ECE 5412. Bayesian Estimation and Stochastic Optimization

Prof. Vikram Krishnamurthy
Electrical & Computer Engineering
Cornell University
email: vikramk@cornell.edu.
Updated October 3, 2023

Books

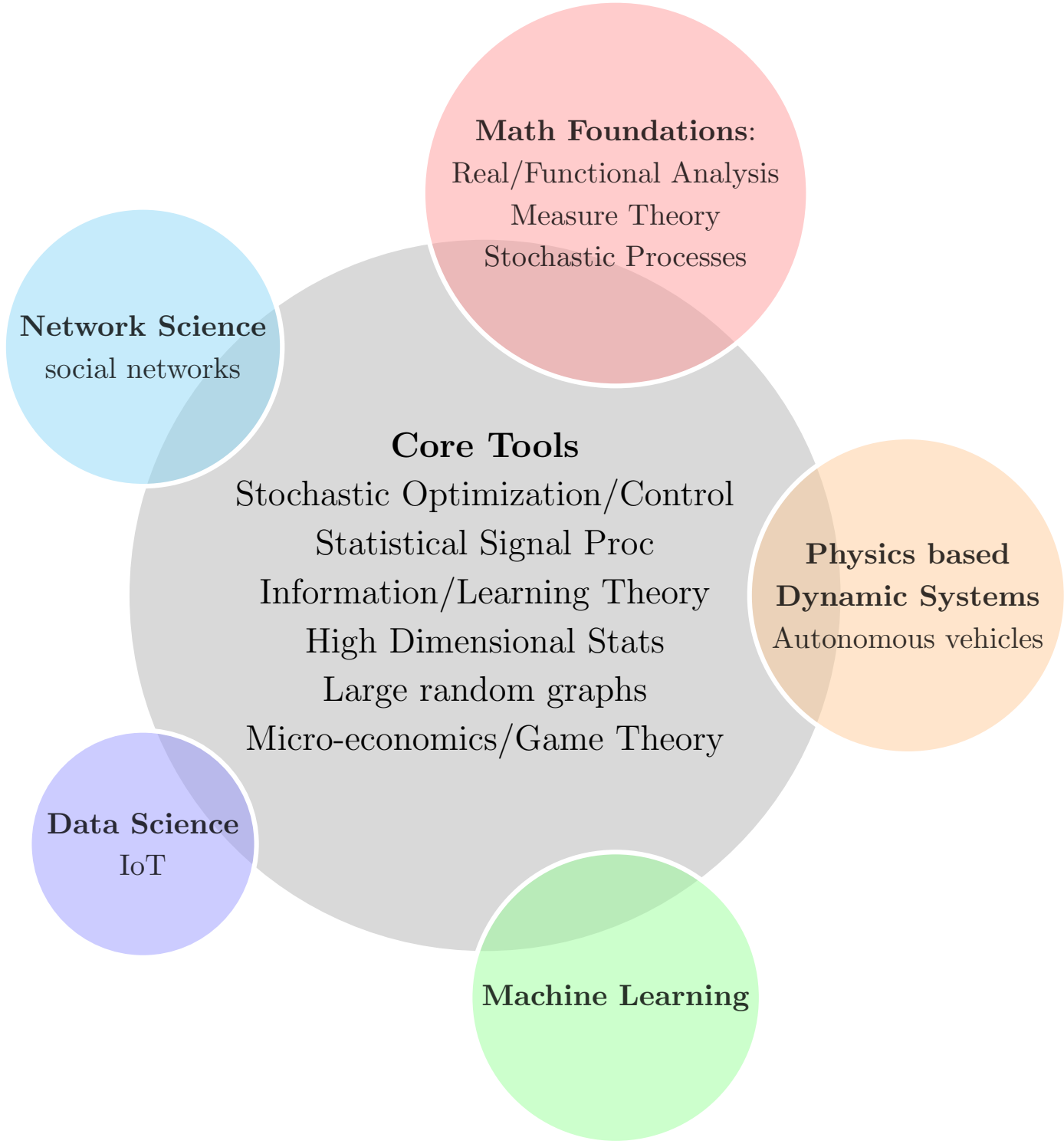
Most material on internet/wikipedia. Graduate level books:

1. V. Krishnamurthy, Partially Observed Markov Decision Processes, Cambridge Univ Press, 2016.
2. L. Lung, System Identification for the user.
3. S. Ross, Simulation
4. Robert & Casella, Monte Carlo Statistical Methods
5. Giraud, Introduction to High Dimensional Statistics

These slides are butchered version of ECE 5412 taught at Cornell comprising 39 lectures.

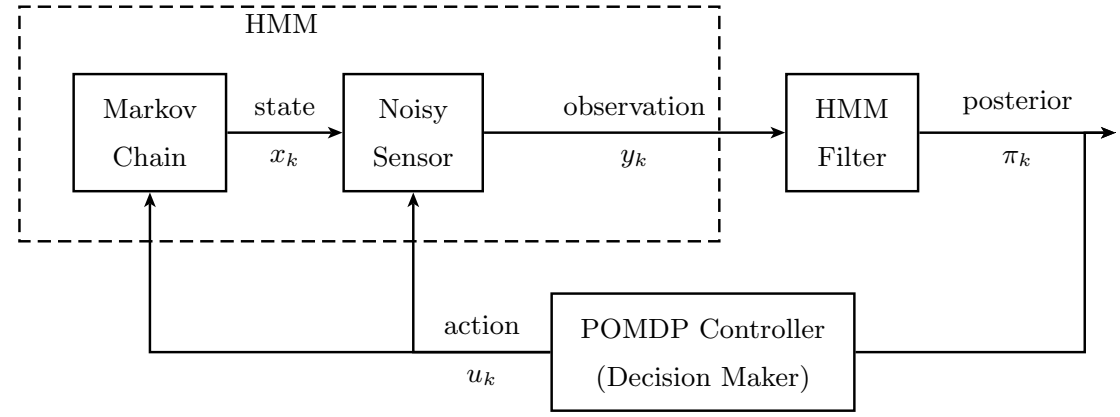
Core Graduate Courses

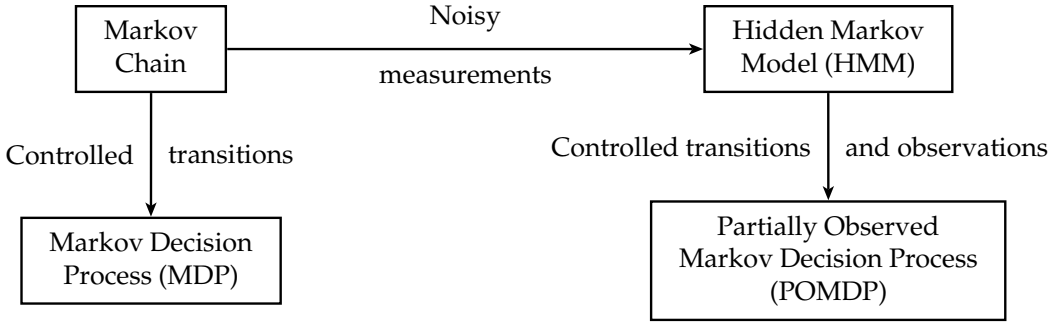
for meaningful PhD research...



Big Picture

Smart (Cognitive) Autonomous Reconfigurable Sensing.





MCMC, Bayesian filtering, ML estimation, stochastic optimization are ubiquitous is sensing, data science, network science, machine learning

---

Outline for Lecture 1 Elementary Background material.

1. Stochastic Simulation
2. Markov chains
3. Optimal Prediction
4. Statistical Inference
5. Hoeffding's Inequality
6. Importance Sampling

IID and Markov processes

IID. $\{X_n\}$ is indpt and identically distributed (iid) on state space $\mathcal{X}$ if conditional densities satisfy:
(i) $p(X_{n+1} = x|x_0, x_1, \ldots, x_n) = p(X_{n+1} = x)$
(ii) $p(X_{n+1} = x)$ has the same pdf/pmf for all n.
Note $\int_{\mathcal{X}} p(X_n = x)dx = 1$.

Markov. $\{X_n\}$ is Markov on state space $\mathcal{X}$ if for all n,

$$p(X_{n+1} = x|x_0, x_1, \ldots, x_n) = p(X_{n+1} = x|x_n) \quad \forall x \in \mathcal{X}$$

Initial density: $\pi_0(x) = p(X_0 = x)$.

Remarks (i) IID processes have memoryless probability laws.
Examples: dice or coin tosses, noise to a first approx
(ii) Markov processes have one-step memory probability laws.
Examples: Most real world signals are Markovian - stock market, speech, video, moving target/vehicle, queuing system
(iii) IID is a special case of Markov

For IID processes, joint distribution of $X_0, \ldots, X_T$ factorizes
$p_{X_0,X_1,\ldots,X_T}(x_0, x_1, \ldots, x_T; 0, 1, \ldots, T) =$

$$p_T(X_T = x_T)\, p_{T-1}(X_{T-1} = x_{T-1}) \times \cdots \times p_0(X_0 = x_0)$$
$$= p(X_T = x_T)\, p(X_{T-1} = x_{T-1}) \times \cdots \times p(X_0 = x_0)$$

For Markov processes, joint distribution of $X_0, \ldots, X_T$ factorizes
$p_{X_0,X_1,\ldots,X_T}(x_0, x_1, \ldots, x_T; 0, 1, \ldots, T) =$

$$p(X_T = x_T|x_{T-1})\, p(X_{T-1} = x_{T-1}|X_{T-2}) \times \cdots$$
$$\times\, p(X_1 = x_1|x_0) \times p(X_0 = x_0)$$

Stochastic Simulation: Scalar RV

$U[0,1]$ uniform pdf with support on $[0,1]$. Matlab `rand(n)` generates an $n \times n$ $U[0,1]$ matrix.

Aim: Given $U[0,1]$ random numbers, generate samples of random variables/processes with specified distributions.

Why? Prototyping; Discrete optimization; Model Validation. In this course: Bayesian inference and Monte-Carlo methods for computing multidimensional integrals efficiently. Given function $\phi : \mathbb{R}^X \to \mathbb{R}$, let $p(\cdot)$ denote pdf on $\mathbb{R}^X$. Then

$$\int_{\mathbb{R}^X} \phi(x)dx = \int_{\mathbb{R}^X} \frac{\phi(x)}{p(x)} p(x)dx = \mathbb{E}_p \left\{ \frac{\phi(x)}{p(x)} \right\}$$

Simulating iid samples $\{x_k\}$, $k = 1, \ldots, N$, from the pdf $p(\cdot)$, Monte-Carlo estimates integral as

$$\boxed{\frac{1}{N} \sum_{k=1}^{N} \frac{\phi(x_k)}{p(x_k)} \text{ where } x_k \sim p(x)} \to \int_{\mathbb{R}^X} \phi(x)dx \text{ as } N \to \infty \text{ w.p.1}$$

Classical Monte-Carlo: iid samples $\{x_k\}$ Markov-chain Monte-Carlo (MCMC): $\{x_k\}$ geometrically ergodic Markov chain with stationary distribution $p(\cdot)$. Covered later.

Simulation of Random Variables

Given $U[0,1]$ random numbers, three elementary methods:

(i) Inverse Transform Method.
Aim: Generate rv $x \sim F$.

Step 1: Generate $u \sim U[0,1]$.
Step 2: Generate $x = F^{-1}(u)$.

Define $F^{-1}(u) = \min\{x : F(x) = u\}$ if $F^{-1}(\cdot)$ is not unique.

