18.175: Lecture 9

More large deviations

Scott Sheffield

MIT

Outline

Outline

DeMoivre-Laplace limit theorem

Weak convergence

Legendre transform

Large deviations

DeMoivre-Laplace limit theorem

- Let X_i be i.i.d. random variables. Write $S_n = \sum_{i=1}^n X_n$.

DeMoivre-Laplace limit theorem

- Let X_i be i.i.d. random variables. Write $S_n = \sum_{i=1}^{n} X_n$.
- Suppose each X_i is 1 with probability p and 0 with probability $q = 1 - p$.

DeMoivre-Laplace limit theorem

- Let X_i be i.i.d. random variables. Write $S_n = \sum_{i=1}^n X_n$.
- Suppose each X_i is 1 with probability p and 0 with probability $q = 1 - p$.
- **DeMoivre-Laplace limit theorem:**

$$\lim_{n\to\infty} P\{a \le \frac{S_n - np}{\sqrt{npq}} \le b\} \to \Phi(b) - \Phi(a).$$

DeMoivre-Laplace limit theorem

- Let X_i be i.i.d. random variables. Write $S_n = \sum_{i=1}^n X_n$.
- Suppose each X_i is 1 with probability p and 0 with probability $q = 1 - p$.
- **DeMoivre-Laplace limit theorem:**

$$\lim_{n\to\infty} P\{a \le \frac{S_n - np}{\sqrt{npq}} \le b\} \to \Phi(b) - \Phi(a).$$

- Here $\Phi(b) - \Phi(a) = P\{a \le Z \le b\}$ when Z is a standard normal random variable.

DeMoivre-Laplace limit theorem

- Let X_i be i.i.d. random variables. Write $S_n = \sum_{i=1}^n X_n$.
- Suppose each X_i is 1 with probability p and 0 with probability $q = 1 - p$.
- **DeMoivre-Laplace limit theorem:**

$$\lim_{n\to\infty} P\{a \le \frac{S_n - np}{\sqrt{npq}} \le b\} \to \Phi(b) - \Phi(a).$$

- Here $\Phi(b) - \Phi(a) = P\{a \le Z \le b\}$ when Z is a standard normal random variable.
- $\frac{S_n - np}{\sqrt{npq}}$ describes "number of standard deviations that S_n is above or below its mean".

DeMoivre-Laplace limit theorem

- Let X_i be i.i.d. random variables. Write $S_n = \sum_{i=1}^n X_n$.
- Suppose each X_i is 1 with probability p and 0 with probability $q = 1 - p$.
- **DeMoivre-Laplace limit theorem:**

$$\lim_{n\to\infty} P\{a \le \frac{S_n - np}{\sqrt{npq}} \le b\} \to \Phi(b) - \Phi(a).$$

- Here $\Phi(b) - \Phi(a) = P\{a \le Z \le b\}$ when Z is a standard normal random variable.
- $\frac{S_n - np}{\sqrt{npq}}$ describes "number of standard deviations that S_n is above or below its mean".
- **Proof idea:** use binomial coefficients and Stirling's formula.

DeMoivre-Laplace limit theorem

- Let X_i be i.i.d. random variables. Write $S_n = \sum_{i=1}^n X_n$.
- Suppose each X_i is 1 with probability p and 0 with probability $q = 1 - p$.
- **DeMoivre-Laplace limit theorem:**

$$\lim_{n\to\infty} P\{a \le \frac{S_n - np}{\sqrt{npq}} \le b\} \to \Phi(b) - \Phi(a).$$

- Here $\Phi(b) - \Phi(a) = P\{a \le Z \le b\}$ when Z is a standard normal random variable.
- $\frac{S_n - np}{\sqrt{npq}}$ describes "number of standard deviations that S_n is above or below its mean".
- **Proof idea:** use binomial coefficients and Stirling's formula.
- Question: Does similar statement hold if X_i are i.i.d. from some other law?

