Recap: Multi-dimensional Gaussian density

- $\mathbf{X} = (X_1, \cdots, X_n)^T$ are said to be jointly Gaussian if

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{n}{2}} \, |\Sigma|^{\frac{1}{2}}} \; e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})}$$

- $E\mathbf{X} = \boldsymbol{\mu}$ and $\Sigma_X = \Sigma$.
- The moment generating function is given by

$$M_{\mathbf{X}}(\mathbf{s}) = e^{\mathbf{s}^T \boldsymbol{\mu} \, + \, \frac{1}{2} \mathbf{s}^T \Sigma \, \mathbf{s}}$$

- When X, Y are jointly Gaussian, the joint densityis given by

$$f_{XY}(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \; e^{-\frac{1}{2(1-\rho^2)}\left(\frac{(x-\mu_x)^2}{\sigma_x^2} + \frac{(y-\mu_y)^2}{\sigma_y^2} - \frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}\right)}$$

Recap: Properties

- The multi-dimensional Gaussian density has some important properties.
- If $X_1, \cdots, X_n$ are jointly Gaussian then they are independent if they are uncorrelated.
- If $X_1, \cdots, X_n$ be jointly Gaussian (with zero means) then there is an orthogonal transform $\mathbf{Y} = A\mathbf{X}$ such that $Y_1, \cdots, Y_n$ are jointly Gaussian and independent.
- $X_1, \cdots, X_n$ are jointly Gaussian if and only if $\mathbf{t}^T\mathbf{X}$ is Gaussian for for all non-zero $\mathbf{t} \in \Re^n$.

- Suppose $\mathbf{X} = (X_1, \cdots, X_n)^T$ be jointly Gaussian and let $W = \mathbf{t}^T\mathbf{X}$.
- Let μ_X and Σ_X denote the mean vector and covariance matrix of $\mathbf{X}$. Then

$$\mu_w \triangleq EW = \mathbf{t}^T\mu_X; \quad \sigma_w^2 \triangleq \mathsf{Var}(W) = \mathbf{t}^T\Sigma_X\mathbf{t}$$

- The mgf of W is given by

$$\begin{aligned} M_W(u) &= E\left[e^{uW}\right] = E\left[e^{u\,\mathbf{t}^T\mathbf{X}}\right] \\ &= M_X(u\mathbf{t}) = e^{u\mathbf{t}^T\mu_x + \frac{1}{2}u^2\mathbf{t}^T\Sigma_x\mathbf{t}} \\ &= e^{u\mu_w + \frac{1}{2}u^2\sigma_w^2} \end{aligned}$$

  showing that W is Gaussian
- Shows density of X_i is Gaussian for each i. For example, if we take $\mathbf{t} = (1, 0, 0, \cdots, 0)^T$ then $\mathbf{t}^T\mathbf{X}$ would be X_1.

- Now suppose $W = \mathbf{t}^T \mathbf{X}$ is Gaussian for all $\mathbf{t} \neq 0$.

$$M_W(u) = e^{u\mu_w + \frac{1}{2}u^2\sigma_w^2} = e^{u\ \mathbf{t}^T \mu_X + \frac{1}{2}u^2\ \mathbf{t}^T \Sigma_X \mathbf{t}}$$

- This implies

$$\begin{aligned} E\left[e^{u\ \mathbf{t}^T \mathbf{X}}\right] &= e^{u\ \mathbf{t}^T \mu_X + \frac{1}{2}u^2\ \mathbf{t}^T \Sigma_X \mathbf{t}},\ \forall u \in \Re, \forall \mathbf{t} \in \Re^n,\ \mathbf{t} \neq 0 \\ E\left[e^{\mathbf{t}^T \mathbf{X}}\right] &= e^{\mathbf{t}^T \mu_X + \frac{1}{2}\mathbf{t}^T \Sigma_X \mathbf{t}},\ \forall \mathbf{t} \end{aligned}$$

  This implies $\mathbf{X}$ is jointly Gaussian.
- This is a defining property of multidimensional Gaussian density

