some basic signal processing models. The results in Part 1 are essential in understanding the results to follow.

Outline

1. Basic Definitions (random variables and processes)
2. Linear State Space Models
3. Markov chains and Hidden Markov Models
4. Statistical Inference and Stochastic Simulation.

Role of measure theoretic probability: Required in algorithm analysis. Not required for algorithm synthesis (e.g. this course).

Here we do an engineering version.

Why measure theoretic probability

Define a counting set $C_n = \{1, 2, \ldots, n\}$.

Defn: A set A has cardinality $\text{card}(A) = n$ if there is a one to one map from A to C_n.

Suppose $A = \{1/n; n = 0, 1, \ldots, \}$. Then $\text{card}(A) = \aleph_0 =$ countable infinity. A set with cardinality $\aleph_0$ is called a countable or denumerable set.

Examples of $\aleph_0$ cardinality sets:

1. $\mathbb{Z} = \{\ldots, -1, 0, 1 \ldots\}$: set of integers.
2. $\mathbb{Z}_+ = \{0, 1, \ldots\}$: set of non-negative integers.
3. $\mathbb{Q} = \{a/b,\ a \in \mathbb{Z}, b \in \mathbb{Z}, b \neq 0\}$.

$\mathbb{R}$ denotes the set of real numbers.

The cardinality of $\mathbb{R}$ is denoted c – uncountable infinity.

Examples: Closed interval $[0, 1]$; open interval (a, b) where $a, b \in \mathbb{R}$; unions of open intervals, $R \times R$, etc.

c is much larger than $\aleph_0$. Roughly speaking $\aleph_0/c = 0$!

Powerset: Given a finite set $A = \{a_1, \ldots, a_n\}$ the power set denoted 2^A comprises of all subsets of A and the empty set $\emptyset$. $\text{card}(2^A) = 2^n$.

Example: Given $A = \{1, 2\}$, $2^A = \{\{1\}, \{2\}, \{1, 2\}, \emptyset\}$.

Theorem: $2^{\aleph_0} = c$.

What is 2^c?? Contains several types of bizarre elements. Only want those subsets of 2^c that are closed under complementation and countable unions.

Defn. Sigma Algebra or Sigma Field: Let Ω be a set. A family $\mathcal{F}$ of subsets of Ω is called a sigma algebra if
(a) $\emptyset \in \mathcal{F}$
(b) $A \in \mathcal{F} \implies A' \in \mathcal{F}$. Here $A' = \Omega - A$ (complement).
(c) If countable sequence of sets $A_i \in \mathcal{F}$, $i = 1, 2, \ldots$, then $\cup_{i=1}^{\infty} A_i \in \mathcal{F}$.
Therefore also $\cap_{i=1}^{\infty} A_i \in \mathcal{F}$ from (b) and (c).
Examples: Given a finite set A, then the following are sigma algebras:
(i) $\mathcal{F} = \{A, \emptyset\}$.
(ii) $\mathcal{F} = 2^A$.
(iii) The intersection of two sigma algebras is a sigma algebra.

Given a family of sets $\mathcal{H}$, then $\mathcal{F}$ is the *smallest* σ-algebra containing $\mathcal{H}$, if for any other sigma algebra $\mathcal{G}$ containing $\mathcal{H}$, $\mathcal{F} \subset \mathcal{G}$.

Back to 2^c. Want a sigma algebra within 2^c.

Defn. Borel Algebra/Field: The Borel algebra $\mathcal{B}$ is the smallest σ-algebra containing all intervals $(-\infty, a)$, $a \in \mathbb{R}$.
Clearly $\mathcal{B} \subset 2^c$. What are the elements of $\mathcal{B}$?
(i) Intervals $(-\infty, a] \in \mathcal{B}$ where $a \in \mathbb{R}$ since
$\cap_{n=1}^{\infty}(-\infty, a + \frac{1}{n}) = (\infty, a]$.
(ii) Open intervals $(a, b) \in \mathcal{B}$. $(a, b) = (-\infty, a]' \cap (-\infty, b)$.
(iii) $[a, b], [a, b), (a, b], [b, \infty) \in \mathcal{B}$.
We are now ready to define probabilistic models.

Probability and Random variables

A Probabilistic model $(\Omega, \mathcal{F}, P)$ is used to characterize a probabilistic experiment:

1. Sample space Ω: Set of possible *outcomes*

In high school probability – *discrete* probability – Ω is finite set.

In continuous-probability consider harder case where Ω is uncountable set e.g. $\Omega = \mathbb{R}$ or $\Omega = [0, 1]$.

Examples of finite sample spaces:

(i) In single coin toss $\Omega = \{H, T\}$.

(ii) In 2 coin tosses

$\Omega = \{H, T\} \times \{H, T\} = \{HH, HT, TH, TT\}$ (Here $\times$ denotes Cartesian product).

Examples of infinite sample spaces:

(i) $\Omega = [0, 1]$: `rand` in Matlab generates uniform rv on [0,1].

(ii)$\Omega = \mathbb{R}$: `randn` in Matlab generates Gaussian rv on $(-\infty, \infty)$.

2 Event Space $\mathcal{F}$: Set of events.

If Ω is finite $\mathcal{F}$ is a set containing some or possibly all subsets of Ω.

e.g., 2^{Ω} (powerset of Ω).

The elements of $\mathcal{F}$ are called *events* – each event is a set.

Choice of $\mathcal{F}$ depends on what probabilities one wants to

compute.

For probabilistic model to be well defined, $\mathcal{F}$ has to satisfy the following conditions:

(a) $\Omega \in \mathcal{F}$

(b) $A \in \mathcal{F}$ implies $A' \in \mathcal{F}$. Here $A' = \Omega - A$.

(c) $A \in \mathcal{F}$, $B \in \mathcal{F}$ implies $A \cup B \in \mathcal{F}$, $A \cap B \in \mathcal{F}$.

(d) For uncountable Ω: If infinite sequence of events $A_i \in \mathcal{F}$, $i = 1, 2, \ldots$, then $\cup_{i=1}^{\infty} A_i \in \mathcal{F}$.

By defn, $\Omega \in \mathcal{F}$. Hence $\emptyset \in \mathcal{F}$ where $\emptyset$ denotes null set.

Notation: $A \cup B$ denotes: event A **or** event B occurs.

$A \cap B$ or simply AB denotes: event A **and** event B occurs.

$\mathcal{F}$ satisfying (a),(b),(d) is called a *sigma algebra*.

Examples:

(i) In single coin toss

$$\mathcal{F} = \{\emptyset, \Omega, \{H\}, \{T\}\}$$

is a valid event space – since it satisfies (a), (b).

(ii) If Ω has N elements, there are a total of 2^N subsets of Ω including $\emptyset$ – called *powerset* of Ω and denoted as 2^{Ω}. The powerset is a valid event space.

(iii) Toss of Dice: $\Omega = \{1, 2, 3, 4, 5, 6\}$. We can choose $\mathcal{F}$ to be powerset $= \{\emptyset, \Omega, \{1\}, \{2\}, \ldots, \{1, 2\}, \ldots\}$ with $2^6 = 64$ events.

Suppose we are only interested whether outcome is *odd* or

even. Then suffices to consider much smaller event space

$$\{\emptyset, \Omega, \{\text{even}\}, \{\text{odd}\}\}$$

Remark: $\mathcal{F}$ is a family of sets, It always contains at least 2 elements: $\emptyset$ and Ω.

3. Probability Measure: For any event $A \in \mathcal{F}$, probability of event A is $P(A)$ and satisfies the following axioms of probability:
(i) $0 \leq P(A) \leq 1$
(ii) $P(A \cup B) = P(A) + P(B)$ if $A \cap B = \emptyset$, i.e.. A, B are *mutually exclusive* events.
(iii) $P(\Omega) = 1$ (certain event), $P(\emptyset) = 0$ (impossible event).

Remarks: 1. P is a function that maps any set $A \in \mathcal{F}$ to $[0, 1]$. Roughly speaking it measures how big the set A is: $A \subset B \implies P(A) < P(B)$, $P(\Omega) = 1$, $P(\emptyset) = 0$.
2. Probability is defined for *events* not *outcomes*.
3. Repeated application of axiom (ii) for mutually exclusive events $A_1, A_2, \ldots$ yields:

$$P(\cup_{i=1}^{n} A_i) = P(A_1 \cup A_2 \cup \cdots \cup A_n) = \sum_{i=1}^{n} P(A_i)$$

Actually, to deal with infinitely many sets, need extra axiom

of infinite additivity:

$$P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$$

for mutually exclusive sets A_i (this is advanced math).

Using axioms (i), (ii), (iii) one can compute probability of any event.

Independent Events: $A \in \mathcal{F}$ and $B \in \mathcal{F}$ are statistically independent if $\boxed{P(A \cap B) = P(A)P(B)}$.

In communication systems, noise is always assumed to be independent of signal.

Q1: Are mutually exclusive events necessarily independent?
Q2: Suppose A, B are independent events, B, C are independent events. Are A, C always independent?

Terminology and Remarks:

1. *Trial*: single performance of probabilistic expt. Results in an outcome $\zeta \in \Omega$.

2. *Occurrence of an Event*: An event $A \in \mathcal{F}$ occurs during a trial if set A contains the outcome ζ. (Depending on the defn of $\mathcal{F}$, single trial yields single outcome which can result in multiple events occuring).

3. Certain event Ω occurs in every trial, impossible event never occurs.

4. *Complement Event*: $\bar{A} = \Omega - A$.

5. In each trial if event $A \in \mathcal{F}$ occurs then $\bar{A}$ does not occur.

6. If A and B are mutually exclusive events, if A occurs then B does not occur. (e.g, A and $\bar{A}$).

7. If $A \subset B$, and A occurs, then B occurs.

8. *Elementary Events*: If $\Omega = \{a, b, c\}$ then $\{a\}, \{b\}, \{c\}$ are called elementary events.
Result: If Ω finite, probability of any event in $\mathcal{F}$ can be computed given probabilities of all elementary events.
Basis of introductory probability.

9. When Ω is infinite set event space $\mathcal{F}$ is more difficult to construct. No such thing as elementary events.

Exercise: Show that if A and B are indpt, then $\bar{A}$ and B are indpt. Also show that $\bar{A}$ and $\bar{B}$ are indpt.