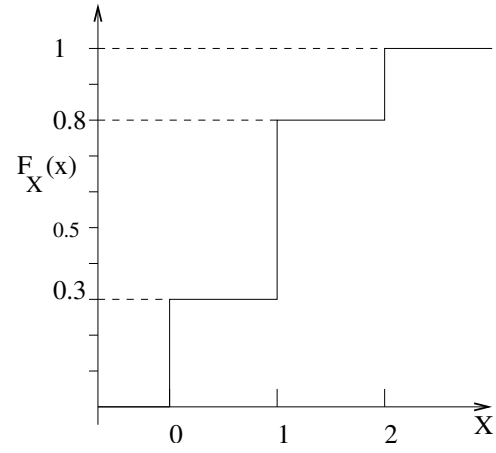
Therefore, if p_i, $i = 1, \ldots, m$ is a probability mass function, then the inverse transform method generates $x \sim p$ as

1. Generate $u \sim U[0,1]$.
2. Generate $x = l^* = \min\{l : u \leq \sum_{k=1}^{l} p_i\}$.

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

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

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



Example 1: Generate rv with cdf $F(x) = x^n$, $0 \leq x \leq 1$.
Soln: $u = F(x) = x^n$ or equivalently, $x = u^{1/n}$. So:
(i) generate rv $u \sim U[0,1]$.
(ii) Compute $u^{1/n}$. This has distribution $F(x)$.

Example 2: Exponentially distributed. $F(x) = 1 - e^{-\lambda x}$, $x \geq 0$, $\lambda > 0$ can be generated as $x = -\frac{1}{\lambda}\log(1-u)$.

Example 3: Generate normal random variables as follows. If $\Theta \sim U[0, 2\pi]$ (uniform pdf) and $R \sim \lambda e^{-\lambda r}$ (exponential pdf) with $\lambda = 1/2$ are independent random variables, then it can be shown that $X = \sqrt{R}\cos\Theta$ and $Y = \sqrt{R}\sin\Theta$ are independent $\boldsymbol{N}(0,1)$ random variables. So if $u_1, u_2 \sim U[0,1]$ are independent, then $x = \sqrt{-2\log u_1}\cos(2\pi u_2)$ and $Y = \sqrt{-2\log u_1}\sin(2\pi u_2)$ are independent $\boldsymbol{N}(0,1)$ random variables.

Example 4: To generate discrete rv, Step 2 comprises of $m-1$ `if` statements and can be inefficient in run time execution if m is large. However, for *discrete uniform mass function* can be implemented efficiently. In this case $p_i = 1/m$ for all $i = 1, 2, \ldots, m$. Then, the above method yields

$$x = l \text{ if } \frac{l-1}{m} \leq u < \frac{l}{m} \text{ or equivalently } x = \text{Int}(mu) + 1. \quad (1)$$

Proof. Let $\bar{F}$ denote the cdf of x generated by the algorithm:

$$\bar{F}(\zeta) = \mathbb{P}\{x \leq \zeta\} = \mathbb{P}\{F^{-1}(u) \leq \zeta\} = \mathbb{P}\{F(F^{-1}(u)) \leq F(\zeta)\}.$$

since F is a monotone non decreasing and so $\alpha \leq \beta$ is equivalent to $F(\alpha) \leq F(\beta)$. Thus, $\bar{F}(\zeta) = \mathbb{P}\{u \leq F(\zeta)\} = F(\zeta)$ where the last equality follows since u is uniformly distributed in $[0,1]$.

How to check if your simulation is correct?

1. Empirical cdf: $\hat{F}_n(x) = \frac{1}{n}\sum_{k=1}^n I(x_k \le x) \to F(x)$
   Matlab: `[F,z] = ecdf(x)` returns empirical cdf `F` evaluated at grid points `z` using the data in the vector `x`.
   Then `plot(z,F)`
2. Empirical pdf: $p_n(x) = \frac{1}{n}\sum_{k=1}^n I(x_k \in [x-\Delta, x+\Delta))$.
   Matlab: Given $x = [x_1, \ldots, x_n]$, use `hist(x, nbins, 1)`

Simulating Gaussian rv (Box Muller eqns) Recall

Result: If $\Theta \sim U(0, 2\pi)$ (uniform pdf) and
$R \sim \lambda e^{-\lambda r}$ exponential pdf with $\lambda = 1/2$ are indpt rvs
then $X = \sqrt{R}\cos\Theta$ and $Y = \sqrt{R}\sin\Theta$ are indpt $N(0,1)$ rvs.

---

1. Generate $U_1 \sim U[0,1]$, $U_2 \sim U[0,1]$ independently
2. Set $d = -2\log U_1$ (inverse transform method for expo)
   Set $\theta = 2\pi U_2$. So obviously $\theta \sim U[0, 2\pi]$
3. Set $X = \sqrt{d}\cos\theta = \sqrt{-2\log U_1}\cos(2\pi U_2)$ and
   $Y = \sqrt{d}\sin\theta = \sqrt{-2\log U_1}\sin(2\pi U_2)$

Then $X \sim N(0,1)$ and $Y \sim N(0,1)$ are indpt rvs.

Simulating Cauchy: $p(x) = \frac{1}{\pi(1+x^2)}$, $F(x) = \frac{1}{2} + \arctan(x)/\pi$.
So $F^{-1}(u) = \tan(\pi(u - \frac{1}{2}))$.

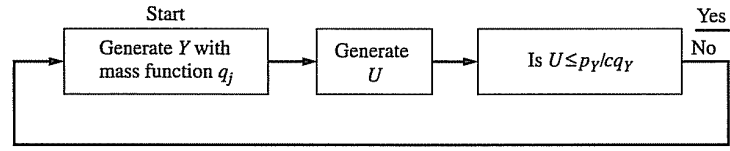
Simulating symmetric α-stable pdfs: $0 \le \alpha \le 2$, $\alpha \ne 1$; see Chambers, Mallows and Stuck, A method for simulating stable rvs, JASA, June 1976.

(ii) Acceptance Rejection Method

Suppose one can generate samples from pdf q. How can random samples be simulated from pdf p? Assume $\max_\zeta \frac{p(\zeta)}{q(\zeta)} < \infty$.

Acceptance Rejection Algorithm Let c denote a constant such that $c \geq \max_\zeta \frac{p(\zeta)}{q(\zeta)}$. Then:

Step 1. Generate $y \sim q$.
Step 2. Generate $u \sim U[0,1]$.
Step 3. If $u < \frac{p(y)}{c\, q(y)}$, set $x = y$.
Otherwise return to step 1.



Example 1. Discrete rv: Want to simulate $X \in \{1, 2, \ldots, 10\}$ 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 (see previous page).
$U_1 \sim U[0,1]$, $Y = \text{Int}(10U_1) + 1$.
Step 2: Generate U
Step 3: If $U < p_Y/(cq_Y)$, set $X = Y$ and stop. Else goto step 1.

Remarks: (i) Acceptance rejection operates on pdf while inverse transform method operates on cdf.
(ii) Self-normalizing: $p(\cdot)$ does not need to be normalized.
(iii) Clearly $c \geq 1$: $c \geq \max_\zeta \frac{p(\zeta)}{q(\zeta)} \implies c \geq \frac{p(\zeta)}{q(\zeta)} \implies c \geq 1$.
(iv) Expected number of iterations is c. Any $c \geq \max p/q$ works.
(v) Matlab code for Acceptance Rejection
`https://www.mathworks.com/examples/statistics/mw/stats-ex85444821-acceptance-rejection-methods`

Each iteration independently yields a probability of acceptance of $\frac{1}{c}$. So the number of iterations to accept is a geometric random variable[a] with mean c and variance $c(c-1)$.

Example 2: Generate rv with cdf $F(x) = x^n, \quad 0 \le x \le 1$.
Soln: Choose $q(x) = U[0,1]$. Then

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

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 3: Generate rv with pdf $p(x) = \frac{2}{\sqrt{2\pi}} e^{-x^2/2}, \quad x \ge 0$. (half normal) from exponential pdf $q(x) = e^{-x}$, $x \ge 0$.

Why? Once we generate $X \sim p(x)$, then $\pm X \sim N(0,1)$.

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

Step 1: simulate exponentially distributed rv y (use inverse transform method). Step 2: generate uniform rv u.
Step 3: Set $x = y$ if $u \le e^{-(y-1)^2/2}$.

---

[a] Geometric distribution models number of trials to the first success when trials are iid with success probability p: So for $n \ge 1$, the probability of n trials to the first success is $p(1-p)^{n-1}$. The expected value and variance of a geometric random variable are $1/p$ and $\frac{1-p}{p^2}$.
B. D. Flury, *Acceptance-Rejection Sampling Made Easy*, SIAM, 1990.

Proof. $\mathbb{P}(x \le \zeta) = \mathbb{P}(y \le \zeta | u \le \dfrac{p(y)}{c\,q(y)}) = \dfrac{\mathbb{P}\left(y \le \zeta, u \le \frac{p(y)}{c\,q(y)}\right)}{\mathbb{P}\left(u \le \frac{p(y)}{cq(y)}\right)}$

$$= \frac{\text{Prob of } y \le \zeta \text{ and accept}}{\text{Prob of accept}} = \frac{\int_{-\infty}^{\zeta} \int_0^{\frac{p(y)}{cq(y)}} du\, q(y) dy}{\int_{-\infty}^{\infty} \int_0^{\frac{p(y)}{cq(y)}} du\, q(y) dy}$$

$$= \frac{\frac{1}{c}\int_{-\infty}^{\zeta} p(y) dy}{\frac{1}{c}\int_{-\infty}^{\infty} p(y) dy}$$

(iii) Composition Method

Aim: Simulate from convex combination of cdfs:
$F(\zeta) = \sum_{i=1}^n p_i F_i(\zeta)$ where $p_i \ge 0$, and $\sum_{i=1}^n p_i = 1$.

1. Generate the integer random sample $i^* \in \{1, \ldots, n\}$ with probability mass function $p_1, \ldots, p_n$
2. Then generate a random sample x from distribution F_{i^*}.

Continuum convex combination: Suppose one can simulate samples from $x \sim p_{x|y}(x|y)$ and $y \sim p_y(y)$. Then samples from

$$p_x(\zeta) = \int_{\mathbb{R}^m} p_{x|y}(\zeta|y) p_y(y) dy \tag{2}$$

can be simulated via the following composition method:

Step 1: Simulate $y^* \sim p_y(\cdot)$.
Step 2: Simulate $x \sim p_{x|y}(\cdot|y^*)$.

The nice property of the above algorithm is that we do not need to compute the integral in (2) in order to simulate from $p_x(\zeta)$.

Example 1. Simulate from cdf $F(x) = \frac{x+x^3+x^5}{3}$, $0 \le x \le 1$.
Example 2. Simulate from $\int_0^\infty x^y e^{-y} dy$, $0 \le x \le 1$.
Soln. (i) Simulate y^* from pdf e^{-y}.
(ii) Simulate x from cdf x^{y^*}, $0 \le x \le 1$.

Example 3. Randomized linear algebra. How to estimate $p'x$ where p is a probability vector and $x \in \mathbb{R}^n_+$?
Standard computations: $\theta = p'x$ requires $O(n)$ multiplications.
Composition method: Note $p'x = \mathbb{E}_p\{x\}$. For $k = 1, 2, \ldots, N$:

1. Generate $i_k \sim p$. Then $x_{i_k} \sim p'x$.
2. Then for sufficiently large N, $\hat{\theta} = \frac{1}{N}\sum_{k=1}^N x_{i_k} \to p'x$

Concentration inequality specifies what N to choose.
Hoeffding inequality. $X_k \in [a, b]$ iid. Then for any $N > 0$,

$$P(\Big|\frac{1}{N}\sum_{k=1}^N x_k - \mathbb{E}\{X_k\}\Big| > \epsilon) \le 2\exp\left(-\frac{2N\epsilon^2}{(b-a)^2}\right)$$

E.g. $a = 0, b = 1$, prob $\le 10^{-4}$, $\epsilon = 10^{-3}$: need $N > 5 \times 10^6$.
So if $n = 10^{10}$, randomized method is more efficient.