- Let $\mathbf{X} = (X_1, \cdots, X_n)^T$ be jointly Gaussian.
- Let A be a $k \times n$ matrix with rank k.
- Then $\mathbf{Y} = A\mathbf{X}$ is jointly Gaussian.
- We will once again show this using the moment generating function.
- Let μ_x and Σ_x denote mean vector and covariance matrix of $\mathbf{X}$. Similarly μ_y and Σ_y for $\mathbf{Y}$
- We have $\mu_y = A\mu_x$ and

$$
\begin{aligned}
\Sigma_y &= E\left[(\mathbf{Y} - \mu_y)(\mathbf{Y} - \mu_y)^T\right] \\
&= E\left[(A(\mathbf{X} - \mu_x))(A(\mathbf{X} - \mu_x))^T\right] \\
&= E\left[A(\mathbf{X} - \mu_x)(\mathbf{X} - \mu_x)^T A^T\right] \\
&= A\, E\left[(\mathbf{X} - \mu_x)(\mathbf{X} - \mu_x)^T\right] A^T = A\Sigma_x A^T
\end{aligned}
$$

- The mgf of $\mathbf{Y}$ is

$$
\begin{aligned}
M_Y(\mathbf{s}) &= E\left[e^{\mathbf{s}^T\mathbf{Y}}\right] \quad (\mathbf{s} \in \Re^k) \\
&= E\left[e^{\mathbf{s}^T A\, \mathbf{X}}\right] \\
&= M_X(A^T\mathbf{s}) \\
&\quad \left(\text{Recall} \quad M_X(\mathbf{t}) = e^{\mathbf{t}^T\mu_x + \frac{1}{2}\mathbf{t}^T\Sigma_x\mathbf{t}}\right) \\
&= e^{\mathbf{s}^T A\mu_x + \frac{1}{2}\mathbf{s}^T A\, \Sigma_x A^T\mathbf{s}} \\
&= e^{\mathbf{s}^T\mu_y + \frac{1}{2}\mathbf{s}^T\Sigma_y\mathbf{s}}
\end{aligned}
$$

  This shows $\mathbf{Y}$ is jointly Gaussian
- Why did we assume A has rank k?

- $\mathbf{X}$ is jointly Gaussian and A is a $k \times n$ matrix with rank k.
- Then $\mathbf{Y} = A\mathbf{X}$ is jointly Gaussian.
- This shows all marginals of $\mathbf{X}$ are gaussian
- For example, if you take A to be

$$A = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \end{bmatrix}$$

then $\mathbf{Y} = (X_1, X_2)^T$

- $\mathbf{X} = (X_1, \cdots, X_n)^T$ is a Gaussian vector when $X_1, \cdots X_n$ are jointly Gaussian.
- They have many interesting special properties.
- X_i are independent iff they are uncorrelated
- $\mathbf{t}^T\mathbf{X}$ being Gaussian for every non-zero $\mathbf{t}$ is a defining property of Gaussian vectors.

- Finding the distribution of a rv by calculating its mgf is useful in many situations.
- Let $X_1, X_2, \cdots$ be iid with mgf $M_X(t)$.
- Let $S_N = \sum_{i=1}^{N} X_i$ where N is a positive integer valued rv which is independent of all X_i.
- We want to find out the distribution of S_N.
- We can calculate mgf of S_N in terms of M_X and distribution of N.
- We can use properties of conditional expectation for this

- The mgf of S_N is $M_{S_N}(t) = E\left[e^{tS_N}\right]$

$$\begin{aligned} E\left[e^{tS_N} \mid N=n\right] &= E\left[e^{t\sum_{i=1}^{N} X_i} \mid N=n\right] \\ &= E\left[e^{t\sum_{i=1}^{n} X_i} \mid N=n\right] \\ &= E\left[e^{t\sum_{i=1}^{n} X_i}\right] = E\left[\prod_{i=1}^{n} e^{tX_i}\right] \\ &= \prod_{i=1}^{n} E\left[e^{tX_i}\right] = \left(M_X(t)\right)^n \end{aligned}$$