Random Variables and Random Processes

Defn: A rv $X(w)$ assigns to each outcome $w \in \Omega$ a real number $X(\omega) \in \mathbb{R}$ so that
(i) The set $X(w) \leq x$ is an event for all $x \in \mathbb{R}$.
(ii) $P(X(w) = \infty) = P(X(w) = -\infty) = 0$.

More precisely: Suppose $(\Omega_1, \mathcal{F}_1)$, $(\Omega_2, \mathcal{F}_2)$ are two probability spaces. Let $f : \Omega_1 \to \Omega_2$. f is called a $(\mathcal{F}_1, \mathcal{F}_2)$ *measurable* function if for every $A \in F_2$,
$\{w : f(w) \in A\} \in \mathcal{F}_1$.

A rv X is a measurable function from $(\Omega, \mathcal{F})$ to $(\mathbb{R}, \mathcal{B})$.

Discrete-time **stochastic process** (random process): family of vector random variables $X(\omega, k)$ indexed with discrete-time $k = 0, 1, 2, \ldots$.
(i) for fixed outcome $\omega = \omega_0$, $X(\omega_0, t)$ is deterministic function of time t. Called a "realization" or "sample path" of X.
(ii) for fixed time $t = t_0$, $X(\omega, t_0)$ is a rv
(iii) for fixed $t = t_0$ and $\omega = \omega_0$, $X(\omega_0, t_0)$ is a constant.
(Similar for discrete-time).

State Space Models

$$x_{k+1} = A_k(x_k) + \Gamma_k(x_k)w_k, \quad x_0 \sim \pi_0(\cdot)$$
$$y_k = C_k(x_k) + D_k(x_k)v_k, \quad w_k \sim p_{w_k}, \quad v_k \sim p_{v_k}.$$

Transition Density form:

1. *State*: The state process $\{x_k\}$ is a Markov process

$$p(x_{k+1}|x_k) = p(x_{k+1}|x_k, x_{k-1}, \ldots, x_0)$$

where x_0 has density π_0, i.e., $p(x_0) = \pi_0$.

2. *Observations*: Given the state x_k, y_k is conditionally independent of the past. Thus

$$p(y_k|x_k) = p(y_k|x_k, x_{k-1}, \ldots, x_0, y_{k-1}, \ldots, y_1).$$

$p(y_k|x_k)$ is called the *observation likelihood*.

$$p(x_{k+1}|x_k) = p_w\left(\Gamma_k^{-1}(x_k)\,[x_{k+1} - A_k(x_k)]\right)$$
$$p(y_k|x_k) = p_v\left(D_k^{-1}(x_k)\,[y_k - C_k(x_k)]\right).$$

Because

$$\mathbb{P}(y_k \le y|x_k = x) = \mathbb{P}(C_k(x) + D_k(x)v_k \le y)$$
$$= \mathbb{P}(v_k \le D_k^{-1}(x)(y - C_k(x)))$$

Optimal Predictor: Chapman Kolmogorov Equation

Recursion for the density of the state at each time k. Let $\pi_k(x) = p(x_k) =$ probability density of the state at time k.

$$\boxed{\pi_k(x) = \int_{\mathcal{X}} p(x_k = x | x_{k-1}) \pi_{k-1}(x_{k-1}) dx_{k-1}}, \quad \text{initialized by } \pi_0.$$

$\pi_k(x)$ is predicted pdf at time k given the initial probability π_0 and no observations. Predict mean, mean square and covariance of the state at any future date k as

$$\hat{x}_k = E\{x_k\} = \int_{\mathcal{X}} x\, \pi_k(x) dx,$$

$$\widehat{x^2}_k = E\{x_k^2\} = \int_{\mathcal{X}} x^2 \pi_k(x) dx,$$

$$\text{cov}(x_k) = \mathbb{E}\{(x_k - \hat{x}_k)(x_k - \hat{x}_k)'\} = \widehat{x_k{}^2} - (\hat{x}_k)^2$$

Examples of Stochastic State Space Models

- Linear Gaussian State Space Model
- Hidden Markov Model
- Jump Markov Linear System

1. Linear Gaussian State Space

$$x_{k+1} = A_k x_k + w_k, \quad x_0 \sim \pi_0 = \boldsymbol{N}(\hat{x}_0, \Sigma_0), \quad w_k \sim \boldsymbol{N}(0, Q_k)$$
$$y_k = C_k x_k + v_k, \quad v_k \sim \boldsymbol{N}(0, R_k).$$

$$p(x_{k+1}|x_k) = p_w\left(x_{k+1} - A_k(x_k)\right) = \boldsymbol{N}(x_{k+1}; A_k x_k, Q_k)$$

$$= (2\pi)^{-X/2}|Q_k|^{-1/2} \exp\left[-\frac{1}{2}(x_{k+1} - A_k x_k)' Q_k^{-1}(x_{k+1} - A_k x_k)\right]$$

$$p(y_k|x_k) = p_v\left(y_k - C_k(x_k)\right) = \boldsymbol{N}(y_k; C_k x_k, R_k)$$

Here $|\cdot|$ denotes the determinant of a matrix.

Evolution of Mean and Covariance: Denote

$$\hat{x}_k = \mathbb{E}\{x_k\}, \quad \Sigma_k = \mathbb{E}\{\ (x_k - \hat{x}_k)(x_k - \hat{x}_k)'\}.$$

$$\boxed{\hat{x}_{k+1} = A_k \hat{x}_k.}$$

$$(x_{k+1} - \hat{x}_{k+1})(x_{k+1} - \hat{x}_{k+1})' =$$

$$A_k(x_k - \hat{x}_k)(x_k - \hat{x}_k)' A_k' + w_k w_k' + A_k(x_k - \hat{x}_k) w_k' + w_k (x_k - \hat{x}_k)' A_k'$$

Since w_k is zero mean and statistically independent of x_k:

$$\boxed{\Sigma_{k+1} = A_k \Sigma_k A_k' + Q_k.}$$ "Lyapunov equation"

Summary: $\hat{x}_k$ is optimal state predictor Σ_k is the accuracy (covariance) of the predicted mean.

Covariance of LTI System

$$x_{k+1} = Ax_k + w_k, \quad x_0 \sim \pi_0, \quad \mathbb{E}\{w_k\} = 0,\ \mathrm{cov}\{w_k\} = Q$$

Assume A is stable.

How do mean and covariance evolve as time $k \to \infty$?

Mean: $\hat{x}_k \to \mathbf{0}$ as $k \to \infty$.

Covariance. Algebraic Lyapunov eqn: $\boxed{\Sigma_\infty = A\Sigma_\infty A' + Q}$.

So asymptotically a weakly stationary system.

How to solve? Solve linear system

$$(I - A \otimes A)\,\mathrm{vec}(\Sigma_\infty) = \mathrm{vec}(Q)$$

Note: $\mathrm{vec}(AXB) = (B' \otimes A)\,\mathrm{vec}(X)$.

Eigenvalues of $(I - A \otimes A)$ are $1 - \lambda_i\lambda_j$. Linear system is solvable if $1 - \lambda_i\lambda_j \neq 0 \quad \forall i, j$. Sufficient condition is $|\lambda_i| < 1$.

Existence of Unique Positive Definite Solution:

Theorem (Theorem 2.1,Chapter 4, Anderson & Moore 1979) Suppose A is stable, $[A, M]$ is completely reachable (where $MM' = Q$). Then the algebraic Lyapunov equation has a unique, positive definite soln $\Sigma_\infty = \sum_{k=0}^{\infty} A^k Q A'^k$.

Reachability: Given a matrix M, the pair $[A, M]$ is said to be completely reachable if $[M, AM, A^2M, \ldots, A^{X-1}M]$ has rank X. Equivalently, the matrix $\sum_{k=0}^{X} A^k M M' A'^k > 0$.

Remark: Gives sufficient conditions for the covariance Σ_∞ of the infinite horizon predictor to be unique, positive definite, and bounded. Predictor is asymptotically independent of the initial condition Σ_0.

Proof. Positive Definiteness: Denote $\bar{\Sigma} = \sum_{k=0}^{\infty} A^k Q A'^k$. Since A is stable, convergent geometric series implies $\bar{\Sigma}_\infty$ exists and finite.

$$\bar{\Sigma} \geq \sum_{k=0}^{X-1} A^k Q A'^k > 0$$

(reachability assumption). Thus $\bar{\Sigma}$ is positive definite.

Uniqueness: Let $\bar{\Sigma}$ and $\tilde{\Sigma}$ satisfy Lyapunov eqn,

$$\bar{\Sigma} - \tilde{\Sigma} - A(\bar{\Sigma} - \tilde{\Sigma})A' = 0.$$

$$A^{k-1}(\bar{\Sigma} - \tilde{\Sigma})A'^{k-1} - A^k(\bar{\Sigma} - \tilde{\Sigma})A'^k = 0$$

$$\bar{\Sigma} - \tilde{\Sigma} - A^k(\bar{\Sigma} - \tilde{\Sigma})A'^k = 0.$$

Finally since $A^k \to 0$ as $k \to \infty$, it follows that $\bar{\Sigma} = \tilde{\Sigma}$. $\square$

Hidden Markov Model

(i) *Finite-state Markov chain*: $\{x_k\} \in \mathcal{X} = \{e_1, e_2, \ldots, e_X\}$.
e_i: unit vector with 1 in the ith position.

Dynamics: $X \times X$ transition probability matrix P.

$$P_{ij} = \mathbb{P}(x_{k+1} = e_j | x_k = e_i), \quad 0 \le P_{ij} \le 1, \quad \sum_{j=1}^{X} P_{ij} = 1.$$

Initial condition: $x_0 \sim \mathbb{P}(x_0 = e_i) = \pi_0(i)$, $i = 1, \ldots, X$.

Note: $P\mathbf{1} = \mathbf{1}$. (stochastic matrix).

(ii) *Noisy Observations*: $y_k = C'x_k + v_k$, $\{v_k\}$ i.i.d.
X-dim vector C: "drift coefficients" or "state levels".