We will return to simulation for:

1. Multivariate distributions (via MCMC)
2. For Bayesian inference (particle filters).

Proof of Composition method. Let W denote generated rv

$$P(W \le w) = \int_{\mathbb{R}} P(I(X \le w)|Y = y)\, p(y)dy = \int_{\mathbb{R}} \int_{-\infty}^{w} p(x|y)dx\, p(y)dy$$

since $y \sim p(y)$ in Step 1.

Finite state Markov chains

1. Finite state space $\mathcal{X} = \{1, \ldots, X\}$, i.e., X alphabets.
2. Initial distribution vector π_0 (X dim col vector)

$$\pi_0 = [P(X_0 = 1), P(X_0 = 2), \ldots, P(X_0 = X)]'$$

   Clearly $\pi_0' \mathbf{1} = 1$ i.e., $\sum_{j=1}^{X} \pi_0(j) = 1$.
3. Transition probabilities: $P_{ij} = P(X_k = j | X_{k-1} = i)$ where $0 \le P_{ij} \le 1$, $\sum_{j=1}^{X} P_{ij} = 1$ for $i = 1, 2, \ldots, X$.
   $X \times X$ transition probability matrix P with elements

$$P_{ij} = P(X_k = j | X_{k-1} = i), \quad 0 \le P_{ij} \le 1, \sum_{j=1}^{X} P_{ij} = 1 \equiv P\mathbf{1} = \mathbf{1}$$

Remarks

1. P always has one eigenvalue at 1 for eigenvector $\mathbf{1}$.
2. P^k is always a stochastic matrix for any integer $k \ge 0$.
3. If all rows of P are identical then iid
4. More precisely, first-order homogeneous Markov chain.

Example: Three state Markov chain with absorbing state:

$$\mathcal{X} = \{1, 2, 3\}, \ \pi_0 = \begin{bmatrix} 0.3 \\ 0.7 \\ 0 \end{bmatrix}, \quad P = \begin{bmatrix} 0.8 & 0.1 & 0.1 \\ 0.6 & 0.4 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

States 1 and 2 are transient; state 3 is absorbing.

Q: Given N point sample path how to estimate transition prob?

$$\hat{P}_{ij} = \frac{\text{number of jumps from } i \text{ to } j}{\text{number of times Markov chain is in } i}$$

Simulating a Markov chain

1. IID process: Repeated use of inverse transform/acceptance rejection where uniform numbers at each step are independent.

2. Markov chain: Let P_i denote the i-th row of P. Since given X_k, state X_{k+1} is conditionally independent of the past

1. Generate $X_0 \sim \pi_0$.
2. For $k = 1, 2, \ldots,$ generate $X_k \sim P_{x_{k-1}}$.

Step 1 and 2 using inverse transform or acceptance rejection.

Markov chain State Properties:

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

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

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

$$v_{ij} = \sum_{k=0}^{\infty} \mathbb{E}\{\mathbf{1}_{X_k=j} | X_0 = i\} = \sum_{k=0}^{\infty} P\{X_k = j | X_0 = i\} = \sum_{k=0}^{\infty} P_{ij}^k$$

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 $P(X_k = i | X_0 = i) = 0$ whenever k is not divisible by n. If period $n = 1$, state is *aperiodic*. If all states are aperiodic, Markov chain is called aperiodic.

Chapman Kolmogorov Theorem - Optimal Prediction

Result: (Chapman Kolmogorov (CK) eqn) Given π_0 and P, the state probability vector at time k

$$\pi_k = [P(X_k = g_1), \ldots, P(X_k = g_X)]'$$

can be computed as

$$\boxed{\pi_{k+1} = P'\pi_k = {P'}^{k+1}\pi_0}$$

Then predicted state mean at time k is

$$\hat{X}_k = \mathbb{E}\{X_k\} = g'\pi_k$$

Predicted mean is $\hat{x}_k$ is MMSE optimal predictor:

$$\mathbb{E}\{(X_k - \hat{X}_k)^2\} \leq \mathbb{E}\{(X_k - \phi(\pi_0))^2\}.$$

Proof of CK: From total probability rule

$$\pi_{k+1}(j) = P(X_{k+1} = g_j) = \sum_{i=1}^{X} P(X_{k+1} = g_j | X_k = g_i)\, P(X_k = g_i)$$

$$= \sum_{i=1}^{X} P_{ij}\pi_k(i)$$

Remark: State probability vector π_k evolves as LTI system with state matrix P. Clearly, $\pi_k'\mathbf{1} = 1$ for all k.

Convergence of Markov Chains

Limiting Distribution *Limiting distribution* is

$$\lim_{k\to\infty} \pi_k = \lim_{k\to\infty} {P'}^k \pi_0.$$

This limiting distribution may not exist. For example if

$$P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \pi_0 = \begin{bmatrix} \pi_0(1) \\ \pi_0(2) \end{bmatrix}, \quad \text{then } \pi_k = \begin{cases} \begin{bmatrix} \pi_0(2) & \pi_0(1) \end{bmatrix}' & k \text{ odd} \\ \begin{bmatrix} \pi_0(1) & \pi_0(2) \end{bmatrix}' & k \text{ even} \end{cases}$$

and so $\lim_{k\to\infty} \pi_k$ does not exist unless $\pi_0(1) = \pi_0(2) = 1/2$.

Stationary Distribution X-dimensional vector π_∞

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

So π_∞ is normalized right eigenvector of P' with eigenvalue 1. Equivalently, choosing $\pi_0 = \pi_\infty$ implies $\pi_k = \pi_\infty$ for all k. Stationary distribution also called the *invariant, equilibrium or steady-state distribution.*

Limiting distributions are a subset of stationary distributions. For example, for above P, $\pi_\infty = \begin{bmatrix} 0.5 & 0.5 \end{bmatrix}'$ is a stationary distribution but there is no limiting distribution.

Markov chain properties:

1. A Markov chain is regular (primitive) if for some $k \geq 1$, all elements of P^k are strictly positive.
2. A Markov chain is irreducible if for each $i, j \in \mathcal{X}$, there exists $k \geq 1$ such that $P^k_{ij} > 0$.

Theorem 1 (Perron-Frobenius). *Consider a finite-state Markov chain with regular transition matrix P. Then:*

1. *The eigenvalue 1 has algebraic & geometric multiplicity of one.*
2. *All remaining eigenvalues of P have modulus strictly smaller than 1.*
3. *The eigenvector of P' corresponding to eigenvalue of 1 can be chosen with non-negative elements.*
4. *$P^k = \mathbf{1}\pi'_\infty + O(|\lambda_2|^k)$ where λ_2 is the second largest eigenvalue modulus (SLEM).*
5. *Limiting distribution and stationary distribution coincide.*

$$\pi_\infty(i) = \lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} I(x_k = i)$$

is fraction of time Markov chain spends in state i.

Statement 4 says if the transition matrix P is regular, state probability vector π_k forgets initial condition geometrically fast.

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

So k-step ahead predictor of a Markov chain forgets initial condition geometrically fast in terms of the second largest eigenvalue modulus, $|\lambda_2|$.
Equivalently, π_k converges geometrically fast to π_∞.

Proof of Statement 2: Define spectral radius $\rho(A) = \max_i |\lambda_i|$
Lemma 1: $\rho(A) \leq \|A\|_\infty$ where $\|A\|_\infty = \max_i \sum_j |a_{ij}|$
Proof: $|\lambda|\|x\| = \|\lambda x\| = \|Ax\| \leq \|A\|\|x\| \implies |\lambda| \leq \|A\| \quad \forall \lambda.$

In our case $\|A\|_\infty = 1$ and A has an eigenvalue at 1. So $\rho(A) = 1$.

Proof of Statement 3:
Lemma 2. For positive matrix A, $A'\pi = \pi$ implies $A'|\pi| = |\pi|$
Proof: $|\pi| = |A'\pi| \leq |A'||\pi| = A'|\pi|$ where $\leq$ follows from the triangle inequality.

So $A'|\pi| - |\pi| \geq 0$.
But $A'|\pi| - |\pi| > 0$ is impossible, since it implies $1'A'|\pi| > 1'|\pi|$, i.e., $1'|\pi| > 1'|\pi|$.
Therefore $A'|\pi| = |\pi|$.

Dobrushin Coefficient of Ergodicity

Result: SLEM $\lambda_2 \leq \rho(P)$ where

$$\boxed{\rho(P) = \frac{1}{2} \sup_{i,j} \sum_k |P_{ik} - P_{jk}| = \sup_{i,j} \|P' e_i - P' e_j\|_{\text{TV}}.}$$

$\rho(P) \in [0,1]$ is max variational dist between two rows of P.
Given pmfs α and β variational distance

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

- If $\rho(P) < 1$, then Markov chain is geometrically ergodic. This implies SLLN: $\mu_n = \frac{1}{n}\sum_{k=1}^n X_k \to \pi_\infty$.
- $\rho(P)$ is a lower bound for convg rate.

Example 1: If $P = \begin{bmatrix} P_{11} & 1 - P_{11} \\ 1 - P_{22} & P_{22} \end{bmatrix}$, then
$\rho(P) = |1 - P_{11} - P_{22}| = |\lambda_2|$.
Example 2: If $P_{ij} \geq \epsilon$ for all i, j, then $\rho(P) \leq 1 - \epsilon$. All non-zero transition probabilities: trivially irreducible and aperiodic.

Example 3. Doeblin/Minorization condition: (advanced).

$$P_{ij} \geq \epsilon \kappa_j, \quad \sum_j \kappa_j = 1, \kappa_j \geq 0 \implies \rho(P) \leq 1 - \epsilon$$

Remark: $\rho(P) = \sup_{i \neq j} \frac{W(P' e_i, P' e_j)}{W(e_i, e_j)}$; W is Wasserstein distance.

Optimal Prediction: Chapman Kolmogorov Equation

Aim: How to optimally predict future state given current state probability?
Given Markov process with transition density $p(x_{k+1}|x_k)$ and initial condition π_0, compute state pdf $\pi_k(x) = p(x_k = x)$ at time k.

We call π_k the *predicted* density.

Chapman Kolmogorov: From total probability rule

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

Therefore predicted state and covariance at time k are

$$\boxed{\begin{aligned} \hat{x}_k = \mathbb{E}\{x_k\} &= \int_{\mathcal{X}} x\, \pi_k(x) dx, \\ \text{cov}(x_k) = \mathbb{E}\{(x_k - \hat{x}_k)(x_k - \hat{x}_k)'\} &= \mathbb{E}\{x_k x_k'\} - \hat{x}_k \hat{x}_k'. \end{aligned}}$$

Predicted mean is $\hat{x}_k$ is optimal in the minimum mean square error sense:

$$\mathbb{E}\{(x_k - \hat{x}_k)^2\} \leq \mathbb{E}\{(x_k - \phi(\pi_0))^2\}.$$

Hence called "optimal predictor".

Simulation-based Optimal State Predictor

Consider general stochastic state evolution model:

$$x_{k+1} = A_k(x_k) + \Gamma_k(x_k)w_k, \quad x_0 \sim \pi_0$$

$$\implies p(x_{k+1}|x_k) = |\Gamma_k^{-1}(x_k)|\; p_w\left(\Gamma_k^{-1}(x_k)\;[x_{k+1} - A_k(x_k)]\right)$$

Chapman Kolmogorov equation for predicted state density:

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

Direct intergration is intractable. Can use stochastic simulation. Given random samples from the predicted state density $\pi_{k-1}(x)$ at time $k-1$, how to simulate samples from the predicted state density $\pi_k(x)$ at time k?