- Hence we get

$$E\left[e^{tS_N} \mid N\right] = \left(M_X(t)\right)^N$$

- We can now find mgf of S_N as

$$\begin{aligned} M_{S_N}(t) &= E\left[e^{tS_N}\right] \\ &= E\left[\ E\left[e^{tS_N} \mid N\right]\ \right] \\ &= E\left[\ (M_X(t))^N\ \right] \\ &= \sum_{n=1}^{\infty} (M_X(t))^n\ f_N(n) \\ &= G_N(\ M_X(t)\ ) \end{aligned}$$

  where $G_N(s) = Es^N$ is the generating function of N
- This method is useful for finding distribution of S_N when we can recognize the distribution from its mgf

- We can also find distribution function of S_N directly using the technique of conditional expectations.
- $F_{S_N}(s) = P[S_N \le s]$ and we know how to find probabilities of events using conditional expectation.

$$\begin{aligned} P\left[\sum_{i=1}^{N} X_i \ \le s\right] &= \sum_{n=1}^{\infty} P\left[\sum_{i=1}^{N} X_i \ \le s \mid N = n\right] \ P[N = n] \\ &= \sum_{n=1}^{\infty} P\left[\sum_{i=1}^{n} X_i \ \le s\right] \ P[N = n] \end{aligned}$$

Jensen's Inequality

- Let $g : \Re \rightarrow \Re$ be a convex function. Then

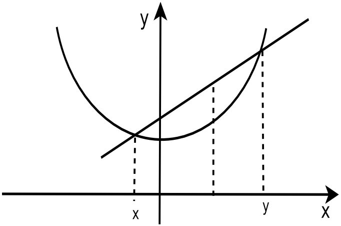
$$g(EX) \leq E[g(X)]$$

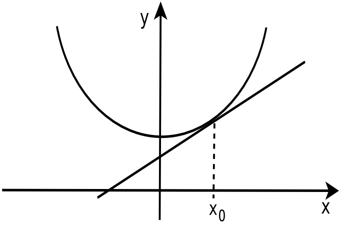
- For example, $(EX)^2 \leq E\left[X^2\right]$
- Function g is convex if

$$g(\alpha x + (1-\alpha)y) \leq \alpha g(x) + (1-\alpha)g(y), \;\; \forall x, y, \;\; \forall 0 \leq \alpha \leq 1$$

- If g is convex, then, given any x_0, exists $\lambda(x_0)$ such that

$$g(x) \geq g(x_0) + \lambda(x_0)(x - x_0), \; \forall x$$





Jensen's Inequality: Proof

- We have: $\forall x_0, \ \exists \lambda(x_0)$ such that

$$g(x) \geq g(x_0) + \lambda(x_0)(x - x_0), \ \forall x$$

- Take $x_0 = EX$ and $x = X(\omega)$. Then

$$g(X(\omega)) \geq g(EX) + \lambda(EX)(X(\omega) - EX), \ \forall \omega$$

- $Y(\omega) \geq Z(\omega), \ \forall \omega \ \Rightarrow \ Y \geq Z \ \Rightarrow EY \geq EZ$
  Hence we get

$$\begin{aligned} g(X) \ &\geq \ g(EX) + \lambda(EX)(X - EX) \\ \Rightarrow \ E[g(X)] \ &\geq \ g(EX) + \lambda(EX) \ E[X - EX] = g(EX) \end{aligned}$$

- This completes the proof

- Consider the set of all mean-zero random variables.
- It is closed under addition and scalar (real number) multiplication.
- $\text{Cov}(X, Y) = E[XY]$ satisfies
  1. $\text{Cov}(X, Y) = \text{Cov}(Y, X)$
  2. $\text{Cov}(X, X) = \text{Var}(X) \geq 0$ and is zero only if $X = 0$
  3. $\text{Cov}(aX, Y) = a\text{Cov}(X, Y)$
  4. $\text{Cov}(X_1 + X_2, Y) = \text{Cov}(X_1, Y) + \text{Cov}(X_2, Y)$
- Thus $\text{Cov}(X, Y)$ is an inner product here.
- The Cauchy-Schwartz inequality ($|\mathbf{x}^T\mathbf{y}| \leq ||\mathbf{x}||\ ||\mathbf{y}||$) gives

$$|\text{Cov}(X, Y)| \leq \sqrt{\text{Cov}(X, X)\ \text{Cov}(Y, Y)} = \sqrt{\text{Var}(X)\ \text{Var}(Y)}$$

- This is same as $|\rho_{XY}| \leq 1$
- A generalization of Cauchy-Schwartz inequality is Holder inequality