Transition Density representation of HMM:

$$B_{iy} = p(y_k = y | x_k = e_i) = p_v(y_k - C'e_i).$$

Example 1. Gaussian noise HMM: $v_k \sim \boldsymbol{N}(0, R)$,

$$B_{iy} = p(y_k = y | x_k = e_i) = \frac{1}{\sqrt{2\pi R}} \exp\left[-\frac{(y_k - C'e_i)^2}{2R} \right].$$

Example 2. Discrete observation HMM: $y_k \in \mathcal{Y} = \{1, \ldots, Y\}$.
Pmf $B_{iy} = \mathbb{P}(y_k = y | x_k = e_i)$ (symbol probabilities).
Summary: HMM specified by (P, B, π_0).

Remark: Martingale representation:

$$x_{k+1} = P'x_k + w_k, \quad \mathbb{E}\{w_k | x_0, x_1, \ldots, x_{k-1}\} = \mathbf{0}$$

Remarks:

1. P always has eigenvalue at 1 correspoding to eigenvector $\mathbf{1}$ because $P\mathbf{1} = \mathbf{1}$.

2. P^k is always a stochastic matrix for any integer $k > 0$.

3 We have defined a first-order homogeneous Markov chain.

Chapman Kolmogorov Eqn

Chapman Kolmogorov eqn for Markov chain is

$$\mathbb{P}(x_{k+1} = e_j) = \sum_{i=1}^{X} \mathbb{P}(x_k = e_i) P_{ij}.$$

Define state prob vector at time k as

$\pi_k = \begin{bmatrix} \mathbb{P}(x_k = e_1) & \dots & \mathbb{P}(x_k = e_X) \end{bmatrix}'$.

$$\boxed{\pi_{k+1} = P'\pi_k \quad \text{initialized by } \pi_0}.$$

Remark: State probability vector π_k evolves as LTI system with state matrix P.

How does π_k behave for large k?

To address this we need to study properties of Markov chain

Properties of Markov Chains

A. State Properties:

1. *Recurrent and Transient States*: A state is *recurrent* if it is visited infinitely often. Otherwise state is called *transient*.

Result: (Algebraic condition): State i is recurrent if $\sum_{k=1}^{\infty} P_{ii}^k = \infty$. Otherwise state i is transient.

Proof: Expected # of visits to state j if started in state i is

$$v_{ij} = \sum_{k=0}^{\infty} \mathbb{E}\{\mathbf{1}_{x_k=j} | x_0 = i\} = \sum_{k=0}^{\infty} \mathbb{P}(x_k = j | x_0 = i) = \sum_{k=0}^{\infty} P_{ij}^k$$

Visit matrix $V = \sum_{k=0}^{\infty} P^k$.

---

2. *Periodic and Aperiodic States*: A state i of a Markov chain has period n if

$$\mathbb{P}(x_k = i | x_0 = i) \neq 0 \quad , k = 0, n, 2n, 3n, \ldots$$
$$\mathbb{P}(x_k = i | x_0 = i) = 0 \quad \text{otherwise}$$

If period $n = 1$, the Markov chain is called *aperiodic*.
Example:

$$P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \text{ period} = 2$$

If all states are aperiodic, Markov chain is called aperiodic.

B. Markov Chain Properties:

1. Irreducible: If every state i communicates with every state j in a finite amount of time. So for each i, j, there exists k such that $P_{ij}^k > 0$.

2. Regular (Primitive): If there exists a positive integer k such that $P_{ij}^k > 0$ for all i, j.

Regular $\implies$ irreducible. Example: $P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ is not regular but is irreducible. Classic book [Seneta, 1981]

Result: Any irreducible Markov chain on a finite state space has a unique stationary distribution.

Result: For an irreducible Markov chain on finite state space: All states are recurrent.
(Reason: Key property of irreducible Markov chain on countable space: All states are either transient or recurrent. For finite state case all states cannot be transient. So for **finite** state irreducible Markov chain all states recurrent).

3. *Ergodic Markov chain*: A Markov chain is ergodic if it is irreducible and aperiodic.

Result: Ergodic Markov chain satisfies law of large numbers:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} h' x_k = h' \pi_\infty \text{ w.p.1}$$

where π_∞ denotes stationary distribution of Markov chain.

Stationary Distribution

How does state probability vector π_k behave for large k?

$$\boxed{\pi_\infty = P'\pi_\infty, \quad \mathbf{1}'\pi_\infty = 1.}$$

Stationary distribution π_∞ is normalized right eigenvector of P' corresponding to the unit eigenvalue.

Theorem: Suppose regular transition matrix P. Then:

1. Eigenvalue 1 has algebraic and geometric multiplicity 1.
2. $1 = \lambda_1 > |\lambda_2| \geq |\lambda_3| \geq \cdots$
3. $P'\pi = \pi$ satisfies $\pi \succeq 0$.
4. $P^k = \mathbf{1}\pi'_\infty + O(k^{m_2-1}|\lambda_2|^k)$ where λ_2 is the second largest eigenvalue modulus and m_2 is its algebraic multiplicity.

Statements 1 to 3 known as the *Perron-Frobenius theorem* for a stochastic matrix.
Statement 4: If P regular, then π_k forgets initial condition geometrically fast since

$$\pi_k = {P'}^k\pi_0 = \pi_\infty \mathbf{1}'\pi_0 + O(k^{m_2-1}|\lambda_2|^k)\pi_0 = \pi_\infty + O(k^{m_2-1}|\lambda_2|^k)\pi_0.$$

So k-step ahead predictor of a Markov chain forgets the initial condition geometrically fast in SLEM.

Proof that $|\lambda_i| \leq 1$: Define spectral radius $r(P) = \max_i |\lambda_i|$
Lemma 1: $r(P) \leq \|P\|_\infty$ where $\|P\|_\infty = \max_i \sum_j P_{ij}$
Proof:
$|\lambda|\|x\| = \|\lambda x\| = \|Px\| \leq \|P\|\|x\| \implies |\lambda| \leq \|P\| \quad \forall \lambda.$

Since $\|P\|_\infty = 1$ and P has an eigenvalue at 1. So $r(P) = 1$.

Proof of 3: For positive matrix P, $P'\pi = \pi$ implies
$P'|\pi| = |\pi|$
Proof: $|\pi| = |P'\pi| \leq |P'||\pi| = P'|\pi|$ So $P'|\pi| - |\pi| \geq 0$.
But $P'|\pi| - |\pi| > 0$ is impossible, since it implies
$1'P'|\pi| > 1'|\pi|$, i.e., $1'|\pi| > 1'|\pi|$.

Upper and lower bounds of SLEM are important.

Dobrushin's Ergodicity Coefficient

Dobrushin coefficient: upper bound for SLEM $|\lambda_2|$.
Useful for showing Markov chain forgets its initial condition geometrically fast. (Even applies to non-linear filtering).

Given pmfs α and β variational distance

$$\|\alpha - \beta\|_{\text{TV}} = \frac{1}{2}\|\alpha - \beta\|_1 = \frac{1}{2}\sum_{i \in \mathcal{X}} |\alpha(i) - \beta(i)|$$

For transition matrix P, define the Dobrushin coefficient

$$\rho(P) = \frac{1}{2}\max_{i,j} \sum_{l \in \mathcal{X}} |P_{il} - P_{jl}| = \max_{i,j} \|P'e_i - P'e_j\|_{\text{TV}}.$$

$\rho(P)$ is max variational dist between two rows of P.

Theorem: $\rho(P)$ satisfies the following properties

1. $0 \leq \rho(P) \leq 1$.
2. $\rho(P) = 0$ if and only if $P = \mathbf{1}\pi'_\infty$.
3. $\rho(P) = 1 - \min_{i,j} \sum_{l \in \mathcal{X}} \min\{P_{il}, P_{jl}\}$.
4. $|\lambda_2| \leq \rho(P)$. (see Bremaud's book).
5. $\|P'\pi - P'\bar{\pi}\|_{\text{TV}} \leq \rho(P) \, \|\pi - \bar{\pi}\|_{\text{TV}}$.
6. Sub-multiplicative: $\rho(P_1 P_2) \leq \rho(P_1)\rho(P_2)$.

Property 2: transition matrix for iid process. $\rho(P) = 0$.

Property 3: If $P_{ij} > 0$, then $\rho(P)$ is strictly smaller than 1.

Prop 5, 6 crucial for Markov chain to forget initial condition.

Markov chain is *geometrically ergodic* if $\rho(P)$ is strictly smaller than 1.

Corollary: π_0 and $\bar{\pi}_0$ two initial distributions.
Corresponding state probabilities: π_k and $\bar{\pi}_k$. Then

$$\begin{aligned}\|\pi_k - \bar{\pi}_k\|_{\text{TV}} = \|{P'}^k \pi_0 - {P'}^k \bar{\pi}_0\|_{\text{TV}} &\leq \|\pi_0 - \bar{\pi}_0\|_{\text{TV}} \rho(P^k) \\ &\leq \|\pi_0 - \bar{\pi}_0\|_{\text{TV}} \, \left(\rho(P)\right)^k, \quad k \geq 0.\end{aligned}$$

So for geometrically ergodic Markov chain, $\|\pi_k - \bar{\pi}_k\|_{\text{TV}} \to 0$ geometrically fast. $\square$

Example 1: If $P = \begin{bmatrix} P_{11} & 1-P_{11} \\ 1-P_{22} & P_{22} \end{bmatrix}$, then
$\rho(P) = |1 - P_{11} - P_{22}|$ which coincides with SLEM $|\lambda_2|$.
Example 2: If $P_{ij} \geq \epsilon$ for all i, j, Property 3 implies
$\rho(P) \leq 1 - \epsilon$. (All non-zero transition probabilities is trivially irreducible and aperiodic.)
Example 3. Uniform Doeblin condition: Suppose $\epsilon \in (0, 1]$

$$\boxed{P_{ij} \geq \epsilon \kappa_j \text{ where } \sum_{j \in \mathcal{X}} \kappa_j = 1, \quad \kappa_j \geq 0.}$$

Then $\rho(P) \leq (1 - \epsilon)$. Doeblin condition weaker than $P_{ij} \geq \epsilon$.