Solution. The composition method for stochastic simulation generates samples as follows:

1. Generate samples $x_{k-1}^{(l)}$, $l = 1, \ldots, L$ from $\pi_{k-1}(x)$.
2. Generate sample $x_k^{(l)} \sim p(x_k|x_{k-1} = x_{k-1}^{(l)})$, for $l = 1, 2, \ldots, L$.

By composition method, simulated samples $x_k^{(l)}, l = 1, \ldots, L$ are from pdf $\pi_k(x)$.

By Glivenko Cantelli Thm, for large L, these samples uniformly approximate pdf $\pi_k(x)$.

Then mean, variance, and other statistics can be estimated.

Example. Linear Gaussian State Space Model

Notation:

$$\boldsymbol{N}(\zeta;\mu,\Sigma) = (2\,\pi)^{-l/2}\,|\Sigma|^{-1/2}\exp\left[-\frac{1}{2}(\zeta-\mu)'\,\Sigma^{-1}(\zeta-\mu)\right].$$

Sometimes shorter notation $\boldsymbol{N}(\mu,\Sigma)$ will be used.

The linear Gaussian state space model

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

$x_k \in \mathcal{X} = \mathbb{R}^X$, $y_k \in \mathcal{Y} = \mathbb{R}^Y$, $A_k \in \mathbb{R}^{X\times X}$, $C_k \in \mathbb{R}^{Y\times X}$.

Transition density form:

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

Optimal Predictor Using the Chapman Kolmogorov equation $\pi_{k+1} = \boldsymbol{N}(\hat{x}_{k+1},\Sigma_{k+1})$ where

$$\boxed{\begin{aligned} \hat{x}_{k+1} &= \mathbb{E}\{x_{k+1}\} = A_k\hat{x}_k \\ \Sigma_{k+1} &= \mathrm{cov}\{x_{k+1}\} = A_k\Sigma_k A_k' + Q_k \\ \mathrm{cov}(x_{k+n},x_k) &= A^n\Sigma_k, \qquad n \geq 0 \end{aligned}}$$

Covariance update is called *Lyapunov equation.*
Same mean and covariance recursions hold for non-gaussian case

Motivation: Predicting target's coordinates (without measurements).

Proof. Evolution of covariance: Let $\tilde{x}_k = x_k - \hat{x}_k$ Then

$$\begin{aligned}
\tilde{x}_{k+1} &= A\tilde{x}_k + w_k \\
\tilde{x}_{k+1}\tilde{x}'_{k+1} &= (A\tilde{x}_k + w_k)(A\tilde{x}_k + w_k)' \\
&= A\tilde{x}_k\tilde{x}'_k A' + A\tilde{x}_k w'_k + w_k\tilde{x}'_k A' + w_k w'_k
\end{aligned}$$

Taking expectations on both sides yields result

Scalar Example In the scalar case

$$\begin{aligned}
\hat{x}_{k+1} &= A\hat{x}_k = A^{k+1}\hat{x}_0 \\
\Sigma_{k+1} &= A^2\Sigma_k + Q \\
&= A^{2k}\Sigma_0 + \frac{1-A^{2k}}{1-A^2}Q
\end{aligned}$$

$$\operatorname{cov}[x_k, x_l] = \begin{cases} A^{l-k}\Sigma_k & l \geq k \\ A^{k-l}\Sigma_k & l < k \end{cases}$$

So if $|A| < 1$, then as $k \to \infty$,

$$\hat{x}_k \to 0,\ \Sigma_k \to Q/(1-A^2),\ \operatorname{cov}(x_k, x_{\tau+k}) = Q\frac{A^\tau}{1-A^2}$$

Thus if $|A| < 1$, x_k becomes a weakly stationary process. This can be generalized to vector processes by requiring that the eigenvalues of A lie within the unit circle for asymptotic weak stationarity.

Statistical Inference

Two fundamental results:
(i) Law of large numbers (ii) Central Limit Theorem.
Law of Large Numbers (LLN). LLN relates

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

For iid process $X_n(\omega)$ two averages can be computed:
1. Expected value (ensemble average) for fixed n: (from pdf)

$$\mathbb{E}\{X_n(\omega)\} = \int_{\mathbb{R}} x f_{X[n]}(x)dx = \int x f_X(x) = \mu$$

Note: Recall because iid, μ is a const.

2. Real life sample path time average (statistic) for fixed ω given N observations: (in real life you live one sample path)

$$\hat{\mu}_N(\omega) = \frac{1}{N}(X[\omega,1] + \ldots + X[\omega,N])$$

Result Strong LLN: For iid process as $N \to \infty$, $\hat{\mu}_N(\omega)$ converges "strongly" to μ (statistic $\hat{\mu}_N$ computed from data converges to mean of random variable μ computed from probabilistic model.)

Calculus 101: A sequence $\{x_n\}$ converges to x if for every $\epsilon > 0$, there exists N_ϵ s.t. $|x_n - x| < \epsilon$ for all $n \geq N_\epsilon$.
Equivalently: For all $n \geq N_\epsilon$, an ϵ excursion $|x_n - x| \geq \epsilon$ never happens.
How to generalize convergence to case when x_n and x are random variables?

Aside. Stochastic Convergence

Consider $(\Omega, \mathcal{F}, P)$. (i) A sequence of rv $\{X_n\}$ converges in probability to rv X if for any $\epsilon > 0$, $\lim_{n\to\infty} p_n = 0$ where

$$p_n = P(\underbrace{\{\omega : |X_n(\omega) - X(\omega)| > \epsilon\}}_{\epsilon \text{ excursion at time } n}).$$

Equivalently: $\boxed{\lim_{n\to\infty} P(|X_n - X| > \epsilon) = 0, \quad \forall \epsilon > 0}$

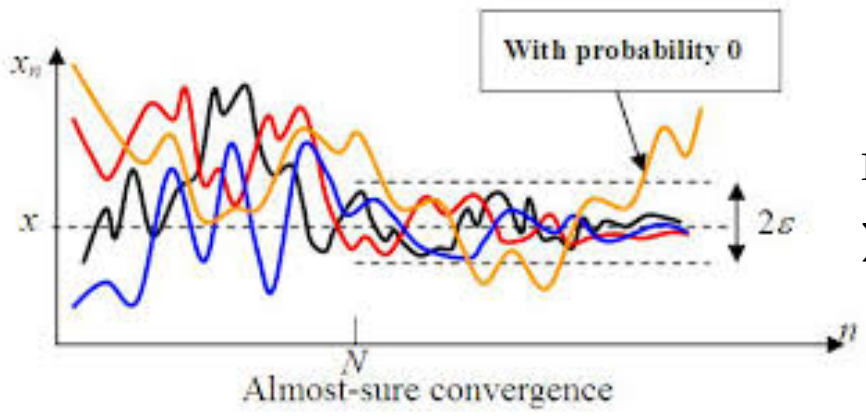
(ii) $\{X_n\}$ converges almost surely (with probability one) to rv X if for any $\epsilon > 0$, $\lim_{n\to\infty} a_n = 0$ where

$$a_n = P(\underbrace{\{\omega : \exists m \geq n \text{ such that } |X_m(\omega) - X(\omega)| > \epsilon\}}_{\epsilon \text{ excursion at or beyond time } n})$$

$$= P(\{\omega : \cup_{m\geq n} |X_m(\omega) - X(\omega)| > \epsilon\}) \quad (\epsilon \text{ excursion at time } n \text{ or } n+1 \text{ or } \ldots)$$

Since $\lim_n P(\cup_{m\geq n}) = P(\lim_n \cup_{m\geq n})$, so X_n converges to X w.p.1 if

$$\boxed{\lim_{n\to\infty} a_n = P(\{\omega : \lim_{n\to\infty} \cup_{m\geq n} |X_m(\omega) - X(\omega)| > \epsilon\}) = 0}$$



Borel Cantelli Lemma

$\sum_n p_n < \infty \implies \lim_{n\to\infty} a_n = 0$

(iii) $\{X_n\}$ converges in distribution if $\lim_{n\to\infty} F_n(x) = F(x)$ for all x for which F is continuous. Here F_n, F are cdfs of X_n, X.

Theorem. a.s. convergence $\implies$ convergence in probability $\implies$ convergence in distribution.

Proof. Obviously event associated with a.s. convergence is superset of event associated with convergence in prob:

$$\underbrace{\{\omega : \cup_{m\geq n}|X_m(\omega) - X(\omega)| > \epsilon\}}_{\epsilon \text{ excursion at } n \text{ or any future time}} \supseteq \underbrace{\{\omega : |X_n(\omega) - X(\omega)| > \epsilon\}}_{\epsilon \text{ excursion at } n}.$$

So $a_n \geq p_n$. As a result $a_n \to 0$ implies $p_n \to 0$. So almost sure convergence always implies convergence in probability

---

A.s. convergence: $\Omega = C \cup \bar{C}$ (disjoint). For each sample path $\omega \in C$, $\lim_n X_n(\omega) = X$ and $P(C) = 1$, $P(\bar{C}) = 0$.
Classical Convergence. $\bar{C} = \emptyset$. Example in appendix.

Under what additional conditions does convergence in probability imply almost sure convergence?
Equivalently: When does $p_n \to 0$ imply $a_n \to 0$?
Recall $p_n = P(\{\omega : |X_n(\omega) - X(\omega)| > \epsilon\})$.

$$\begin{aligned} a_n &= P(\{\omega : \cup_{m\geq n}|X_m(\omega) - X(\omega)| > \epsilon\}) \\ &\leq \sum_{m\geq n} P(\{\omega : |X_m(\omega) - X(\omega)| > \epsilon\}) = \sum_{m\geq n} p_n \end{aligned}$$

where the inequality follows since $\mathbb{P}(\cup_m A_m) \leq \sum_m \mathbb{P}(A_m)$. So if as $n \to \infty$, $\sum_{m\geq n} p_m \to 0$, then $a_n \to 0$. Therefore

Borel Cantelli Lemma

$$\boxed{\sum_{n=1}^{\infty} p_n < \infty \implies a_n \to 0.}$$

Summary: $\sum_n P(\{\omega : |X_n(\omega) - X(\omega)| > \epsilon\}) < \infty$ implies $X_n \xrightarrow{\text{a.s.}} X$.

Borel Cantelli lemma: for proving a.s. convergence

1. ϵ-excursion event at time n:

$$B_n = \{w : |X_n(w) - X(w)| > \epsilon\}; \quad p_n = P(B_n)$$

ϵ-excursion event from time n to ∞:

$$A_n = \{w : \cup_{m \geq n} |X_n(w) - X(w)| > \epsilon\}; \quad a_n = P(A_n)$$

Then $\sum_n p_n < \infty$ implies $a_n \to 0$, i.e, $X_n \xrightarrow{\text{a.s.}} X$

2. *Converse.* If $\{B_n\}$ are independent, then $\sum_n P(B_n) = \infty$ implies X_n does not converge a.s.

Example 1: Suppose $P(X_n = 1) = p_n$ and $P(X_n = 0) = 1 - p_n$.

1. Then if $p_n \to 0$, $X_n \to 0$ in probability.
2. For $X_n \xrightarrow{wp1} 0$, need $\sum_n p_n < \infty$, e.g. $p_n = 1/n^{1+\alpha}$, $\alpha > 0$.
3. X_n indpt, $p_n = 1/n$: then X_n does not converge to 0 wp1.

Example 2. Weak Law of Large numbers.
Theorem. Assume $\{X_n\}$ iid with $\mathbb{E}\{X\} = \mu$, $\text{Var}(X) = \sigma^2 < \infty$. Let $\mu_n = \frac{1}{n} \sum_{k=1}^n X_n$. Then $\mu_n \to \mu$ in probability .