Holder Inequality

- For all p, q with $p, q > 1$ and $\frac{1}{p} + \frac{1}{q} = 1$
$$E[|XY|] \leq (E|X|^p)^{\frac{1}{p}} \ (E|Y|^q)^{\frac{1}{q}}$$
(We assume all the expectations are finite)
- If we take $p = q = 2$
$$E[|XY|] \leq \sqrt{E[X^2] \ E[Y^2]}$$
- This is same as Cauchy-Schwartz inequality. This implies $|\rho_{XY}| \leq 1$.
$$\begin{aligned} \left|\text{Cov}(X,Y)\right| &= \left|E[(X-EX)(Y-EY)]\right| \\ &\leq E\left[\left|(X-EX)(Y-EY)\right|\right] \\ &\leq \sqrt{E[(X-EX)^2] \ E[(Y-EY)^2]} \\ &= \sqrt{\text{Var}(X) \ \text{Var}(Y)} \end{aligned}$$

Proof

- First we will show, for $p, q > 1$ and $\frac{1}{p} + \frac{1}{q} = 1$

$$|xy| \le \frac{|x|^p}{p} + \frac{|y|^q}{q}, \quad \forall x, y \in \Re$$

- For $x > 0$, $g(x) = -\log(x)$ is convex because $g''(x) = 1/x^2 \ge 0$, $\forall x$.
- Hence, for all $x_1, x_2 > 0$ and $0 \le t \le 1$,

$$\begin{aligned}
-\log(tx_1 + (1-t)x_2) &\le -t\log(x_1) - (1-t)\log(x_2) \\
\Rightarrow \log(tx_1 + (1-t)x_2) &\ge \log\left(x_1^t \, x_2^{(1-t)}\right) \\
\Rightarrow tx_1 + (1-t)x_2 &\ge x_1^t \, x_2^{(1-t)}
\end{aligned}$$

- We have for all $x_1, x_2 > 0$ and $0 \leq t \leq 1$,

$$tx_1 + (1-t)x_2 \ \geq \ x_1^t \ x_2^{(1-t)}$$

- Take $x_1 = |x|^p$, $x_2 = |y|^q$, $t = \frac{1}{p}$ (and hence $1 - t = \frac{1}{q}$)

$$\begin{aligned} (|x|^p)^{\frac{1}{p}} \ (|y|^q)^{\frac{1}{q}} \quad &\leq \quad \frac{1}{p}\,|x|^p + \frac{1}{q}\,|y|^q \\ \Rightarrow \quad |xy| \quad &\leq \quad \frac{|x|^p}{p} + \frac{|y|^q}{q}, \ \ \forall x, y \end{aligned}$$

$$|xy| \leq \frac{|x|^p}{p} + \frac{|y|^q}{q}, \ \ \forall x, y$$

- Take $x = X(\omega)\,(E|X|^p)^{-\frac{1}{p}}$, $y = Y(\omega)\,(E|Y|^q)^{-\frac{1}{q}}$

$$\frac{|X(\omega)Y(\omega)|}{(E|X|^p)^{\frac{1}{p}}\ (E|Y|^q)^{\frac{1}{q}}} \leq \frac{|X(\omega)|^p\ (E|X|^p)^{-1}}{p} + \frac{|Y(\omega)|^q\ (E|Y|^q)^{-1}}{q}$$

$$\Rightarrow \frac{|XY|}{(E|X|^p)^{\frac{1}{p}}\ (E|Y|^q)^{\frac{1}{q}}} \leq \frac{|X|^p\ (E|X|^p)^{-1}}{p} + \frac{|Y|^q\ (E|Y|^q)^{-1}}{q}$$

$$\Rightarrow \frac{E|XY|}{(E|X|^p)^{\frac{1}{p}}\ (E|Y|^q)^{\frac{1}{q}}} \leq \frac{1}{p} + \frac{1}{q} = 1$$