Coupling Interpretation: If $X \sim \pi$, $Y \sim \bar{\pi}$, and $X, Y \sim p(x, y)$.
Coupling inequality: $\boxed{\|\pi - \bar{\pi}\|_{\text{TV}} \leq \mathbb{P}(X \neq Y)}$.
Markov chain $X_n \sim P_{ij} = \epsilon \pi_\infty + (1 - \epsilon) \frac{P_{ij} - \epsilon \pi_\infty(j)}{1 - \epsilon}$. Suppose $\tau = \inf\{n : X_n \sim \pi_\infty\}$. Then

$$\|P'^n \pi_0 - \pi_\infty\|_{\text{TV}} \leq \mathbb{P}(X_n \neq Y_n) = \mathbb{P}(n < \tau) = \mathbb{P}(\tau > n) = (1-\epsilon)^n$$

Proof that $\rho(P) \leq 1 - \epsilon$: Using Property 3,
$\rho(P) = 1 - \min_{i,j} \sum_{l \in \mathcal{X}} \min\{P_{il}, P_{jl}\}$. From the Doeblin condition, $\min\{P_{il}, P_{jl}\} \geq \epsilon \kappa_l$. Therefore
$\sum_l \min\{P_{il}, P_{jl}\} \geq \sum_l \epsilon \kappa_l = \epsilon$. Then using Property 3 that $\rho(P) = 1 - \min_{i,j} \sum_{l \in \mathcal{X}} \min\{P_{il}, P_{jl}\}$ yields $\rho(P) \leq 1 - \epsilon$.

Proof of Properties 5 and 6 of Dobrushin Coefficient:
Property 5: Suppose $\pi = e_i$ and $\bar{\pi} = e_j$, $i \neq j$.
LHS: $\|P'e_i - P'e_j\|_{\mathrm{TV}}$.
RHS: Since $\|e_i - e_j\|_{\mathrm{TV}} = 1$,

$$\rho(P)\,\|e_i - e_j\|_{\mathrm{TV}} = \max_{ij} \|P'e_i - P'e_j\|_{\mathrm{TV}}.$$

So for unit vectors, Property 5 holds.

To prove Property 5 for general π, $\bar{\pi}$, express $\pi - \bar{\pi}$ in terms of unit vectors. Define $\phi(i) = \min\{\pi(i), \bar{\pi}(i)\}$. Then

$$\pi - \bar{\pi} = \sum_{i,j} \gamma_{ij}(e_i - e_j), \qquad \text{where } \gamma_{ij} = \frac{(\pi(i) - \phi(i))\,(\pi(j) - \phi(j))}{\|\pi - \bar{\pi}\|_{\mathrm{TV}}}.$$

where $\sum_{i,j} \gamma_{ij} = \|\pi - \bar{\pi}\|_{\mathrm{TV}}$.

Then to prove Property 5:

$$\|P'\pi - P'\bar{\pi}\|_1 = \|\sum_{i,j} \gamma_{ij} P'(e_i - e_j)\|_1 \leq \sum_{i,j} \gamma_{ij} \max_{i,j} \|P'(e_i - e_j)\|_1$$

Using the definition $\rho(P) = \frac{1}{2} \max_{i,j} \|P'(e_i - e_j)\|_1$ and
$\sum_{i,j} \gamma_{ij} = \|\pi - \bar{\pi}\|_{\mathrm{TV}}$,

$$\|P'\pi - P'\bar{\pi}\|_{\mathrm{TV}} \leq \rho(P)\,\|\pi - \bar{\pi}\|_{\mathrm{TV}}.$$

Property 6:

$$\begin{aligned}
\rho(P_1 P_2) &= \frac{1}{2} \max_{ij} \|P_2' P_1'(e_i - e_j)\|_1 \\
&\leq \frac{1}{2} \max_{ij} \rho(P_2) \|P_1'(e_i - e_j)\|_1 \qquad \text{(by Property 5)} \\
&\leq \rho(P_2)\rho(P_1) \qquad \text{(by definition)}
\end{aligned}$$

Statistical Inference For IID processes

statistics (real world) $\leftrightarrows$ probability (mathematical model)

IID process: $\{x_k\}$ is iid if distribution of x_k is indpt of time k and x_k and x_n are independent for $k \neq n$. Note $\mathbb{E}\{x_k\} = \mu$ (indpt of time) for an iid process.

Two major results (i) Law of large numbers (strong convergence)
(ii) Central Limit Theorem (convergence in distribution)

For iid process $\{x_n(\omega)\}$ two averages can be computed:
1. Expected value for fixed n:

$$\mathbb{E}\{x_n(\omega)\} = \int_{\mathcal{X}} x p_x(x) dx = \mu$$

2. Sample path average for fixed ω given N observations:

$$\hat{\mu}_N(\omega) = \frac{1}{N}\left(x_1(\omega) + \ldots + x_N(\omega)\right)$$

Result Kolmogorov Strong Law or Large Numbers: For iid process iff $\mathbb{E}\{|x_n|\} < \infty$ (integrable rv), then

$$P(\omega : \lim_{N\to\infty} \hat{\mu}_N(\omega) = \mu\}) = 1.$$

Comment: If not iid, then SLLN may not hold. Here is a counterexample. Consider random process:

$$x_n = x_{n-1}, \quad x_0 = \begin{cases} 0 & \text{with prob } 0.8 \\ 1 & \text{with prob } 0.2 \end{cases}$$

Then there are only two possible outcomes
$\omega = [0,0,0,0,0,0,\ldots,]$ or $\omega = [1,1,1,1,1\ldots,]$.
So time average $\hat{\mu}_N = 0$ or 1, but $\mu = \mathbb{E}\{x_n\} = 0.2$. Thus $\mu \neq \hat{\mu}_N$, i.e., SLLN does not hold.
Here x_n are identically distributed but not indpt.

Central Limit Theorem: For iid sequence, if $\mathbb{E}\{x_n^2\} < \infty$, then $\sqrt{n}(\hat{\mu}_n - \mu) \to \boldsymbol{N}(0, \sigma^2)$ where $\sigma^2 = \mathrm{Var}(x_n)$.

Note: $\mathbb{E}\{x_n^2\} < \infty$ implies $\mathbb{E}\{|x_n|\} < \infty$ – so SLLN holds.

Markov Chain SLLN: If regular (aperiodic + irreducible):

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} h' x_k = h' \pi_\infty \text{ w.p.1}$$

where π_∞ denotes stationary distribution of Markov chain.

$$CLT: \sqrt{n} \left(\frac{1}{n} \sum_{k=1}^{n} g' x_k - g' \pi_\infty \right) \to \boldsymbol{N}(0, \sigma^2)$$

$$\sigma^2 = \mathrm{Var}_{\pi_\infty}(g' x_0) + 2 \sum_{k=1}^{\infty} \mathrm{cov}_{\pi_\infty}(g' x_0, g' x_k)$$

$$= \boxed{g' \mathrm{diag}(\pi_\infty) \left(2Z - (I + \mathbf{1}\pi_\infty')\right) g; \quad Z = (I - P - \mathbf{1}\pi_\infty')^{-1}}$$

How long to simulate a Markov chain?

$$\phi_n = \frac{1}{n}\sum_{k=1}^{n} h' x_k$$

$$\begin{aligned}
\text{Bias}(\phi_n) &= \mathbb{E}\{\phi_n\} - h'\pi_\infty \\
\text{Var}(\phi_n) &= \mathbb{E}\{\phi_n - \mathbb{E}\{\phi_n\}\}^2 \\
\text{MSD}(\phi_n) &= \mathbb{E}\{\phi_n - h'\pi_\infty\}^2 = \text{Var}(\phi_n) + (\text{Bias}(\phi_n))^2.
\end{aligned}$$

Theorem: n-point sample path of a Markov chain $\{x_k\}$ with regular transition matrix P. Then

$$|\,\text{Bias}(\phi_n)| \le \frac{|\max_i h_i|}{n(1-\rho)}\,\|\pi_0 - \pi_\infty\|_{\text{TV}}$$

$$\text{MSD}(\phi_n) \le \frac{|\max_i h_i|^2}{n(1-\rho)} \sum_{i\in\mathcal{X}} (\|e_i - \pi_\infty\|_{\text{TV}})^2\, \pi_\infty(i) + O(\frac{1}{n^2}).$$

$$\|\pi - \bar{\pi}\|_{\text{TV}} = \frac{1}{2}\|\pi - \bar{\pi}\|_1 = \frac{1}{2}\sum_i |\pi(i) - \bar{\pi}(i)|$$

ρ is Dobrushin coefficient of transition matrix P. □

Bias small if: π_0 close to π_∞ or n is large, or ρ is small.

MSD: also depends on sum of the squares of the variational distance between π_∞ and unit state vectors.

Proof. Recall $\pi_k = \mathbb{E}\{x_k\} = {P'}^k \pi_0$ and $\pi_\infty = {P'}^k \pi_\infty$. Then

$$\begin{aligned}
\left| \mathbb{E}\left\{ \frac{1}{n} \sum_{k=1}^{n} h' x_k \right\} - h'\pi_\infty \right| &= \frac{1}{n} \left| \sum_{k=1}^{n} h'(\pi_k - \pi_\infty) \right| \\
&\leq \frac{1}{n} \sum_{k=1}^{n} \max_i |h_i| \, \|\pi_k - \pi_\infty\|_{\text{TV}} \quad \text{(Holder's inequality)} \\
&= \frac{1}{n} \max_i |h_i| \sum_{k=1}^{n} \|{P'}^k \pi_0 - {P'}^k \pi_\infty\|_{\text{TV}} \\
&\leq \frac{1}{n} \max_i |h_i| \, \|\pi_0 - \pi_\infty\|_{\text{TV}} \sum_{k=1}^{n} \rho^k \\
&\leq \frac{1}{n} \max_i |h_i| \, \|\pi_0 - \pi_\infty\|_{\text{TV}} \, \frac{1}{1-\rho}
\end{aligned}$$

where the last inequality follows because $\rho < 1$ since P is regular. □

Proof of SLLN (iid case)

$\mu_n = \frac{1}{n}\sum_{k=1}^n x_k$. Then convergence in probability says

$$\mathbb{P}(|\mu_n - \mu| > \delta) \to 0, \quad \text{as } n \to \infty$$

SLLN says that $|\mu_n - \mu|$ is larger than δ only a finite number of times (with probability 1).
That is: $\sum_{n=1}^\infty I(|\mu_n - \mu| > \delta)$ converges. That is, there exists a finite n_0, such that for all $n > n_0$, $|\mu_n - \mu| < \delta$.