DeMoivre-Laplace limit theorem

- Let X_i be i.i.d. random variables. Write $S_n = \sum_{i=1}^n X_n$.
- Suppose each X_i is 1 with probability p and 0 with probability $q = 1 - p$.
- **DeMoivre-Laplace limit theorem:**

$$\lim_{n\to\infty} P\{a \le \frac{S_n - np}{\sqrt{npq}} \le b\} \to \Phi(b) - \Phi(a).$$

- Here $\Phi(b) - \Phi(a) = P\{a \le Z \le b\}$ when Z is a standard normal random variable.
- $\frac{S_n - np}{\sqrt{npq}}$ describes "number of standard deviations that S_n is above or below its mean".
- **Proof idea:** use binomial coefficients and Stirling's formula.
- Question: Does similar statement hold if X_i are i.i.d. from some other law?
- **Central limit theorem:** Yes, if they have finite variance.

Local $p = 1/2$ DeMoivre-Laplace limit theorem

- **Stirling:** $n! \sim n^n e^{-n}\sqrt{2\pi n}$ where $\sim$ means ratio tends to one.

Local $p = 1/2$ DeMoivre-Laplace limit theorem

- **Stirling:** $n! \sim n^n e^{-n}\sqrt{2\pi n}$ where $\sim$ means ratio tends to one.
- **Theorem:** If $2k/\sqrt{2n} \to x$ then $P(S_{2n} = 2k) \sim (\pi n)^{-1/2} e^{-x^2/2}$.

Local $p = 1/2$ DeMoivre-Laplace limit theorem

- **Stirling:** $n! \sim n^n e^{-n}\sqrt{2\pi n}$ where $\sim$ means ratio tends to one.
- **Theorem:** If $2k/\sqrt{2n} \to x$ then $P(S_{2n} = 2k) \sim (\pi n)^{-1/2} e^{-x^2/2}$.
- Recall $P(S_{2n} = 2k) = \binom{2n}{n+k} 2^{-2n} = 2^{-2n} \frac{(2n)!}{(n+k)!(n-k)!}$.

Outline

DeMoivre-Laplace limit theorem

Weak convergence

Legendre transform

Large deviations

Outline

DeMoivre-Laplace limit theorem

Weak convergence

Legendre transform

Large deviations

Weak convergence

- Let X be random variable, X_n a sequence of random variables.

Weak convergence

- Let X be random variable, X_n a sequence of random variables.
- Say X_n **converge in distribution** or **converge in law** to X if $\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$ at all $x \in \mathbb{R}$ at which F_X is continuous.

Weak convergence

- Let X be random variable, X_n a sequence of random variables.
- Say X_n **converge in distribution** or **converge in law** to X if $\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$ at all $x \in \mathbb{R}$ at which F_X is continuous.
- Also say that the $F_n = F_{X_n}$ converge weakly to $F = F_X$.

Weak convergence

- Let X be random variable, X_n a sequence of random variables.
- Say X_n **converge in distribution** or **converge in law** to X if $\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$ at all $x \in \mathbb{R}$ at which F_X is continuous.
- Also say that the $F_n = F_{X_n}$ converge weakly to $F = F_X$.
- **Example:** X_i chosen from $\{-1, 1\}$ with i.i.d. fair coin tosses: then $n^{-1/2} \sum_{i=1}^n X_i$ converges in law to a normal random variable (mean zero, variance one) by DeMoivre-Laplace.

Weak convergence

- Let X be random variable, X_n a sequence of random variables.
- Say X_n **converge in distribution** or **converge in law** to X if $\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$ at all $x \in \mathbb{R}$ at which F_X is continuous.
- Also say that the $F_n = F_{X_n}$ converge weakly to $F = F_X$.
- **Example:** X_i chosen from $\{-1, 1\}$ with i.i.d. fair coin tosses: then $n^{-1/2}\sum_{i=1}^n X_i$ converges in law to a normal random variable (mean zero, variance one) by DeMoivre-Laplace.
- **Example:** If X_n is equal to $1/n$ a.s. then X_n converge weakly to an X equal to 0 a.s. Note that $\lim_{n\to\infty} F_n(0) \neq F(0)$ in this case.