$$\Rightarrow E|XY| \leq (E|X|^p)^{\frac{1}{p}}\ (E|Y|^q)^{\frac{1}{q}}$$

- **Jensen's Inequality**: If g is convex and EX and $E[g(X)]$ exist
$$g(EX) \leq E[g(X)]$$
- **Holder Inequality**: For $p, q > 1$ and $\frac{1}{p} + \frac{1}{q} = 1$
$$E|XY| \leq (E|X|^p)^{\frac{1}{p}} \ (E|Y|^q)^{\frac{1}{q}}$$
(assuming all expectations exist)
- For $p = q = 2$, the above is Cauchy-Schwartz inequality
- This implies $|\rho_{XY}| \leq 1$
- **Minkowski's Inequality**:
$$(E|X+Y|^r)^{\frac{1}{r}} \leq (E|X|^r)^{\frac{1}{r}} + (E|Y|^r)^{\frac{1}{r}}$$

Chernoff Bounds

- Recall Markov inequality. If h is positive, strictly increasing

$$P[X > a] \; = \; P[h(X) > h(a)] \; \leq \; \frac{E[h(X)]}{h(a)}$$

- Take $h(x) = e^{sx}, \; s > 0$. Then

$$P[X > a] \; \leq \; \frac{E[e^{sX}]}{e^{sa}} \; = \; \frac{M_X(s)}{e^{sa}}, \; \forall s > 0$$

- The RHS is a function of S. We can get a tight bound by using a value of s which minimizes RHS.

Hoeffding Inequality

- Often we need to deal with sums of iid random variables.
- Here is a simple version of an inequality very useful in such situations.
- Let X_i be iid and let $X_i \in [a,\ b],\ \forall i$. Let $EX_i = \mu$

$$P\left[\left|\sum_{i=1}^{n} X_i - n\mu\right| \geq \epsilon\right] \leq 2e^{-\frac{2\epsilon^2}{n(b-a)}},\ \epsilon > 0$$

- Note we do not need knowledge of any moments of X_i to calculate the bound

- Let $X_1, X_2, \cdots$ be iid random variables
- Let $EX_i = \mu$ and let $\text{Var}(X_i) = \sigma^2$
- Define $S_n = \sum_{i=1}^n X_i$. Then

$$ES_n = \sum_{i=1}^n EX_i = n\mu; \quad \text{and} \quad \text{Var}(S_n) = \sum_{i=1}^n \text{Var}(X_i) = n\sigma^2$$

- Consider $\frac{S_n}{n}$, average of $X_1, \cdots, X_n$.

$$\begin{aligned} E\left[\frac{S_n}{n}\right] &= \frac{1}{n} ES_n = \mu, \ \ \forall n \\ \text{Var}\left(\frac{S_n}{n}\right) &= \left(\frac{1}{n}\right)^2 \text{Var}(S_n) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}, \ \ \forall n \end{aligned}$$

Weak Law of large numbers

- X_i are iid, $EX_i = \mu$, $\text{Var}(X_i) = \sigma^2$, $S_n = \sum_{i=1}^n X_i$

$$E\left[\frac{S_n}{n}\right] = \mu; \quad \text{and} \quad \text{Var}\left(\frac{S_n}{n}\right) = \frac{\sigma^2}{n}$$

- As n becomes large, variance of $\frac{S_n}{n}$ becomes close to zero
- $\frac{S_n}{n}$ 'converges' to its expectation, μ, as $n \to \infty$
- By Chebyshev Inequality

$$P\left[\left|\frac{S_n}{n} - \mu\right| \geq \epsilon\right] \leq \frac{\text{Var}(\frac{S_n}{n})}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}, \quad \forall \epsilon > 0$$

- Thus, we get

$$\lim_{n\to\infty} P\left[\left|\frac{S_n}{n} - \mu\right| \geq \epsilon\right] = 0, \quad \forall \epsilon > 0$$

- Known as weak law of large numbers

- Suppose we are tossing a (biased) coin repeatedly
- $X_i = 1$ if i^{th} toss came up head and is zero otherwise.
- $EX_i = p$ where p is the probability of heads.
- $S_n = \sum_{i=1}^n X_i$ is the number of heads in n tosses
- $\frac{S_n}{n}$ is the fraction of heads in n tosses.
- We are saying $\frac{S_n}{n}$ 'converges' to p
- The probability of head is the limiting fraction of heads when you toss the coin infinite times

$$\lim_{n\to\infty} P\left[\left|\frac{S_n}{n} - p\right| \geq \epsilon\right] = 0, \ \ \forall \epsilon > 0$$