Defintions: (i) A sequence of events $\{A_n\}$ happens infinitely often (denoted as $\{A_n \text{ i.o.}\}$ if

$$\{w : w \in A_n \text{ for infinite number of } n\} = \{w : w \in \cap_{n=1}^\infty \cup_{k\geq n} A_k\}$$

(ii) $X_n \overset{\text{w.p.1}}{\to} X$ if $P(w : \lim_{n\to\infty} X_n(w) = X(w)) = 1$ (a.s. convergence). Equivalently: if $P(|X_n - X| \geq \epsilon \text{ i.o}) = 0$.
(iii) $X_n \to X$ in probability if $\lim_{n\to\infty} P(|X_n - X| \geq \epsilon) = 0$

Remark: Denote $G_n = \cup_{k\geq n} A_k$. Then G_n occurs implies there is some A_k for $k \geq n$ which occurs. So $\cap_{n=1}^\infty G_n$ occurs implies an infinitely many A_k occur.

Borel-Cantelli Lemma: Sufficient condition for wp1 convergence: $\sum_{n=1}^\infty P(A_n) < \infty \implies P(A_n \text{ i.o}) = 0$.
Converse: $\{A_n\}$ iid & $\sum_n P(A_n) = \infty \implies P(A_n \text{ i.o. } = 1)$.

Example 1: Assume $X_1, X_2, X_3, \ldots$ are iid. Denote

$$U_n = \begin{cases} 1 & \text{if } X_n > X_j, j = 1, \ldots, n-1 \\ 0 & \text{otherwise} \end{cases}$$

U_n denotes record at time n. $P(U_n = 1) = 1/n$ by symmetry.
(i) Prove that infinitely many records happen as $n \to \infty$.
Define $A_n = \{U_n = 1\}$. Clearly A_n are indpt. So $\sum_{n=1}^{\infty} P(A_n) = \sum_n \frac{1}{n} = \infty$. So from converse of Borel Cantelli, $P(A_n \text{ i.o.} = 1)$. So infinitely many records happen.
(ii) Prove that only a finite number of double records happen, i.e. two records in a row.

$$\sum_{n=1}^{\infty} P(A_n \cap A_{n+1}) = \sum_{n=1}^{\infty} P(A_n)P(A_{n+1}) = \sum_n \frac{1}{n(n+1)} < \infty$$

Example 2: $X_1, X_2, \ldots$ are iid. Prove $P(|X_n| > n \text{ i.o.}) = 0$.
Hint: Start with $\mathbb{E}|X| = \int_0^{\infty} P(|X| > x)dx =$
$\sum_{n=0}^{\infty} \int_n^{n+1} P(|X| > x)dx \leq 1 + \sum_{n=1}^{\infty} P(|X_n| > n)$

Baby SLLN: $\{X_n\}$ iid and $\sigma^2 = \text{Var}(X_n)$ is bounded, then $\hat{\mu}_n \to \mu$ w.p.1.
Assume $\mu = 0$. Define $S_n = \sum_{i \leq n} X_i$. Note $\text{Var}(S_n) = n\sigma^2$
Aim: Prove $\hat{\mu}_n = S_n/n \overset{\text{w.p.1}}{\to} 0$.

Recall Chebyshev (Markov inequality) states that for $p \geq 1$,

$$\boxed{P(|X| > \alpha) \leq \frac{\mathbb{E}|X|^p}{\alpha^p}}$$

Initial attempt: Using Chebyshev inequality.

$$P(|S_n| > n\epsilon) \leq \frac{\mathrm{Var}(S_n)}{(n\epsilon)^2} = \frac{n\sigma^2}{(n\epsilon)^2} = \frac{\sigma^2}{n\epsilon^2}$$

Unfortunately $\sum_n 1/n$ is not convergent.

So we consider subsequence $\{S_{n^2}\}$, $n = 1, 2, \ldots$.

$$P(|S_{n^2}| > n^2\epsilon) \leq \frac{\mathrm{Var}(S_{n^2})}{(n^2\epsilon)^2} = \frac{\sigma^2}{n^2\epsilon^2}$$

So from Borel Cantelli, $P(\frac{S_{n^2}}{n^2} > 0 \text{ i.o.}) = 0$, i.e. $\frac{S_{n^2}}{n^2} \overset{\text{w.p.1}}{\to} 0$.
But what happens between S_{n^2} and $S_{(n+1)^2}$? Define

$$D_n = \max_{n^2 \leq k < (n+1)^2} |S_k - S_{n^2}|$$

Prove that $\mathbb{E}\{D_n^2\} \leq 4n^2\sigma^2$. Then Chebyshev implies

$$P(D_n > n^2\epsilon) = \frac{4\sigma^2}{n^2\epsilon^2}$$

This implies $D_n/n^2 \overset{\text{w.p.1}}{\to} 0$. So for k between n^2 and $(n+1)^2$,

$$\left|\frac{S_k}{k}\right| \leq \frac{|S_{n^2}| + D_n}{k} \leq \frac{|S_{n^2}| + D_n}{n^2} \overset{\text{w.p.1}}{\to} 0$$

To relax the assumption of finite variance requires more work. It uses the truncated variable

$$\bar{X}_n = \begin{cases} X_n & \text{if } |X|_n \leq n \\ 0 & \text{otherwise} \end{cases}$$

Details are omitted.

Simulation of RVs

How can one generate a rv with a specified density using a computer?

Why simulation?

1. How do you efficiently solve nonlinear Bayesian filtering problem (particle filters later). To do multidimensional integration efficiently via Monte-Carlo methods. E.g.,

$$\int_{\mathbb{R}^N} f(x)dx = \int \frac{f(x)}{\pi(x)}\pi(x)dx \approx \frac{1}{N}\sum_{i=1}^{N} \frac{f(x_i)}{\pi(x_i)} \text{ where } x_i \sim \pi(x) \ .$$

2. Simulation of discrete-event systems. How do you simulate a queue/buffer with complicated arrival and departure distributions? .

3. Simulation based optimization in stochastic control: Q Learning, Temporal difference methods, etc.

4. Randomized algorithms such as *simulated annealing* are globally optimal.

Simulation, Ross; Casella & Roberts, MCMC

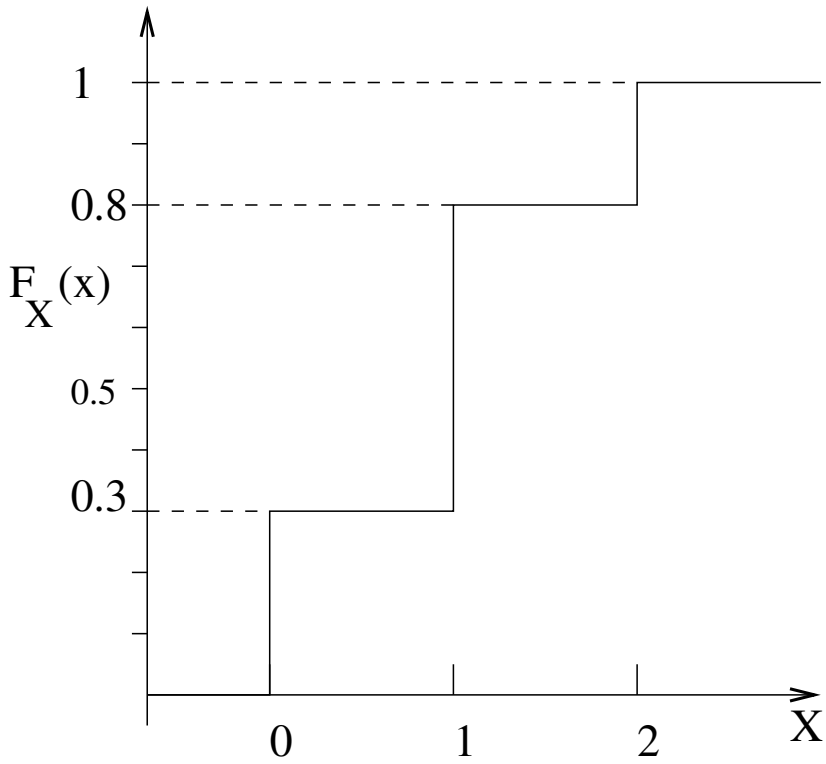
Inverse Transform Method – Discrete rvs

First consider Inverse Transform Method for generating discrete valued rvs.

Example: Given $U[0,1]$ generator, generate a rv X with $P(X=0)=0.3$, $P(X=1)=0.5$, $P(X=2)=0.2$.

Solution: Generate $u \sim U[0,1]$.

$$\text{Set } X = \begin{cases} 0 & \text{if } u < 0.3 \\ 1 & \text{if } 0.3 \le u < 0.8 \\ 2 & \text{otherwise} \end{cases}$$



Examples: Discrete RVs

1. Suppose $X \in [1, 2, 3, 4] \sim [0.20\ 0.15\ 0.25\ 0.4]$.
(requires 3 **if** decisions).
Efficiency.

2. Discrete uniform rv: Suppose $P(X = j) = 1/n$,
$j = 1, \ldots, n$. In this special case no **if** statements required!

$$X = j \text{ if } \frac{j-1}{n} \leq U < \frac{j}{n}$$

i.e., $X = j$, if $j - 1 \leq nU < j$, i.e.

$$X = \text{Int}(nU) + 1$$

Inverse Transform Method – Cont rvs

Result: Suppose $u \sim U[0, 1]$. Then for any continuous distribution function F, the rv X defined by

$$X = F^{-1}(u)$$

has distribution F. ($F^{-1}(u)$ is defined to be that value of x such that $F(x) = u$.).

Example 1: Generate rv with distribution

$$F(x) = x^n, \qquad 0 \le x \le 1$$

Soln: $u = F(x) = x^n$ or equivalently, $x = u^{1/n}$. So generate rv $u \sim U[0,1]$. Then $u^{1/n}$ has distribution $F(x)$.