Proof. Clearly $\text{Var}\{\mu_n\} = \sigma^2/n$. Chebyshev inequality implies

$$p_n = P(|\mu_n - \mu| > \epsilon) \leq \frac{\sigma^2}{n\epsilon^2} \to 0 \text{ as } n \to \infty$$

Example 3. Strong Law of Large numbers. How to prove $\mu_n \to \mu$ a.s.?
Example 2 $\implies$ $\sum_n p_n = \infty$; so cannot directly use Borel Cantelli
But for sequence $\{p_{n^2}\}$ clearly $\sum_n p_{n^2} < \infty$. So $\mu_{n^2} \to \mu$ a.s.
More advanced course: use this to show $\mu_n \to \mu$ a.s.

Result 1. Strong Law of Large Numbers (SLLN)

Define the sample average as $\mu_n = \frac{1}{n}\sum_{k=1}^{n} X_k$. SLLN states

Theorem 2. *(IID) Suppose $\{X_k\}$ is an i.i.d. sequence of vector random variables. Then $\lim_{n\to\infty} \mu_n = \mathbb{E}\{X_1\}$ almost surely iff $\mathbb{E}\{|X_1|\} < \infty$ where $\mathbb{E}\{X_1\} = \mu$.*

(Finite-state Markov) Suppose $\{X_n\}$ is an X-state Markov chain with state space of X-dimensional unit vectors. Assume transition matrix P is regular. Then $\lim_{n\to\infty} \mu_n = \pi_\infty$ almost surely.

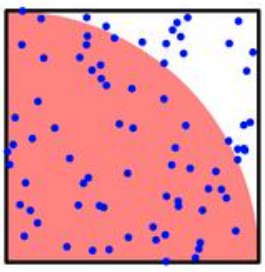
Remarks: SLLN is basis of Monte Carlo (MC) inference.

1. The i.i.d. version also called Kolmogorov's SLLN.
2. For Markov chain $\pi_\infty(i)$ is fraction of time spent in state i.
3. A random process for which SLLN holds is called "ergodic". So any iid process is ergodic. For Markov chain, need geometric ergodicity for SLLN.

Example 1. Multidimensional Monte-Carlo integration: Compute difficult integrals numerically by stochastic simulation and LLN. As $n \to \infty$,

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

A dumb approximation of π



$$P(x,y) = \begin{cases} 1 & 0<x<1 \text{ and } 0<y<1 \\ 0 & \text{otherwise} \end{cases}$$

$$\pi = 4\iint \mathbb{I}\left((x^2+y^2)<1\right)P(x,y)\,dx\,dy$$

In Matlab avoid "for loops". Use vectorized code; much faster

Example 2. If not iid, then LLN may not hold.

1. 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
   $w = [0, 0, 0, 0, 0, 0, \ldots,]$ or $w = [1, 1, 1, 1, 1 \ldots,]$.
   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., LLN does not hold.
   Here $X[n]$ are identically distributed but not indpt. For non-iid sources, need to be more careful.
2. SLLN does not hold for Cauchy density since $\mathbb{E}\{|X|\}$ is not finite.

Example 3: Estimation of a cdf from random samples.
Suppose $\{X_n\}$ is an i.i.d. sequence simulated from an unknown cdf F. The empirical cdf is defined as

$$\boxed{F_n(x) = \frac{1}{n} \sum_{k=1}^{n} I(X_k \leq x)}$$

$F_n(x)$ is a natural estimator of F when F is not known. By SLLN $\lim_{n\to\infty} F_n(x) = F(x)$ almost surely for each x. Actually for estimation of cumulative distributions uniform almost sure convergence. (Glivenko-Cantelli Theorem)

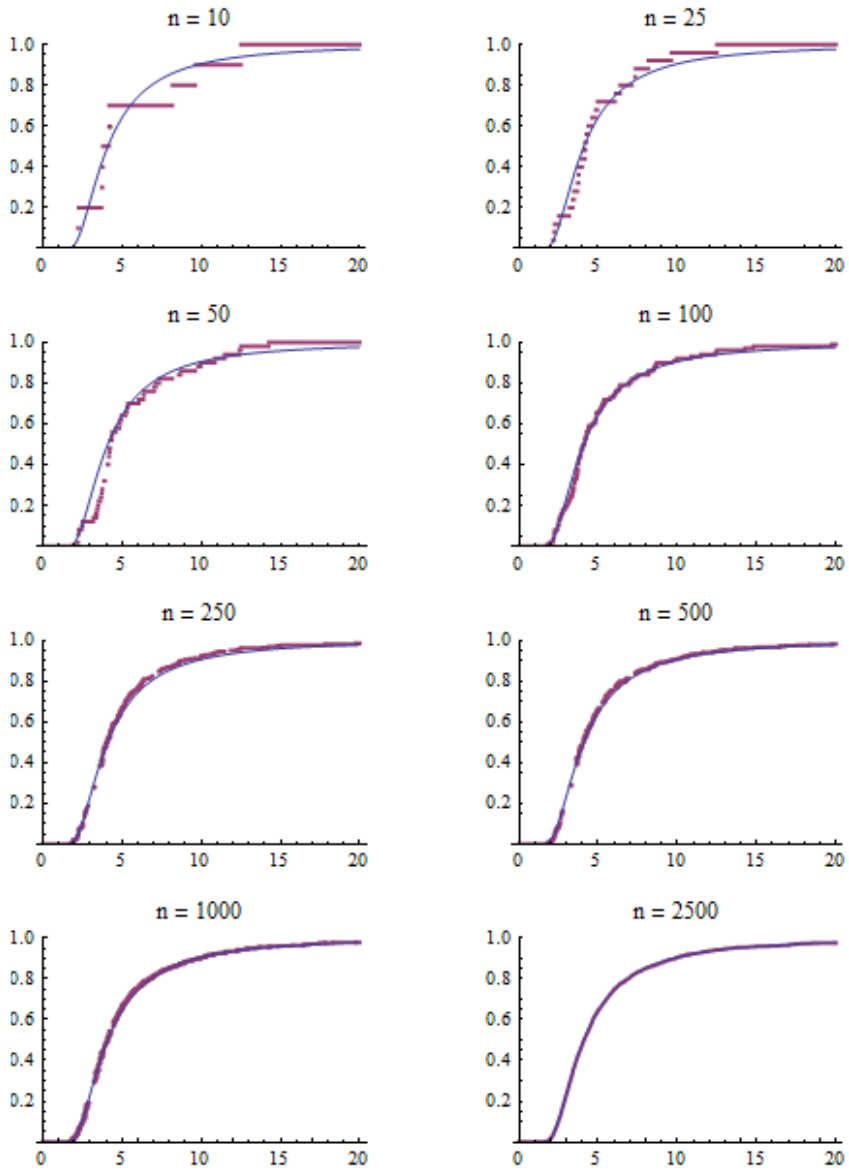
In Matlab: `[F,z] = ecdf(x)` returns empirical cdf `F` evaluated at grid points `z` using the data in the vector `x`. Then `plot(z,F)`

Theorem 3 (Glivenko-Cantelli Theorem). *Suppose* $\{X_n\}$ *is an i.i.d. sequence with cdf* F. *Then uniform SLLN holds:*

$$\lim_{n\to\infty} \sup_{x\in\mathbb{R}} |F_n(x) - F(x)| = 0 \quad \textit{almost surely.}$$

"Large number of random samples uniformly approximates cdf."

Essential idea in stochastic simulation and particle filtering.



Dvoretsky-Kiefer-Wolfowitz inequality:

$$P(\sup_x |F_n(x) - F(x)| > \epsilon) \leq 2e^{-2n\epsilon^2}$$

Example 4. Shannon-McMillan-Breiman Theorem.
Assume

1. $y_k \sim p(y|\theta^o)$ iid, $k = 1, 2 \ldots, n$.
2. We dont know θ^o and assume model θ.

SLLN gives iid version of Shannon-McMillan-Breiman theorem:

$$\frac{1}{n}\sum_{k=1}^{n} \log p(y_k|\theta) \xrightarrow{\text{wp1}} \mathbb{E}_{\theta^o}\{\log p(y|\theta)\} = \int \log p(y|\theta)\, p(y|\theta^o) dy$$

$$\frac{1}{n}\sum_{k=1}^{n} \log p(y_k|\theta^o) \xrightarrow{\text{wp1}} \mathbb{E}_{\theta^o}\{\log p(y|\theta^o)\} = \int \log p(y|\theta^o)\, p(y|\theta^o) dy$$

Negative of KL divergence is

$$\begin{aligned} -K(\theta, \theta^o) &= \mathbb{E}_{\theta^o}\{\log p(y|\theta)\} - \mathbb{E}_{\theta^o}\{\log p(y|\theta^o)\} \\ &= \int \log \frac{p(y|\theta)}{p(y|\theta^o)}\, p(y|\theta^o) dy \leq 0 \end{aligned}$$

since by Jensen inequality: $\mathbb{E}\{\log(X)\} \leq \log(\mathbb{E}\{X\})$
Also clearly $K(\theta^o, \theta^o) = 0$.
So to estimate true model, we should maximize $-K(\theta, \theta^o)$ wrt θ.
Equivalently maximize $\mathbb{E}_{\theta^o}\{\log p(y|\theta)\}$ wrt θ
Equivalently maximize $\frac{1}{n}\sum_{k=1}^{n} \log p(y_k|\theta)$ wrt θ for large n.
Equivalently maximize $p(y_1, \ldots, y_n|\theta)$ wrt θ
This is the basis of maximum likelihood estimation (Part IV).

Result 2. Central Limit Theorem

Theorem 4. *With $\mu_n = \frac{1}{n}\sum_{k=1}^n X_k$ and $\stackrel{\mathcal{L}}{=}$ denoting convergence in distribution, the following hold:*

(IID) Suppose $\{X_k\}$ is iid with zero mean and finite variance σ^2. Then $\lim_{n\to\infty}\sqrt{n}\mu_n = \frac{1}{\sqrt{n}}\sum_{k=1}^n X_k \stackrel{\mathcal{L}}{=} \boldsymbol{N}(0,\sigma^2)$.

Equivalently: $\mu_n \stackrel{\mathcal{L}}{=} \boldsymbol{N}(0,\sigma^2/n)$.

(Finite-state Markov) Suppose $\{X_n\}$ is an X-state Markov chain with state space comprising of X-dimensional unit vectors. Suppose P is regular implying a unique stationary distribution π_∞ exists. Then for any $g \in \mathbb{R}^X$,

$$\lim_{n\to\infty}\sqrt{n}g'(\mu_n - \pi_\infty) \stackrel{\mathcal{L}}{=} \boldsymbol{N}(0,\sigma^2)$$

$$\sigma^2 = 2g'\mathrm{diag}(\pi_\infty)Zg - g'\mathrm{diag}(\pi_\infty)(I + \mathbf{1}\pi_\infty')g.$$

$Z = (I - (P - \mathbf{1}\pi_\infty'))^{-1}$ is called fundamental matrix.

Remarks:

1. The CLT is best understood as follows: Suppose $X_1, X_2, \ldots$ are iid with $\mathbb{E}\{X_i\} = 0$, and variance $\mathrm{var}(X_i) = \sigma^2$. Then

$$\text{SLLN: } \frac{1}{n}\sum_{i=1}^n X_i \xrightarrow{\text{wp1}} 0 \qquad \text{CLT: } \frac{1}{\sqrt{n}}\sum_{i=1}^n X_i \xrightarrow{\mathcal{L}} N(0,\sigma^2)$$

$$\text{Law of iterated logarithm: } \frac{1}{\sqrt{2n\log\log n}}\sum_{i=1}^n X_i \in [-1,1]$$

2. Functional CLT (advanced): Scaled sum of stochastic processes converge to a Gaussian stochastic process

Example 1. Multipath yields Rayleigh fading wireless channel:

$$X(t) = \frac{1}{\sqrt{N}} \sum_{k=1}^{N} A_k \cos(w_c t + \Theta_k)$$

A_k: attenuation of signal on kth path. Assume iid in $[-1, 1]$.
Θ_k: propagation delay due to kth path. Assume iid $U(0, 2\pi)$.