Weak convergence

- Let X be random variable, X_n a sequence of random variables.
- Say X_n **converge in distribution** or **converge in law** to X if $\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$ at all $x \in \mathbb{R}$ at which F_X is continuous.
- Also say that the $F_n = F_{X_n}$ converge weakly to $F = F_X$.
- **Example:** X_i chosen from $\{-1, 1\}$ with i.i.d. fair coin tosses: then $n^{-1/2} \sum_{i=1}^n X_i$ converges in law to a normal random variable (mean zero, variance one) by DeMoivre-Laplace.
- **Example:** If X_n is equal to $1/n$ a.s. then X_n converge weakly to an X equal to 0 a.s. Note that $\lim_{n\to\infty} F_n(0) \neq F(0)$ in this case.
- **Example:** If X_i are i.i.d. then the empirical distributions converge a.s. to law of X_1 (Glivenko-Cantelli).

Weak convergence

- Let X be random variable, X_n a sequence of random variables.
- Say X_n **converge in distribution** or **converge in law** to X if $\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$ at all $x \in \mathbb{R}$ at which F_X is continuous.
- Also say that the $F_n = F_{X_n}$ converge weakly to $F = F_X$.
- **Example:** X_i chosen from $\{-1, 1\}$ with i.i.d. fair coin tosses: then $n^{-1/2}\sum_{i=1}^n X_i$ converges in law to a normal random variable (mean zero, variance one) by DeMoivre-Laplace.
- **Example:** If X_n is equal to $1/n$ a.s. then X_n converge weakly to an X equal to 0 a.s. Note that $\lim_{n\to\infty} F_n(0) \neq F(0)$ in this case.
- **Example:** If X_i are i.i.d. then the empirical distributions converge a.s. to law of X_1 (Glivenko-Cantelli).
- **Example:** Let X_n be the nth largest of $2n+1$ points chosen i.i.d. from fixed law.

Convergence results

- **Theorem:** If $F_n \to F_\infty$, then we can find corresponding random variables Y_n on a common measure space so that $Y_n \to Y_\infty$ almost surely.

Convergence results

- **Theorem:** If $F_n \to F_\infty$, then we can find corresponding random variables Y_n on a common measure space so that $Y_n \to Y_\infty$ almost surely.
- **Proof idea:** Take $\Omega = (0,1)$ and $Y_n = \sup\{y : F_n(y) < x\}$.

Convergence results

- **Theorem:** If $F_n \to F_\infty$, then we can find corresponding random variables Y_n on a common measure space so that $Y_n \to Y_\infty$ almost surely.
- **Proof idea:** Take $\Omega = (0,1)$ and $Y_n = \sup\{y : F_n(y) < x\}$.
- **Theorem:** $X_n \implies X_\infty$ if and only if for every bounded continuous g we have $Eg(X_n) \to Eg(X_\infty)$.

Convergence results

- **Theorem:** If $F_n \to F_\infty$, then we can find corresponding random variables Y_n on a common measure space so that $Y_n \to Y_\infty$ almost surely.
- **Proof idea:** Take $\Omega = (0,1)$ and $Y_n = \sup\{y : F_n(y) < x\}$.
- **Theorem:** $X_n \implies X_\infty$ if and only if for every bounded continuous g we have $Eg(X_n) \to Eg(X_\infty)$.
- **Proof idea:** Define X_n on common sample space so converge a.s., use bounded convergence theorem.

Convergence results

- **Theorem:** If $F_n \to F_\infty$, then we can find corresponding random variables Y_n on a common measure space so that $Y_n \to Y_\infty$ almost surely.
- **Proof idea:** Take $\Omega = (0,1)$ and $Y_n = \sup\{y : F_n(y) < x\}$.
- **Theorem:** $X_n \implies X_\infty$ if and only if for every bounded continuous g we have $Eg(X_n) \to Eg(X_\infty)$.
- **Proof idea:** Define X_n on common sample space so converge a.s., use bounded convergence theorem.
- **Theorem:** Suppose g is measurable and its set of discontinuity points has μ_X measure zero. Then $X_n \implies X_\infty$ implies $g(X_n) \implies g(X)$.

Convergence results

- **Theorem:** If $F_n \to F_\infty$, then we can find corresponding random variables Y_n on a common measure space so that $Y_n \to Y_\infty$ almost surely.
- **Proof idea:** Take $\Omega = (0,1)$ and $Y_n = \sup\{y : F_n(y) < x\}$.
- **Theorem:** $X_n \implies X_\infty$ if and only if for every bounded continuous g we have $Eg(X_n) \to Eg(X_\infty)$.
- **Proof idea:** Define X_n on common sample space so converge a.s., use bounded convergence theorem.
- **Theorem:** Suppose g is measurable and its set of discontinuity points has μ_X measure zero. Then $X_n \implies X_\infty$ implies $g(X_n) \implies g(X)$.
- **Proof idea:** Define X_n on common sample space so converge a.s., use bounded convergence theorem.