Example 2: Generate exponentially distributed rv.

$$F(x) = 1 - e^{-x}$$

Soln: $u = F(x) = 1 - e^{-x}$. Thus $x = -\log(1-u)$. So generate rv $u \sim U[0,1]$. Then $-\log(1-u)$ has expo dist. Note Poisson rvs can be generated from expo rv.

Why does the method work?

Why Inverse Transform Method works

Result: Suppose $u \sim U[0,1]$. Then for any continuous distribution function F, the rv X defined by

$$X = F^{-1}(u)$$

has distribution F.

Reason: Let $F_X(x)$ denote distribution of X. Then

$$F_X(x) = P(X \leq x) = P(F^{-1}(u) \leq x)$$

But F is a distribution function. Therefore F is monotone increasing in x.
So $a \leq b \equiv F(a) \leq F(b)$.
Therefore

$$F_x(x) = P\{F(F^{-1}(u) \leq F(x))\} = P(u \leq F(x)) = F(x)$$

since $u \sim U(0,1)$.

Remarks:

1. In inverse transform method you invert the probability distribution function – not the density function.

2. For Gaussian rv, there is no explicit formula for distribution. So cannot use inverse transform method to generate Gaussian.

Advanced Example: Skorohod Representation (e.g. see Kushner & Yin's book or D. Pollard's book on weak convergence)

Result: If $Y_n \sim F_n \to Y \sim F$ weakly on $(\Omega, \mathcal{F}, P)$.

Then there exists $(\bar{\Omega}, \bar{\mathcal{F}}, \bar{P})$ such that

$$\bar{Y}_n \sim F_n \to \bar{Y} \sim F \text{ almost surely}$$

An explicit construction of $\bar{Y}_n$ and $\bar{Y}$ is based on inverse transform method:

Define $\bar{Y}_n = F_n^{-1}(u)$, $\bar{Y} = F^{-1}(u)$, $u \sim U[0,1]$, i.e. $\bar{P}$ is Lebesgue measure on $[0,1]$. Then inverse transform method implies $\bar{Y}_n \sim F_n$ and $\bar{Y} \sim F$.

Also since $F_n \to F$ pointwise monotonically, $F_n^{-1} \to F^{-1}$ pointwise. This means that $\bar{Y}_n \to \bar{Y}$ a.s.

Simulating a Markov chain

For a Markov chain x, given x_k, x_{k+1} is conditionally independent of the past. So the following procedure is obvious: Let P_i denote the i-th row of transition matrix.

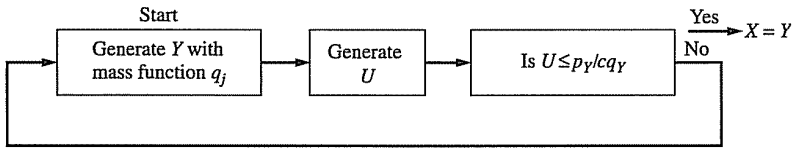
1. Generate $x_0 \sim \pi_0$.
2. For $k = 1, \ldots,$ generate $x_k \sim P_{x_{k-1}}$.

Here Step 1 and 2 can be implemented via inverse transform or Acceptance rejection method.

Acceptance Rejection Method

Discrete rv: Suppose $X \sim q_j$, $j = 1, \ldots, M$ is easy to generate. Aim: Generate $Y \sim p_j$, $j = 1, \ldots, M$.

$$c = \max_j \frac{p_j}{q_j}$$



Step 1: Generate $X \sim q_j$

Step 2: Generate $U \sim U[0,1]$.

Step 3: If $U < \frac{p_X}{cq_X}$, set $Y = X$ and stop. Else goto Step 1.

Example: Want to simulate $Y \in \{1, 2, \ldots, 10\}$ with probs $\{0.11, 0.12, 0.08, 0.12, 0.10, 0.09, 0.10, 0.10\}$.

Generate $q_j = 1/10$, $j = 1, \ldots, 10$.

$$c = \max p_j/q_j = 1.2$$

Step 1: Generate uniform discrete rv (as explained earlier). $U_1 \sim U[0,1]$, $X = \text{Int}(10U_1) + 1$.

Step 2: Generate U

Step 3: If $U < p_X/(cq_X)$, set $Y = X$ and stop. Else goto step 1.

Remarks 1: Average number of iterations $= c$. (Note c can never be less than 1 – why?)

2. Only p_j/c is used (normalization term not required).

3. Bound c need not be tight. e.g. $2c$ will work.

Cont rv: Generate $Y \sim p(Y)$ assuming we can simulate from $X \sim q(X)$. Suppose $p(x)/q(x) \leq c$. Then:

Step 1. Generate $X \sim q(X)$.
Step 2: Generate $U \sim U[0,1]$.
Step 3: If $U < p(X)/(cq(X))$, set $Y = X$.
Otherwise go back to step 1.

Example 1: Generate from a $U[0,1]$ variable, rv from

$$F(x) = x^n, \quad 0 \leq x \leq 1.$$

Soln: $q(x) = U[0,1]$. Then

$$\sup_{\zeta} \frac{p(\zeta)}{q(\zeta)} = \max_{\zeta \in [0,1]} n\zeta^{n-1} = n$$

So choosing $c = n$, Step 1 and Step 2 generate two independent $U[0,1]$ samples y and u. Step 3 sets $x = y$ if $u < y^{n-1}$.

Example 2: Generate a unit normal variable $\boldsymbol{N}(0,1)$ from exponential pdf $q(x) = e^{-x}$, $x \geq 0$.

Soln: First generate rv

$$p(x) = \frac{2}{\sqrt{2\pi}} e^{-x^2/2}, \quad x \geq 0.$$

This corresponds to absolute value of a $\boldsymbol{N}(0,1)$ rv. Note

$$\sup_{\zeta} \frac{p(\zeta)}{q(\zeta)} = \sup_{\zeta \in \mathbb{R}} \sqrt{\frac{2}{\pi}} e^{\zeta - \zeta^2/2} = \sqrt{\frac{2e}{\pi}}.$$

Choosing $c = \sqrt{2e/\pi}$, Step 1 generates an exponentially distributed random variable y (using for example the inverse transform method). Step 2 generates a uniform random variable u. Finally, Step 3 sets $x = y$ if $u \leq e^{-(Y-1)^2/2}$. Finally to generate a $\boldsymbol{N}(0,1)$ random variable, we multiply x by ± 1 where 1 and -1 are chosen with probability 1/2 each.

Proof :$P(Y \leq y) = P(X \leq y | U \leq \dfrac{p(X)}{cq(X)})$

$$= \frac{P\left(X \leq y, U \leq \frac{p(X)}{cq(X)}\right)}{P\left(U \leq \frac{p(X)}{cq(X)}\right)} = \frac{\text{Prob of } X \leq y \text{ and accept}}{\text{Prob of accept}}$$

$$= \frac{\int_{-\infty}^{y} \int_0^{\frac{p(x)}{cq(x)}} du\, q(x) dx}{\int_{-\infty}^{\infty} \int_0^{\frac{p(x)}{cq(x)}} du\, q(x) dx} = \frac{\frac{1}{c}\int_{-\infty}^{y} p(x) dx}{\frac{1}{c}\int_{-\infty}^{\infty} p(x) dx}$$

Each iteration independently yields Prob of accept $= \frac{1}{c}$. Number iterations to accept is geometric RV with mean c. Recall geometric rv models number of trials to first success when each trial is iid with success prob p: So for $n \geq 1$, $P(X = n) = p(1-p)^{n-1}$. So $\mathbb{E}\{X\} = 1/p$, $\text{var}\{X\} = \frac{1-p}{p^2}$

B. D. Flury, *Acceptance-Rejection Sampling Made Easy*, SIAM Review, 1990.

Composition Method

Discrete rv: Suppose easy to generate rvs from $F_i(x)$, $i = 1, 2, \ldots, n$. How to generate rv from

$$F(x) = \sum_{i=1}^{n} p_i F_i(x), \quad p_i \geq 0, \sum_{i=1}^{n} p_i = 1$$

Soln: Step 1: generate a rv $i^* \in \{1, \ldots, n\}$ with probabilities $p_1, \ldots, p_n$.
Step 2: Then generate a rv with distribution F_{i^*}.

Cont rv: Suppose easy to generate rvs from $x \sim f_{X|Y}(x|y)$ and $y \sim f_Y(y)$. How to generate samples from

$$f_X(x) = \int_{-\infty}^{\infty} f_{X|Y}(x|y) f_Y(y) dy$$

Step 1: Simulate $y^* \sim f_Y(y)$.
Step 2: Simulate $x \sim f_{X|Y}(x|y^*)$.

Application: Simulation based optimal predictor.

$$\pi_k(x) = \int_{\mathcal{X}} p(x_k = x|x_{k-1}) \pi_{k-1}(x_{k-1}) dx_{k-1}.$$

How to simulate samples from $\pi_k(x)$? Composition method:

Step 1: Simulate $x_{k-1}^{(l)}$, $l = 1, \ldots, L$ from $\pi_{k-1}(x)$.
Step 2: Simulate $x_k^{(l)} \sim p(x_k|x_{k-1} = x_{k-1}^{(l)})$, $l = 1, 2, \ldots, L$.

Importance Sampling

$$\mathbb{E}_p\{c(x)\} = \int_{\mathcal{X}} c(x)p(x)dx \approx \frac{1}{N}\sum_{k=1}^{N} c(x_k) \quad \text{where } x_k \sim p$$

In importance sampling use a proposal distribution $q(x)$:

$$\mathbb{E}_p\{c(x)\} = \underbrace{\int_{\mathcal{X}} c(x)\frac{p(x)}{q(x)}\, q(x)dx}_{\mathbb{E}_q\{c(x)\frac{p(x)}{q(x)}\}} \approx \underbrace{\frac{1}{N}\sum_{k=1}^{N} c(x_k)\frac{p(x_k)}{q(x_k)}}_{\hat{c}}, \quad x_k \sim q$$

Denote $w(x) = \frac{p(x)}{q(x)}$. Then: $\mathbb{E}_p\{c(x)\} \approx \frac{\sum_{k=1}^{N} c(x_k)w(x_k)}{\sum_{k=1}^{N} w(x_k)}$, $x_k \sim q$.