From basic trigonometry $X(t) = X_I \cos w_c t - X_Q \sin w_c t$

$$X_I = \frac{1}{\sqrt{N}} \sum_{k=1}^{N} A_k \cos \Theta_k, \quad X_Q = \frac{1}{\sqrt{N}} \sum_{k=1}^{N} A_k \sin \Theta_k$$

As $N \to \infty$, CLT implies $X_I \sim N(0, \sigma^2)$, $X_Q \sim N(0, \sigma^2)$ where X_I and X_Q are indpt.

Then $X(t)$ expressed in terms of envelope and phase is:

$$X(t) = R \cos(w_c t + \Psi), \quad \text{where } R = \sqrt{X_I^2 + X_Q^2} \sim \text{Rayleigh}$$

and $\Psi \sim U[0, 2\pi]$.

Example 2. Empirical cdf is $F_n(x) = \frac{1}{n} \sum_{k=1}^{n} I(X_k \le x)$.
Central limit theorem says

$$\lim_{n \to \infty} \sqrt{n}(F_n(x) - F(x)) \stackrel{\mathcal{L}}{=} \boldsymbol{N}\big(0, F(x)(1 - F(x))\big) \text{ for each } x$$

Homework: Show that the variance is $F(x)(1 - F(x))$.
Hint: Define $Z_k = I(X_k \le x)$. Show $\text{Var}(Z_k) = F(x) - F^2(x)$.
Since Z_k are iid, $\text{Var}(\frac{1}{\sqrt{n}} \sum_{k=1}^{n} Z_k) = (F(x) - F^2(x))$.

CLT is the basis of statistical significance and confidence tests.

Aside: Big Data Perspective

1. Classical setting: $x_k \in \mathbb{R}^n$ iid. Then as number of time points $N \to \infty$, (so $n \ll N$) following asymptotic results hold:

$$\frac{1}{N}\sum_{k=0}^{N} x_k = \hat{\mu}_N \to \mathbb{E}\{X_k\} = \mu \quad \text{(SLLN)}$$

$$\frac{1}{\sqrt{N}}\sum_{k=0}^{N}(x_k - \mu) \to N(0, \Sigma) \quad \text{(CLT)}$$

Big Data: $n = N$ or $n > N$. Asymptotic statistics dont apply.

Main tool. *Concentration Inequality* (finite N analysis).

Hoeffding inequality. $X_k \in [a, b]$ iid. Then for any $N > 0$,

$$\boxed{P(|\hat{\mu}_N - \mu| > \epsilon) \le 2\exp\left(-\frac{2N\epsilon^2}{(b-a)^2}\right)}$$

Also applied to martingales. By Borel-Cantelli: $\hat{\mu}_N \to \mu$ wp1.

Least Squares: $Y_N = \Psi_{N\times n}\theta_n + \epsilon$.

Classical setting: $n \ll N$ (overdetermined case)

Big data: $n/N =$ const or $n \gg N$ (underdetermined case).

Example 1: $y_k = \theta + \epsilon_k$, $k = 1, \ldots, N$ where $\epsilon_k \in [-0.5, 0.5]$ iid.

$\boxed{\theta_{LS}(N) = \sum_{k=1}^{N} y_k/N}$ Classical: as $N \to \infty$, $\theta_{LS}(N) \to \theta$ w.p1.

Hoeffding inequality: $P(|\theta_{LS}(N) - \theta| > \epsilon) \le 2\exp(-2N\epsilon^2)$.

How much data N for $P(|\theta_{LS}(N) - \theta| > \epsilon) < \alpha$?

$\alpha < 2\exp(-2N\epsilon^2)$. So choose $N > -\frac{1}{2\epsilon^2}\ln(\alpha/2)$.

$\alpha = \epsilon = 0.01$, $N \approx 26500$. (works for any bounded rv ϵ_k)

Machine Learning. Uniform Concentration Inequality

Supervised learning: Given labeled training data
$X_i, Y_i, i = 1, 2, \ldots, n$ and classifier h s.t. $\hat{Y}_i = h(X_i) \in \{0, 1\}$.
Risk of classifier h: $R(h) = P(Y \neq \hat{Y}) = P(Y \neq h(X))$
Empirical Risk (training error): $\hat{R}(h) = \frac{1}{n} \sum_{i=1}^{n} I\big(Y_i \neq h(X_i)\big)$

Define $h^* = \text{argmin}_h \, R(h)$ (oracle); $\hat{h} = \text{argmin}_h \, \hat{R}(h)$ (empirical)
To prove $R(h^*)$ and $R(\hat{h})$ are close, we need uniform concentration:

$$\boxed{P(\sup_{h \in \mathcal{F}} |\hat{R}(h) - R(h)| > \epsilon) \leq \delta_n, \quad \delta_n \downarrow 0} \qquad \text{(UL)}$$

Result: If UL holds, then wp $\geq 1 - \delta_n$, $|R(\hat{h}) - R(h^*)| \leq 2\epsilon$.

Proof: $R(\hat{h}) - R(h^*) = R(\hat{h}) - \hat{R}(\hat{h}) + \hat{R}(\hat{h}) - R(h^*)$

$\leq R(\hat{h}) - \hat{R}(\hat{h}) + \hat{R}(h^*) - R(h^*)$ since $\hat{h}$ minimizes $\hat{R}$

$\implies |R(\hat{h}) - R(h^*)| \leq |R(\hat{h}) - \hat{R}(\hat{h})| + |\hat{R}(h^*) - R(h^*)|$

Uniform law: wp $\geq 1 - \delta_n$, $|R(\hat{h}) - \hat{R}(\hat{h})| \leq \epsilon$, $|\hat{R}(h^*) - R(h^*)| \leq \epsilon$.

Hoeffding: $P(|\hat{R}(h) - R(h)| > \epsilon) \leq 2\, e^{-2n\epsilon^2}$

Vapnik Chervonenkis inequality (uniform conc inequality):

$$\boxed{P(\sup_{h \in \mathcal{F}} |\hat{R}(h) - R(h)| > \epsilon) \leq 8\, \mathcal{S}(\mathcal{F}, n)\, e^{-n\epsilon^2/32}}$$

shattering (growth) function = max # of distinct labels for n points

$\mathcal{S}(\mathcal{F}, n) = \max_{x_1, \ldots, x_n} |\{\big(h(x_1), \ldots, h(x_n)\big),\ h \in \mathcal{F}\}| \leq (n+1)^{V_{\mathcal{F}}}$ (Sauer Lem)

if VC dimension $V_{\mathcal{F}} \stackrel{\text{defn}}{=} \sup\{n : \mathcal{S}(\mathcal{F}, n) = 2^n\}$ is finite

Define $Z_i = (X_i, Y_i)$, $f(Z_i) = I(Y_i \neq h(X_i))$.

Proof: Suppose $Z'_1, \ldots, Z'_n$ are iid ghost samples with same distribution as $Z_1, \ldots, Z_n$.

$$P(\sup_h |\hat{R}(h) - R(h)| > \epsilon) = P(\sup_f |\frac{1}{n}\sum_{i=1}^n f(Z_i) - \mathbb{E}\{f(Z)\}| > \epsilon)$$

$$\leq 2P(\sup_f |\frac{1}{n}\sum_{i=1}^n f(Z_i) - \frac{1}{n}\sum_{i=1}^n f(Z'_i)| > \frac{\epsilon}{2}) \qquad \text{(symmetrization lemma)}$$

$$\leq 2P(\max_{v,v'} |\frac{1}{n}\sum_{i=1}^n (v_i - v'_i)| > \frac{\epsilon}{2}) \qquad \text{where } v_i, v'_i \in \{0,1\}$$

$$\leq 2\sum_{v,v'} P(|\frac{1}{n}\sum_{i=1}^n (v_i - v'_i)| > \frac{\epsilon}{2}) \qquad \text{union bound } P(\max_v A_v) \leq \sum_v P(A_v)$$

Note $v_i - v'_i \in [-1, 1]$ and zero mean. Then Hoeffding inequality

$$\leq 2\sum_{v,v'} 2\exp(-n\epsilon^2/8) = 4\,\mathcal{S}(\mathcal{F}, 2n)\,\exp(-n\epsilon^2/8)$$

Binary Classifier: Let h be a fixed classifier: $I(y \neq h(x)) \in \{0,1\}$. Hoeffding implies $P(|\hat{R}(h) - R(h)| > \epsilon) \leq 2\exp(-2n\epsilon^2)$.

Next, suppose $\mathcal{F} = \{h_\theta, \theta \in \mathbb{R}\}$, $h_\theta(x) = 1, x > \theta$, $h_\theta(x) = 0, x \leq \theta$ be a family of step function classifiers. Then if $x_1 < x_2 < x_3$, $\mathcal{F} = \{(0,0,0), (0,0,1), (0,1,1), (1,1,1)\}$. It can be shown $\mathcal{S}(\mathcal{F}, n) = \sup_{x_1,\ldots,x_n} |\mathcal{F}_{x_1,\ldots,x_n}| = n + 1$. VC inequality

$$P(\sup_h |\hat{R}(h) - R(h)| > \epsilon) \leq 8\,\underbrace{\mathcal{S}(\mathcal{F}, n)}_{n+1}\exp(-n\epsilon^2/32)$$

Also, it can be shown that $V_\mathcal{F} = 1$; recall $\mathcal{S} \leq (n+1)^{V_\mathcal{F}}$.

Symmetrization lemma For $n\epsilon^2 \geq 1$, $f(z_i) \in \{0,1\}$ iid,

$$P(\sup_f |\frac{1}{n}\sum_{i=1}^n f(Z_i) - \mathbb{E}\{f(Z)\}| > \epsilon)$$

$$\leq 2P(\sup_f |\frac{1}{n}\sum_{i=1}^n f(Z_i) - \frac{1}{n}\sum_{i=1}^n f(Z_i')| > \frac{\epsilon}{2})$$

Proof: Call $f^* = \sup_f []$. Then

$$\underbrace{|\frac{1}{n}\sum_{i=1}^n f^*(z_i) - \mathbb{E}\{f^*(z)\}| > \epsilon}_{|a-b|} \text{ and } \underbrace{|\frac{1}{n}\sum_{i=1}^n f^*(z_i') - \mathbb{E}\{f^*(z)\}| \leq \tfrac{\epsilon}{2}}_{|a-c|}$$

$$\implies \underbrace{\frac{1}{n}|\sum_{i=1}^n [f^*(z_i) - f^*(z_i')]|}_{|b-c|} > \tfrac{\epsilon}{2}. \text{ (since } |a-b| \leq |a-c| + |b-c|)$$

$$\text{So } I(|a-b| > \epsilon) \times I(|a-c| \leq \frac{\epsilon}{2}) \leq I(|b-c| > \frac{\epsilon}{2})$$

$$\implies I(|a-b| > \epsilon) \times P'(|a-c| \leq \frac{\epsilon}{2}) \leq P'(|b-c| > \frac{\epsilon}{2})$$

But $P'(|a-c| \leq \frac{\epsilon}{2}) > 1/2$ from Chebyshev inequality.

$$\text{So } P(|a-b| > \epsilon) \leq 2\, P(|b-c| > \frac{\epsilon}{2})$$

Using Chebyshev and assuming $n\epsilon^2 > 1$

$$P'(|a-c| > \frac{\epsilon}{2}) \leq \frac{4\,\text{Var}\,|a-c|}{\epsilon^2} \leq \frac{4}{n\epsilon^2} \times \text{Var}[f(Z_i)] \leq \frac{4}{n\epsilon^2}\frac{1}{4} \leq \frac{1}{n\epsilon^2} \leq \frac{1}{2}$$

Recall Markov $\phi : \mathbb{R} \to \mathbb{R}_+ \implies P(X \geq \epsilon) \leq \frac{1}{\phi(\epsilon)}\mathbb{E}\{\phi(X)\}$, $\epsilon \in \mathbb{R}$

Chebyshev: $\phi(X) = X^2 \implies P(|X-\mu| \geq \epsilon) \leq \frac{1}{\epsilon^2}\mathbb{E}\{|X-\mu|^2\}$

Two famous Probabilistic Inequalities

Markov Inequality: For $\phi : \mathbb{R} \to \mathbb{R}_+$ increasing and $X \in \mathbb{R}$

$$P(X \geq t) \leq \frac{1}{\phi(t)} \mathbb{E}\{\phi(X)\}$$

Proof: $P(X \geq t) = \mathbb{E}\{\mathbf{1}_{X \geq t}\} \leq \mathbb{E}\{\frac{\phi(X)}{\phi(t)} \mathbf{1}_{X \geq t}\} \leq \frac{1}{\phi(t)} \mathbb{E}\{\phi(X)\}$

Example: *Chebyshev ineq*: $P(|X - \mu| \geq t) \leq \mathbb{E}|X - \mu|^2 / t^2$.