Compactness

- **Theorem:** Every sequence F_n of distribution has subsequence converging to right continuous nondecreasing F so that $\lim F_{n(k)}(y) = F(y)$ at all continuity points of F.

Compactness

- **Theorem:** Every sequence F_n of distribution has subsequence converging to right continuous nondecreasing F so that $\lim F_{n(k)}(y) = F(y)$ at all continuity points of F.
- Limit may not be a distribution function.

Compactness

- **Theorem:** Every sequence F_n of distribution has subsequence converging to right continuous nondecreasing F so that $\lim F_{n(k)}(y) = F(y)$ at all continuity points of F.
- Limit may not be a distribution function.
- Need a "tightness" assumption to make that the case. Say μ_n are **tight** if for every ϵ we can find an M so that $\mu_n[-M, M] < \epsilon$ for all n. Define tightness analogously for corresponding real random variables or distributions functions.

Compactness

- **Theorem:** Every sequence F_n of distribution has subsequence converging to right continuous nondecreasing F so that $\lim F_{n(k)}(y) = F(y)$ at all continuity points of F.
- Limit may not be a distribution function.
- Need a "tightness" assumption to make that the case. Say μ_n are **tight** if for every ϵ we can find an M so that $\mu_n[-M, M] < \epsilon$ for all n. Define tightness analogously for corresponding real random variables or distributions functions.
- **Theorem:** Every subsequential limit of the F_n above is the distribution function of a probability measure if and only if the F_n are tight.

Total variation norm

- If we have two probability measures μ and ν we define the **total variation distance** between them is $||\mu - \nu|| := \sup_B |\mu(B) - \nu(B)|$.

Total variation norm

- If we have two probability measures μ and ν we define the **total variation distance** between them is $||\mu - \nu|| := \sup_B |\mu(B) - \nu(B)|$.
- Intuitively, it two measures are close in the total variation sense, then (most of the time) a sample from one measure looks like a sample from the other.

Total variation norm

- If we have two probability measures μ and ν we define the **total variation distance** between them is $||\mu - \nu|| := \sup_B |\mu(B) - \nu(B)|$.
- Intuitively, it two measures are close in the total variation sense, then (most of the time) a sample from one measure looks like a sample from the other.
- Convergence in total variation norm is much stronger than weak convergence.

Outline

DeMoivre-Laplace limit theorem

Weak convergence

Legendre transform

Large deviations

Outline

DeMoivre-Laplace limit theorem

Weak convergence

Legendre transform

Large deviations

Legendre transform

- Define **Legendre transform** (or Legendre dual) of a function $\Lambda : \mathbb{R}^d \to \mathbb{R}$ by

$$\Lambda^*(x) = \sup_{\lambda \in \mathbb{R}^d} \{(\lambda, x) - \Lambda(\lambda)\}.$$

Legendre transform

- Define **Legendre transform** (or Legendre dual) of a function $\Lambda : \mathbb{R}^d \to \mathbb{R}$ by

$$\Lambda^*(x) = \sup_{\lambda \in \mathbb{R}^d} \{(\lambda, x) - \Lambda(\lambda)\}.$$

- Let's describe the Legendre dual geometrically if $d = 1$: $\Lambda^*(x)$ is where tangent line to Λ of slope x intersects the real axis. We can "roll" this tangent line around the convex hull of the graph of Λ, to get all Λ^* values.

Legendre transform

- Define **Legendre transform** (or Legendre dual) of a function $\Lambda : \mathbb{R}^d \to \mathbb{R}$ by

$$\Lambda^*(x) = \sup_{\lambda \in \mathbb{R}^d} \{(\lambda, x) - \Lambda(\lambda)\}.$$

- Let's describe the Legendre dual geometrically if $d = 1$: $\Lambda^*(x)$ is where tangent line to Λ of slope x intersects the real axis. We can "roll" this tangent line around the convex hull of the graph of Λ, to get all Λ^* values.
- Is the Legendre dual always convex?