Result. Optimal choice of $q(x)$ to minimize variance of $\hat{c}$ is $q(x) \propto c(x)p(x)$.

Proof:

Variance Reduction by dissection: Suppose $F(x) = \sum_{i=1}^{n} p_i F_i(x), \quad p_i \geq 0, \sum_{i=1}^{n} p_i = 1$. Then: Single sample $x \sim F(x)$ has larger variance than $\hat{x} = \sum_i p_i x_i$ where $x_i \sim F_i(x)$, $i = 1, 2, \ldots, n$. (weight avg of n samples).

Outline

1. Basic Definitions (random variables and processes)
2. Simulation and Stochastic Optimization
   - Inverse transform; Acceptance Rejection; Composition Method.
   - Metropolis Hastings Algorithm.
   - **Stochastic Optimization via simulation.**
3. Stochastic Difference Equations

References:

1. Pflug, Optimization of Stochastic Models, 1995.
2. Kushner and Yin, 1997.
3. Benveniste, Metvier and Priouret, 1990.
4. Any book on Adaptive filtering

Stochastic Optimization Problem

Aim: Compute control $\theta \in \Theta$ which specifies randomized policy π_θ to optimize

$$C(\theta) = E_{\pi_\theta}\{c(X_n, \theta)\} \text{ where } X_n \sim \pi_\theta$$

E.g. $P(X_{n+1}|X_n) = A_\theta$ with stationary measure π_θ.

Remarks: 1. π_θ not known. Otherwise deterministic opt

$$\min_\theta \mathbb{E}_{\pi_\theta}\{c(X_n, \theta)\} = \min_\theta \int c(X, \theta)\pi_\theta(X)dX$$

2. **No dynamics for** θ: Either constant or slowly evolving.

3. **Discrete Stoch optimization**: If $\theta \in \Theta$ finite.

Evaluate $\hat{C}_N(\theta) = \frac{1}{N}\sum_{n=1}^N c(X_n, \theta)$ for each $\theta \in \Theta$.

Then pick $\theta^* = \arg\min_{\theta\in\Theta} \hat{C}_N(\theta)$.

Since Θ finite, $\arg\min_\theta \lim_{N\to\infty} \hat{C}_N(\theta) = \arg\min_\theta C(\theta)$.

Highly inefficient since equal effort at each element of Θ.

$N|\Theta|$ simulations to determine θ^*.

Key Issue: Efficiency. How to construct algorithm that spends more time at θ^* than other values in Θ?

4. **Continuous Stoch Optimization**: $\Theta \subset \mathbb{R}^p$ compact.

Aim: Compute control $\theta \in \Theta$ which specifies randomized policy π_θ to optimize

$$C(\theta) = E_{\pi_\theta}\{c(X_n, \theta)\} \text{ where } X_n \sim \pi_\theta$$

Andradottir 1996, Multiarmed bandits - Agrawal 1995

$\hat{\nabla}_\theta C_n(\theta)$: unbiased estimate of gradient.
Stochastic gradient (approximation) algorithm

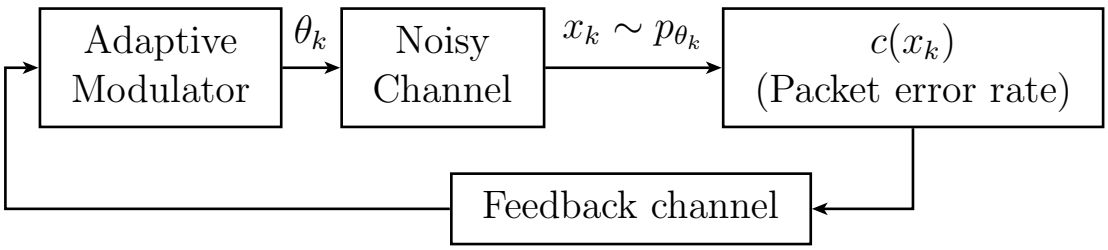
$$\theta_{n+1} = \theta_n - \epsilon_n \hat{\nabla}_\theta C_n(\theta), \quad \text{step size } \epsilon_n$$

converges to local stationary point of $C(\theta)$ under conditions.
(a) Decreasing step size: $\epsilon_n = 1/n$ (wp1 convergence).
(b) Constant step size: $\epsilon_n = \epsilon << 1$ (weak convergence) for tracking time varying θ.

EX 1: LMS algorithm to min $C(\theta) = \mathbb{E}\{(y_n - \psi_n'\theta)^2\}$

$$\theta_{n+1} = \theta_n + \epsilon\psi_n(y_n - \psi_n'\theta)$$

Trivial because: derivative obvious and π_θ is indpt of θ.
EX 2: Q-learning (Reinforcement learning)
EX 3: Policy Gradient learning stochastic control.:
$\min C(\theta) = \mathbb{E}_{\pi_\theta}\{c(X_n)\}$ where $X_n \sim \pi_\theta$ and $c(X_n)$ is black-box (e.g., queuing system).
How to evaluate unbiased estimate $\hat{C}_n(\theta)$ and $\hat{\nabla}_\theta C_n(\theta)$?



Cost: Because $C(\theta) = \int c(X_n)\pi_\theta(X_n)dX_n$ by SLLN

$$\text{Simulate } X_k \sim \pi_\theta \text{ and } \hat{C}_n(\theta) = \frac{1}{N}\sum_{k=1}^{N} c(X_k) \text{ for any } N$$

Gradient: $\nabla_\theta C(\theta) = \int c(X_n)\nabla_\theta \pi_\theta(X_n)dX_n$ but $\nabla_\theta \pi_\theta$ is not a pdf!

1. **Finite difference methods**: Suppose $\theta \in \mathbb{R}^p$.

(a) Kiefer Wolfowitz algorithm: Use the following gradient estimate in stochastic approx alg (e_i is unit indicator vector).

$$\frac{\hat{d}}{d\theta_i}C_n(\theta) = \frac{1}{2\epsilon}\left[\hat{C}_n(\theta + \epsilon e_i) - \hat{C}_n(\theta - \epsilon e_i)\right], \quad i = 1, \ldots, p$$

Disadvantages: $2p$ simulations required.

Bias $\propto \epsilon^2$. If $\hat{C}_n(\theta + \epsilon e_i)$ and $\hat{C}_n(\theta + \epsilon e_i)$ are sampled indpt, then variance $\propto 1/\epsilon^2$; ill-conditioning if ϵ small.

(b) Simultaneous Perturbation Stoch Approx (SPSA): [Spall] Generate the p dimensional vector d_n with random elements $d_n(i)$, $i = 1, \ldots, p$ simulated as follows:

$$d_n(i) = \begin{cases} -1 & \text{with probability 0.5} \\ +1 & \text{with probability 0.5.} \end{cases}$$

gradient estimator

$$\hat{\nabla} C = \frac{\hat{C}_n(\theta_n + \epsilon d_n) - \hat{C}_n(\theta_n - \epsilon d_n)}{2\epsilon} d_n.$$

Only 2 simulations required. Asymptotically efficient as Kiefer Wolfowitz.

Unbiased Gradient Estimation Methods for RVs:

1. **Score function method** Using single sample path:

$$\nabla_\theta C(\theta) = \int c(X)\nabla_\theta \pi_\theta(X)dX = \int c(X)\frac{\nabla_\theta \pi_\theta(X)}{\pi_\theta(X)}\pi_\theta(X)dX$$

So simulate $X_k \sim \pi_\theta$ and compute for any N

$$\hat{\nabla}_\theta C_n = \frac{1}{N}\sum_{k=1}^{N} c(X_k)\frac{\nabla_\theta \pi_\theta(X_k)}{\pi_\theta(X_k)}$$

Example: $\pi_\theta(x) = \theta e^{-\theta x}$, then $\frac{\nabla_\theta \pi_\theta(X)}{\pi_\theta(X)} = \nabla_\theta \log \pi_\theta = \frac{1}{\theta} - X$

2. **Weak Derivative method**: Hahn Jordan decomp: Any signed measure $\nabla_\theta \pi_\theta(X) = a_\theta(f_\theta(x) - g_\theta(x))$ where f_θ and g_θ are orthogonal. So

$$\nabla_\theta C(\theta) = \int c(X)\nabla_\theta \pi_\theta(X)dX = a_\theta \int c(X)[(f_\theta(X) - g_\theta(X)]dX$$

Always has smaller variance than score function method.

Example: $\pi_\theta(x) = \theta e^{-\theta x}$. Then

$$\nabla_\theta \pi_\theta(x) = e^{-\theta x}(1 - x\theta)I(x > \frac{1}{\theta}) - e^{-\theta x}(x\theta - 1)I(x \le \frac{1}{\theta})$$

$a_\theta = \theta e$, $f_\theta(x) = \theta e^{-\theta x}(1 - x\theta)I(x > \frac{1}{\theta})$,
$g_\theta(x) = \theta e^{-\theta x}(x\theta - 1)I(x \le \frac{1}{\theta})$.

3. **Process Derivative**: From inverse transform method, $X \sim F_\theta^{-1}(U)$. So $\frac{dX}{d\theta} = \frac{d}{d\theta}F_\theta^{-1}(U)$. Example: $\pi_\theta(x) = \theta e^{-\theta x}$. Then $X = -\frac{1}{\theta}\log(1 - x)$. So $\frac{dX}{d\theta} = \frac{1}{\theta^2}\log(1 - x)$.