Example: $X = \#$ items produced by factory. $\mathbb{E}\{X\} = 500$.

Then $P(X \geq 1000) \leq \mathbb{E}\{X\}/1000 = 0.5$

If we know variance =100. Then

$P(X \geq 1000) \leq \mathbb{E}|X|^2/10^6 = (\mathrm{Var}(X) + \mathbb{E}^2\{X\})/10^6 \approx 0.25.$

Jensen's Inequality: For convex $\phi : \mathbb{R}^n \to \mathbb{R}$ and $X \in \mathbb{R}^n$,

$$\phi(\mathbb{E}\{X\}) \leq \mathbb{E}\{\phi(X)\}$$

Example 1. $\mathbb{E}^2\{X\} \leq \mathbb{E}\{X^2\}$. So $\mathrm{Var}(X) \geq 0$.

Example 2. $\mathbb{E}\{X\} \leq \mathbb{E}\{|X|\}$

Example 3. $\mathbb{E}\{\log X\} \leq \log \mathbb{E}\{X\}$ since log is concave.

KL Divergence

$D(p||q) = \mathbb{E}_p\{\log \frac{p}{q}\} = \int p(x) \log[p(x)/q(x)] dx \geq 0.$

Example 4. $\max_i \mathbb{E}\{X_i\} \leq \mathbb{E}\{\max_i X_i\}$.

Example 5. Geometric mean is smaller than arithmetic mean

$$(x_1 x_2 \cdots x_n)^{1/n} \leq \frac{x_1 + x_2 + \ldots + x_n}{n}$$

Example 6. Cauchy Schwartz: $|\mathbb{E}\{XY\}|^2 \leq \mathbb{E}\{X^2\}\mathbb{E}\{Y^2\}$

Aside. Convex Functions

Definition (i) $f : \mathbb{R}^d \to \mathbb{R}$ is convex if for all $x_1, x_2 \in \mathbb{R}^d$,

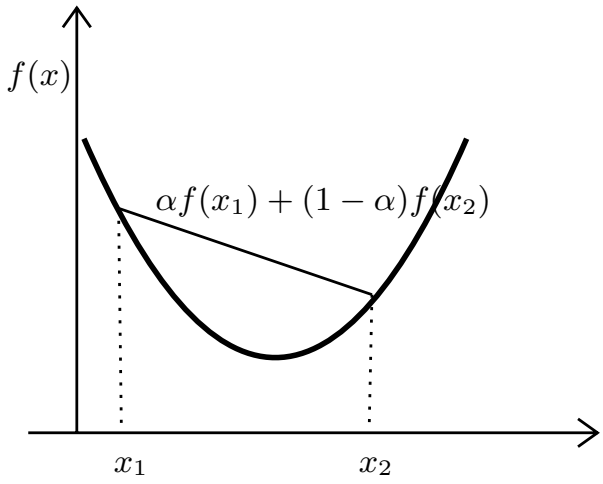
$$f(\alpha x_1 + (1-\alpha)x_2) \leq \alpha f(x_1) + (1-\alpha) f(x_2) \text{ for } \alpha \in [0,1]$$

(ii) Differentiable function f is convex if

$$f(y) \geq f(x) + (y-x)' \nabla f(x), \quad \text{for all } x, y \in \mathbb{R}^d$$

(iii) Twice differentiable function f is convex if Hessian matrix $\nabla^2 f(x)$ is positive semidefinite for all $x \in \mathbb{R}^d$.

Examples of convex functions: Ax, e^x, $-\log(x)$, $x'Qx$ where Q is positive definite, etc.



Proof of Jensen's Inequality:

Convexity implies $f(X) \geq f(a) + (X-a)\nabla f(a)$.

Choose $a = \mathbb{E}\{X\}$. Then

$$f(X) \geq f(\mathbb{E}\{X\}) + (X - \mathbb{E}\{X\})\, \nabla f(\mathbb{E}\{X\})$$

$$\implies \mathbb{E}\{f(X)\} \geq f(\mathbb{E}\{X\}) + \underbrace{\mathbb{E}\{(X - \mathbb{E}\{X\})\}}_{0}\, \nabla f(\mathbb{E}\{X\})$$

Importance Sampling in Simulation

Why? Variance Reduction in Simulation.

Monte-Carlo evaluation of an integral: Simulate N i.i.d. samples of $x_k; k = 1, 2, \ldots, N$ from some pdf p. Then as $N \to \infty$, SLLN

$$\frac{1}{N}\sum_{k=1}^{N} c(x_k) \to \mathbb{E}_p\{c(x)\} = \int_{\mathcal{X}} c(x)p(x)dx \quad \text{with probability 1.}$$

$\frac{1}{N}\sum_{k=1}^{N} c(x_k)$ is an unbiased estimate of $\mathbb{E}_p\{c(x)\}$ for any N; however, the variance of the estimate can be large.

How to reduce the variance of the estimate?

Let $p(x)$ denote a target distribution. Then for any density $q(x)$

$$\mathbb{E}_p\{c(x)\} = \int_{\mathcal{X}} c(x)\frac{p(x)}{\cancel{q(x)}}\cancel{q(x)}\,dx$$

as long as $q(x)$ is chosen so that $p(x)/q(x)$ is finite for all x.
Importance sampling estimate of $\mathbb{E}_p\{c(x)\}$:
(i) sample $x_k; k = 1, 2, \ldots, N$ from importance distribution ("instrumental distribution") $q(x)$ st $p(x)/q(x) < \infty$ for all x.
(ii) Then importance sampling estimate is

$$\boxed{\hat{c}_N = \frac{1}{N}\sum_{k=1}^{N} c(x_k)\frac{p(x_k)}{q(x_k)}, \quad x_k \sim q.}$$

Clearly unbiased estimate.

Result. If the sequence $\{x_k\}$ is i.i.d., then via the strong law of large numbers, as $N \to \infty$, the importance sampled estimate $\hat{c}_N \to \mathbb{E}_p\{c(x)\}$ almost surely. Also, by Central Limit Theorem

$$\lim_{N\to\infty} \sqrt{N}\,(\hat{c}_N - \mathbb{E}_p\{c(x)\}) \stackrel{\mathcal{L}}{=} \boldsymbol{N}(0, \mathrm{Var}_q(c(x))),$$

$$\text{where } \mathrm{Var}_q(c(x)) = \int c^2(x)\frac{p^2(x)}{q(x)}dx - \mathbb{E}^2\{c(x)\}. \quad (3)$$

Example. Rare Event Estimation: For fixed real number α, evaluate

$$c = \mathbb{P}(x > \alpha) \quad \text{where } x \sim \boldsymbol{N}(0,1)$$

Standard MC estimator based on N i.i.d. samples is

$$\hat{c} = \frac{1}{N}\sum_{k=1}^{N} I(x_k > \alpha), \quad x_k \sim \boldsymbol{N}(0,1) \text{ i.i.d.} \quad (4)$$

If α is large we will get very few samples $x_k > \alpha$. Then standard MC has high variance (inaccurate estimate).

Choosing the importance density $q = \boldsymbol{N}(\mu, 1)$, yields

$$\hat{c} = \frac{1}{N}\sum_{k=1}^{N} I(x_k > \alpha)\exp\left(\frac{\mu^2}{2} - \mu x_k\right), \quad x_k \sim \boldsymbol{N}(\mu,1) \text{ i.i.d.}$$

Choose $\mu = \alpha$ in importance sampling estimator. Let us estimate $\mathbb{P}(x > \alpha)$ where $\alpha = 8$, $x \sim \boldsymbol{N}(0,1)$.

Standard MC estimate: Matlab simulation for $N = 50000$ points yields $\hat{c} = 0$ (in double precision) which is useless.

Importance sampling estimate: Matlab simulation for $N = 50000$ points with $\mu = \alpha = 8$, yields estimate $\hat{c} = 6.25 \times 10^{-16}$.

Example 2 (HW): Suppose w_k is a random walk on integers:

$$w_{k+1} = \begin{cases} w_k + 1 & \text{wp } 1/2 \\ w_k - 1 & \text{wp } 1/2 \end{cases}, \quad w_0 = K > 0.$$

Define hitting time $\tau = \min\{k : w_k = 0\}$.

Aim. Estimate $P(\tau \le T)$ for fixed integer T.

Standard MC: Let τ^i be hitting time for ith simulation. Then

$$\hat{\tau} = \frac{1}{N} \sum_{k=1}^{N} I(\tau^i \le T)$$

If $P(\tau \le T)$ then standard MC requires many iterations N.

Importance sampling: Consider asymmetric random walk

$$v_{k+1} = \begin{cases} v_k + 1 & \text{wp } 1-\alpha \\ v_k - 1 & \text{wp } \alpha \end{cases}, \ \alpha \in [0,1], \quad v_0 = K > 0.$$

Define $\sigma^i = \#\{k < \tau^i : v_{k+1} < v_k\}$ = number of times downhill.

Let τ^i be hitting time for ith simulation. Then IS estimate is

$$\tilde{\tau} = \frac{1}{N} \sum_{i=1}^{N} I(\tau^i \le T) \frac{1}{2^{\tau^i} \alpha^{\sigma^i} (1-\alpha)^{\tau^i - \sigma^i}}$$

Example: Choose $T = 20$, $K = 10$, $N = 100$

α	0.5 (MC)	0.6	0.7	0.8	0.9
$\text{Var}(\tilde{\tau})$	0.0268	0.00436	0.00101	0.00058	0.0044

Self-normalized Importance Sampling

Aim. Simulate from un-normalized $p(x)$ by simulating from $q(x)$. Often $p(x)$ is un-normalized and normalization is intractable. Define importance weight $w(x) = \frac{p(x)}{q(x)}$.

Self normalized IS:

$$\boxed{\hat{c}_N = \frac{\sum_{k=1}^N c(x_k)w(x_k)}{\sum_{k=1}^N w(x_k)}, \quad x_k \sim q.}$$

The self-normalized estimate is biased for any finite N.

Suppose $\int_{\mathbb{R}} p(x)dx = \alpha$. For an i.i.d. sequence $\{x_k\}$, from strong law of large numbers

$$\frac{1}{N}\sum_{k=1}^N c(x_k)w(x_k) \to \alpha\, \mathbb{E}_p\{c(x)\}, \quad \frac{1}{N}\sum_{k=1}^N w(x_k) \to \alpha$$

implying that $\hat{c}_N \to \mathbb{E}_p\{c(x)\}$ with probability one.