- This is true of any event.
- Consider repeatedly performing a random experiment
- X_i be the indicator of event A on i^{th} repetition
- Then $EX_i = P(A), \forall i$
- $\frac{S_n}{n}$ is the fraction of times the event A occurred.
- The fraction of times an event occurs 'converges' to its probability as you repeat the experiment infinite times

- X is a random variable and we want to find EX.
- Make multiple independent observations of X. Call them $X_1, \cdots, X_n$.
- These are called samples of X. $S_n = \sum_{i=1}^{n} X_i$
- $\frac{S_n}{n}$ is the sample mean – average of all samples.
- $\frac{S_n}{n}$ has the same expectation as X but has much smaller variance.
- Sample mean 'converges' to expectation ('population mean')
- This is the principle of sample surveys
- In general one can get an approximate value of expectation of X through simulations/experiments
- Known as Monte Carlo simulations

- X_i are iid, $EX_i = \mu$, $\text{Var}(X_i) = \sigma^2$, $S_n = \sum_{i=1}^n X_i$

$$E\left[\frac{S_n}{n}\right] = \mu; \quad \text{and} \quad \text{Var}\left(\frac{S_n}{n}\right) = \frac{\sigma^2}{n}$$

- As n becomes large, variance of $\frac{S_n}{n}$ becomes close to zero
- We would like to say $\frac{S_n}{n} \to \mu$.
- We need to properly define convergence of a sequence of random variables
- One way of looking at this convergence is

$$\lim_{n\to\infty} P\left[\left|\frac{S_n}{n} - \mu\right| \geq \epsilon\right] = 0, \ \ \forall \epsilon > 0$$

- There are other ways of defining convergence of random variables

- Recall convergence of real number sequences.
- A sequence of real numbers x_n is said to converge to x_0, $x_n \to x_0$, if

$$\forall \epsilon > 0,\ \exists N < \infty,\ s.t.\ |x_n - x_0| \leq \epsilon,\ \forall n \geq N$$

- To show a sequence converges using this definition, we need to know (or guess) the limit.
- Convergent sequences of real numbers satisfy the Cauchy criterion

$$\forall \epsilon > 0,\ \exists N < \infty,\ s.t. |x_n - x_m| \leq \epsilon,\ \forall n, m \geq N$$

- Now consider defining sequence of random variables X_n converging to X_0
- These are not numbers. They are, in fact functions.
- There are different notions of convergence of a sequence of functions to a function.

- Consider a sequence of functions g_n mapping $\Re$ to $\Re$.
- We can say $g_n \to g_0$ if $g_n(x) \to g_0(x), \ \forall x$.
- This is known as point-wise convergence
- Or we can ask for $\int |g_n(x) - g_0(x)|^2 \ dx \to 0$.
- There are multiple notions of convergence that are reasonable for a sequence of functions.
- Thus there would be multiple ways to define convergence of sequence of random variables.

Convergence in Probability

- A sequence of random variables, X_n, is said to **converge in probability** to a random variable X_0 is

$$\lim_{n\to\infty} P\left[|X_n - X_0| > \epsilon\right] = 0, \ \forall\epsilon > 0$$

  This is denoted as $X_n \stackrel{P}{\to} X_0$
- We would mostly be considering convergence to a constant.
- By the definition of limit, the above means

$$\forall\delta > 0, \ \exists N < \infty, \ s.t. \ P\left[|X_n - X_0| > \epsilon\right] < \delta, \ \forall n > N$$

- We only need marginal distributions of individual X_n to decide whether a sequence converges to a constant in probability

Example: Partial sums of iid random variables

- X_i are iid, $EX_i = \mu$, $\text{Var}(X_i) = \sigma^2$, $S_n = \sum_{i=1}^n X_i$
- Then we saw

$$P\left[\left|\frac{S_n}{n} - \mu\right| \geq \epsilon\right] \leq \frac{\sigma^2}{n\epsilon^2}, \quad \forall \epsilon > 0$$

- Hence we have $\frac{S_n}{n} \stackrel{P}{\to} \mu$
- Weak law of large numbers says that sample mean converges in probability to the expectation