Derivatives of Markov Chains

Given a Markov process with transition matrix P^θ estimate

$$\nabla_\theta \mathbb{E}_{\pi^\theta_\infty}\{c(x)\} = \int_{\mathcal{X}} c(x)\, \nabla_\theta \pi^\theta_\infty(x) dx$$

Clearly for large k, by the law of large numbers

$$\int c(x_k) p^\theta(x_0, x_2, \ldots, x_k) dx_{0:k} \to \int c(x) \pi^\theta_\infty(x) dx$$

$$\int c(x_k) \nabla_\theta p^\theta(x_0, x_2, \ldots, x_k) dx_{0:k} \to \int c(x) \nabla_\theta \pi^\theta_\infty(x) dx$$

$$\begin{aligned}
&\int c(x_k) \nabla_\theta p^\theta(x_0, x_2, \ldots, x_k) dx_{0:k} \\
&\quad = \int c(x_k) \frac{\nabla_\theta p^\theta(x_0, x_2, \ldots, x_k) dx_{0:k}}{p^\theta(x_0, x_2, \ldots, x_k)} p^\theta(x_0, x_2, \ldots, x_k) dx_{0:k} \\
&\quad = \int c(x_k) \nabla_\theta \log p^\theta(x_0, x_2, \ldots, x_k)\, p^\theta(x_0, x_2, \ldots, x_k)\, dx_{0:k}
\end{aligned}$$

$$\{x_k\} \text{ Markov} \implies p^\theta(x_0, x_2, \ldots, x_k) = \pi_0(x_0) \prod_{n=1}^{k} P^\theta_{x_{n-1} x_n}$$

$$\nabla_\theta \log p^\theta(x_0, x_2, \ldots, x_k) = \sum_{n=1}^{k} \frac{\nabla_\theta P^\theta_{x_{n-1} x_n}}{P^\theta_{x_{n-1} x_n}}$$

$$\begin{aligned}
&\int c(x_k) \nabla_\theta p^\theta(x_0, x_2, \ldots, x_k) dx_{0:k} \\
&\quad = \int c(x_k) \sum_{n=1}^{k} \frac{\nabla_\theta P^\theta_{x_{n-1} x_n}}{P^\theta_{x_{n-1} x_n}} p^\theta(x_0, x_2, \ldots, x_k)\, dx_{0:k}
\end{aligned}$$

Step 1. Simulate Markov chain $x_0, \dots x_k$ with transition matrix P^θ.

Step 2. Compute $l_k^\theta = \sum_{n=1}^k \frac{\nabla_\theta P^\theta_{x_{n-1}x_n}}{P^\theta_{x_{n-1}x_n}}$ This can be evaluated recursively as

$$l_n^\theta = \frac{\nabla_\theta P^\theta_{x_{n-1}x_n}}{P^\theta_{x_{n-1}x_n}} + l_{n-1}^\theta, \quad n = 1, \dots, k.$$

Step 3 . Evaluate estimate $\hat{\nabla}_\theta C(\theta) = \frac{1}{k}\sum_{n=1}^k c(x_n) l_n^\theta$

Example from Pflug, pg 257:

$$P^\theta = \frac{1}{6}\begin{bmatrix} 3-3\theta & 2+4\theta & 1-\theta \\ 3-3\theta & 2-2\theta & 1+5\theta \\ 3+3\theta & 2-2\theta & 1-\theta \end{bmatrix}$$

$$\frac{\nabla_\theta P^\theta_{ij}}{P^\theta_{ij}} = \begin{bmatrix} \frac{-1}{1-\theta} & \frac{2}{1+2\theta} & \frac{-1}{1-\theta} \\ \frac{-1}{1-\theta} & \frac{-1}{1-\theta} & \frac{5}{1+5\theta} \\ \frac{1}{1\theta} & \frac{-1}{1-\theta} & \frac{-1}{1-\theta} \end{bmatrix}$$

Convergence of Stochastic Approx Algorithms

$$\theta_{n+1} = \theta_n + \epsilon H(\theta_n, X_n)$$

Until 1976: X_n independent, $\epsilon = 1/n$ (decreasing step size). Basic martingale convergence theorems to show that $\theta_n \to \theta^*$ a.s.

1976: L. Ljung – ODE approach – X_n: geometrically ergodic Markov process parametrized by θ. $\epsilon = 1/n$. Then associated ODE obtained from "averaging theory" is:

$$\frac{d\theta}{dt} = \mathbb{E}_{\pi_\theta(X)}\{H(\theta, X_n)\}$$

Ljung: If ODE has attraction point θ^* then $\theta_n \to \theta^*$ a.s.

1980s: Kushner. X_n: geometrically ergodic Markov process parametrized by θ. Assume ϵ constant step size

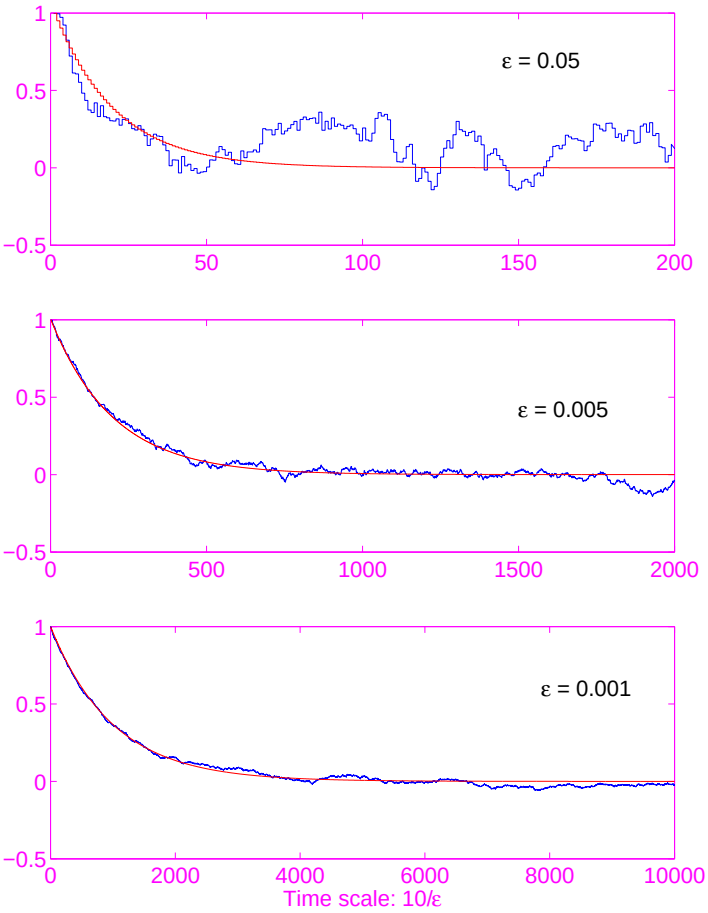
$$\text{Define interpolated process: } \theta_t^\epsilon = \theta_n \quad n \le \frac{t}{\epsilon} < n+1$$

Then provided $H(\theta_n, X_n)$ is uniformly integrable (tightness assumption) the trajectory $\theta_t^\epsilon \to \theta_t$ weakly as $\epsilon \downarrow 0$.

$$\lim_{\epsilon \downarrow 0} P\Big(\sup_{0 \le t < T} |\theta_t^\epsilon - \theta_t| > \rho\Big) \to 0$$

Kushner & Yin, 2003.

Example of Weak Convergence:



Stoch approx algorithm:

$$\theta_{n+1} = \theta_n + \epsilon(-\theta_n + w_n),$$

$w_n \sim N(0,1)$. ODE obtained by averaging

$$\frac{d\theta}{dt} = -\theta_t$$

Define interpolated process

$$\theta_t^\epsilon = \theta_n \quad n \le t/\epsilon < n+1$$

Weak convergence: As $\rho \to 0$,

$$\lim_{\epsilon \downarrow 0} P\Big(\sup_{0 \le t < T} |\theta_t^\epsilon - \theta_t| > \rho \Big) \to 0$$

i.e. convergence of trajectory

Simplistic guide to Weak Convergence

Why: Only constant step size stoch grad algorithms are used in practise to track time varying parameters – so no a.s. convergence. Need to develop stochastic averaging theory.

Convergence in distribution: $X_n \sim F_n(x)$, $X \sim F(x)$, then $X_n \to X$ means $\lim_{n\to\infty} F_n(x) \to F(x)$ or equaivalently: $\int f dP_n \to \int f dP$ for any test fn f.

Weak Convergence: Generalization to fn space – usually uniform metric or Skorohod metric on $D[0,\infty)$.
Defn: $X_t^\epsilon \to X_t$ as $\epsilon \downarrow 0$ weakly if:
(i) Finite distributions converge:
$F(X_{t_1}^\epsilon, X_{t_2}^\epsilon, \ldots, X_{t_n}^\epsilon) \to F(X_{t_1}, X_{t_2}, \ldots, X_{t_n})$
(ii) Behaviour between grid points $t_1, t_2, \ldots,$ is nice – *tight*.
A sufficient condition (Aldous condition) is equi-continuity.

$$P(\sup_{|t-s|<\delta} |X_t^\epsilon - X_s^\epsilon| \geq \rho) \leq \eta \quad \text{for } \epsilon < \epsilon*$$

A sufficient condition is uniform integrability of $H(\theta_n, X_n)$. in

$$\theta_{n+1} = \theta_n + \epsilon H(\theta_n, X_n)$$

Outline of convergence proof

$$\theta_{n+1} = \theta_n + \epsilon H(\theta_n, X_n)$$

$X_n|X_{n-1}, \theta_n$ is parameterized Markov chain.

$$\begin{aligned}\theta_{n+1} &= \theta_n + \epsilon h(\theta_n) + \epsilon \left(H(\theta_n, X_n) - h(\theta_n)\right) \\ &= \theta_0 + \epsilon \sum_{k=1}^{n} h(\theta_n) + \epsilon \sum_{k=1}^{n} \left[H(\theta_k, X_k) - h(\theta_k)\right]\end{aligned}$$

Define interpolated process $\theta_t^\epsilon = \theta_n \quad n \le \frac{t}{\epsilon} < n+1$

$$\theta_t^\epsilon = \theta_0^\epsilon + \epsilon \sum_{k=1}^{t/\epsilon} h(\theta_k) + \epsilon \sum_{k=1}^{t/\epsilon} \left[H(\theta_k, X_k) - h(\theta_k)\right]$$

Then because θ_t^ϵ is weakly compact, extract a convergent subsequence – hop along this subsequence that yields

$$\theta_t = \theta_0 + \int_0^t h(\theta_t) dt$$

Show the limit is unique.

Rate of convergence: Functional central limit theorem holds
Define $u_n = (\theta_n - \theta^*)/\sqrt{\epsilon}$. Suppose $\theta_{n+1} = \theta_n + \epsilon Y_n$. Then

$$du_t = Au_t dt + dw_t$$

where w_t is Brownian motion.