Derivation:

$$\frac{\int c(x)p(x)dx}{\int p(x)dx} = \frac{\int c(x)\frac{p(x)}{q(x)}q(x)dx}{\int \frac{p(x)}{q(x)}q(x)dx} = \frac{\mathbb{E}_q\{c(x)w(x)\}}{\mathbb{E}_q\{w(x)\}}$$

Many other variance reduction methods. (not covered)

1. Variance Reduction by Conditioning: Uses property that $\mathrm{Var}(X) \geq \mathrm{Var}(\mathbb{E}\{X|Z\})$.
2. Variance Reduction by Stratified Sampling: Uses property that $\mathrm{Var}(X) \geq \mathbb{E}\{\mathrm{Var}(X|Z)\}$.

How to choose importance density to estimate $\mathbb{E}\{c(x)\}$?

$$\text{Importance Sampling:} \quad \text{Var}_q(c(x)) = \int c^2(x)\frac{p^2(x)}{q^2(x)}q(x)dx - \mathbb{E}^2\{c(x)\}$$

$$\text{Standard MC:} \quad \text{Var}_p(c(x)) = \int c^2(x)p(x)dx - \mathbb{E}^2\{c(x)\}.$$

So to estimate $\mathbb{E}\{c(x)\}$, in order to get a variance reduction $\text{Var}_q(c(x)) < \text{Var}_p(c(x))$ requires we choose $q(x)$ st

$$\int c^2(x)\, p(x) \left(1 - \frac{p(x)}{q(x)}\right) dx > 0.$$

Guideline. In regions of x where $c^2(x)p(x)$ is large choose $p(x)/q(x) < 1$; i.e., choose $q(x)$ to be large.

Optimal: $q(x) = c(x)p(x)/\int c(x)p(x)dx$. Then $\text{Var}_q(c(x)) = 0$. Check this. But result is not practical.

Example: Suppose $p(x) = U(a, b)$. Then choose $q(x) \propto c(x)$.

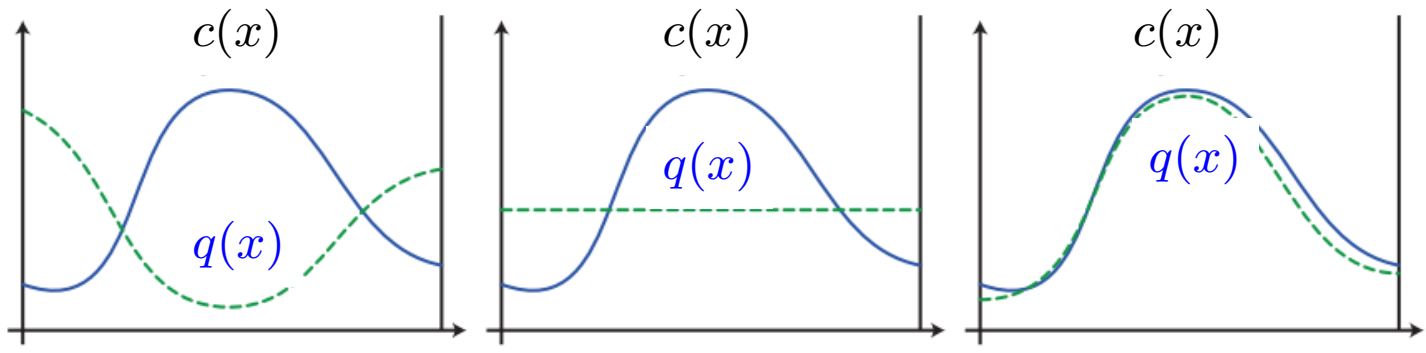


Figure A.2: Comparison of three probability density functions. The PDF on the right provides variance reduction over the uniform PDF in the center. However, using the PDF on the left would significantly increase variance over simple uniform sampling.

Martingales

Definition. Suppose $\mathcal{F}_n = (X_1, \ldots, X_n)$. Then $\{Z_n\}$ is a mtg process wrt $\mathcal{F}_n$ if $\mathbb{E}\{Z_{n+1}|\mathcal{F}_n\} = Z_n$ and $\mathbb{E}|Z_n| < \infty$.

Therefore mtg has constant mean: $\mathbb{E}\{Z_{n+1}\} = \mathbb{E}\{Z_n\} = \mathbb{E}\{Z_0\}$.
If $\mathbb{E}\{Z_{n+1}|\mathcal{F}_n\} \leq Z_n$ then $\{Z_n\}$ is a supermartingale
Martingale difference: $M_n = Z_n - \mathbb{E}\{Z_n|\mathcal{F}_{n-1}\} = Z_n - Z_{n-1}$
Note $\mathbb{E}\{M_n|\mathcal{F}_n\} = 0$ and $\text{cov}\{M_n M_m\} = 0$; so mtg diff is uncorrelated process; more general than iid.
Also $Z_n = Z_0 + \sum_{k=1}^n M_k$.

Martingales are useful (ML, finance, Statistics, OR) because

1. Martingale representation theorem: Every Markov process can be decomposed into martingales.
   Example. Doob Decomposition. Any random process $\{X_n\}$ can be decomposed into $X_n = A_n + Z_n$ where A_n is predictable and Z_n is a mtg.
   How? Clearly $M_n = X_n - \mathbb{E}\{X_n|\mathcal{F}_{n-1}\}$ is a mtg diff.
   So $Z_n = X_0 + \sum_{k=1}^n M_k$ is a mtg
   Let $\Delta_k = \mathbb{E}\{X_k|\mathcal{F}_{k-1}\} - X_{k-1} \implies A_n = \sum_{k=1}^n \Delta_k$ predictable.
   Therefore $X_n = Z_n + A_n$
   Example. Doob-Meyer Decomposition. Supermtg = sum of mtg and decreasing predictable process.
2. **Martingale convergence theorem**: If Z_n is a martingale and $\sup_k \mathbb{E}|Z_n| < \infty$, then Z_n converges to a rv Z_∞ a.s.
   Corollary: Every non-negative martingale converges since $\mathbb{E}|Z_n| = \mathbb{E}\{Z_n\} = \mathbb{E}\{Z_0\} < \infty$.

3. Martingale statistical inference theorems: Law of large numbers Central limit theorem, Concentration inequalities

Martingale Representation of Markov Chain

Suppose $\mathcal{X} = \{e_1, e_2, \ldots, e_X\}$. Then Markov chain with transition prob matrix P can be expressed as linear difference equation

$$x_{k+1} = P' x_k + M_{k+1}, \qquad \text{where } \mathbb{E}\{M_{k+1} | X_1, \ldots, X_k\} = 0$$

M_k is a mtg difference, i.e., $Z_k = \sum_{n=1}^{k} M_n$ is a mtg process.

Proof. Define $M_k = X_{k+1} - P' X_k$. Then
$\mathbb{E}\{M_k | \mathcal{F}_k\} = \mathbb{E}\{X_{k+1} | X_k\} - P' X_k = P' X_k - P' X_k = 0.$
So M_k is a mtg diff process.

Examples of martingales

(i) Random walk: $\{X_n\}$ are iid zero mean. Then

$$Z_n = \sum_{k=1}^{n} X_k \text{ is a mtg.}$$

If X_n are iid with mean 1, then

$$Z_n = \prod_{k=1}^{n} X_k \text{ is a mtg}$$

(iii) Likelihood Ratio Martingales: Suppose X_k iid with pdf g.

$$\text{Then } Z_n = \prod_{k=1}^{n} \frac{f(X_k)}{g(X_k)}, k \geq 1 \text{ is a mtg}$$

Clearly

$$\mathbb{E}\{Z_{n+1}|X_{1:n}\} = \mathbb{E}\{Z_n \frac{f(X_{n+1})}{g(X_{n+1})} \mid X_{1:n}\} = Z_n \int \frac{f(x)}{g(x)} g(x)dx = Z_n$$

(iii) Doob Martingale. Given a sequence of rvs $\{X_n\}$, let $Z_k = \mathbb{E}\{Y|X_1, \ldots, X_k\}$. Then Z_k is a mtg wrt $X_1, \ldots, X_k$. Localization estimate $\mathbb{E}\{\theta|X_1, \ldots, X_k\}$ is a Doob mtg.

(iv) *Wald's Equation.* Suppose $\{X_n\}$ is iid and T is a stopping time. Then

$$\mathbb{E}\{\sum_{k=1}^{T} X_k\} = \mathbb{E}\{T\}\,\mathbb{E}\{X_1\}$$

Ex. $X_n \in \{-1, 1\}$ iid symmetric. Let $T = \min\{n : X_n = 1\}$. What is $\mathbb{E}\{S_T\} = \mathbb{E}\{\sum_{k=1}^{T} X_k\}$?

Intuition: Since we stop after $+1$, $\mathbb{E}\{S_T\} > 0$. This is incorrect.

From Wald Eqn: $\mathbb{E}\{S_T\} = 0$.

(iv) **Bertrand's Ballot Theorem**. A and B contest an election. A receives a votes, B receives b votes and $a > b$. Compute prob that while counting votes, A remains always ahead of B?
Ans:

$$\frac{a-b}{a+b}$$

Let S_k be difference in votes for A vs B after k votes counted.
So $S_n = a - b$ where n is total # votes.
Define $Z_k = S_{n-k}/(n-k)$. Show that Z_k is a mtg.
Define stopping time

$$T = \min\{k : Z_k = 0\}, \qquad \text{or } T = n-1 \text{ otherwise}$$

2 possibilities:
Case 1. A is always ahead: $T = n-1$, $Z_T = Z_{T-1} = S_1 = 1$.
Case 2. A is not always ahead. Then at some point Z_k must be zero; so $Z_T = 0$.
So $\mathbb{E}\{Z_T\} = p \times 1 + (1-p) \times 0$. By optional sampling theorem:

$$p = \mathbb{E}\{Z_T\} = \mathbb{E}\{Z_0\} = \mathbb{E}\{S_n/n\} = \frac{a-b}{a+b}$$

Doob Martingale inequality: Stronger than Chebyshev

$$P(\max_{1\le k\le n}|Z_k|\ge\lambda)\le\frac{\mathbb{E}\{Z_k^2\}}{\lambda^2}$$

Azuma-Hoeffding inequality: Suppose $Z_n=\sum_{k=1}^n M_k+Z_0$ where $\{M_k\}$ is a martingale difference process with bounded differences satisfying $|M_k|\le\Delta_k$ almost surely where Δ_k are finite constants. Then for any $\epsilon>0$

$$\mathbb{P}(Z_n-Z_0\ge\epsilon)\le\exp\left(-\frac{\epsilon^2}{2\sum_{k=1}^n\Delta_k^2}\right)$$

$$\mathbb{P}(Z_n-Z_0\le-\epsilon)\le\exp\left(-\frac{\epsilon^2}{2\sum_{k=1}^n\Delta_k^2}\right).$$

Therefore

$$\boxed{\mathbb{P}(|Z_n-Z_0|\ge\epsilon)\le 2\exp\left(-\frac{\epsilon^2}{2\sum_{k=1}^n\Delta_k^2}\right).}$$

Example: Suppose $M_k\in\{-1,1\}$ are iid. Then $Z_n=\sum_{k=0}^n M_k$ is a mtg. Also $|M_k|\le 1$. So

$$P(Z_n\ge\epsilon)\le\exp(-\frac{\epsilon^2}{2n})$$

Equivalently the empirical mean satisfies

$$P(Z_n/n\ge\epsilon)\le\exp(-\frac{n\epsilon^2}{2})$$