Legendre transform

- Define **Legendre transform** (or Legendre dual) of a function $\Lambda : \mathbb{R}^d \to \mathbb{R}$ by
$$\Lambda^*(x) = \sup_{\lambda \in \mathbb{R}^d} \{(\lambda, x) - \Lambda(\lambda)\}.$$
- Let's describe the Legendre dual geometrically if $d = 1$: $\Lambda^*(x)$ is where tangent line to Λ of slope x intersects the real axis. We can "roll" this tangent line around the convex hull of the graph of Λ, to get all Λ^* values.
- Is the Legendre dual always convex?
- What is the Legendre dual of x^2? Of the function equal to 0 at 0 and ∞ everywhere else?

Legendre transform

- Define **Legendre transform** (or Legendre dual) of a function $\Lambda : \mathbb{R}^d \to \mathbb{R}$ by

$$\Lambda^*(x) = \sup_{\lambda \in \mathbb{R}^d} \{(\lambda, x) - \Lambda(\lambda)\}.$$

- Let's describe the Legendre dual geometrically if $d = 1$: $\Lambda^*(x)$ is where tangent line to Λ of slope x intersects the real axis. We can "roll" this tangent line around the convex hull of the graph of Λ, to get all Λ^* values.
- Is the Legendre dual always convex?
- What is the Legendre dual of x^2? Of the function equal to 0 at 0 and ∞ everywhere else?
- How are derivatives of Λ and Λ^* related?

Legendre transform

- Define **Legendre transform** (or Legendre dual) of a function $\Lambda : \mathbb{R}^d \to \mathbb{R}$ by

$$\Lambda^*(x) = \sup_{\lambda \in \mathbb{R}^d} \{(\lambda, x) - \Lambda(\lambda)\}.$$

- Let's describe the Legendre dual geometrically if $d = 1$: $\Lambda^*(x)$ is where tangent line to Λ of slope x intersects the real axis. We can "roll" this tangent line around the convex hull of the graph of Λ, to get all Λ^* values.
- Is the Legendre dual always convex?
- What is the Legendre dual of x^2? Of the function equal to 0 at 0 and ∞ everywhere else?
- How are derivatives of Λ and Λ^* related?
- What is the Legendre dual of the Legendre dual of a convex function?

Legendre transform

- Define **Legendre transform** (or Legendre dual) of a function $\Lambda : \mathbb{R}^d \to \mathbb{R}$ by
$$\Lambda^*(x) = \sup_{\lambda \in \mathbb{R}^d} \{(\lambda, x) - \Lambda(\lambda)\}.$$
- Let's describe the Legendre dual geometrically if $d = 1$: $\Lambda^*(x)$ is where tangent line to Λ of slope x intersects the real axis. We can "roll" this tangent line around the convex hull of the graph of Λ, to get all Λ^* values.
- Is the Legendre dual always convex?
- What is the Legendre dual of x^2? Of the function equal to 0 at 0 and ∞ everywhere else?
- How are derivatives of Λ and Λ^* related?
- What is the Legendre dual of the Legendre dual of a convex function?
- What's the higher dimensional analog of rolling the tangent line?

Outline

DeMoivre-Laplace limit theorem

Weak convergence

Legendre transform

Large deviations

Outline

DeMoivre-Laplace limit theorem

Weak convergence

Legendre transform

Large deviations

Recall: moment generating functions

- Let X be a random variable.

Recall: moment generating functions

- Let X be a random variable.
- The **moment generating function** of X is defined by $M(t) = M_X(t) := E[e^{tX}]$.

Recall: moment generating functions

- Let X be a random variable.
- The **moment generating function** of X is defined by $M(t) = M_X(t) := E[e^{tX}]$.

Recall: moment generating functions

- Let X be a random variable.
- The **moment generating function** of X is defined by $M(t) = M_X(t) := E[e^{tX}]$.
- When X is discrete, can write $M(t) = \sum_x e^{tx} p_X(x)$. So $M(t)$ is a weighted average of countably many exponential functions.

Recall: moment generating functions

- Let X be a random variable.
- The **moment generating function** of X is defined by $M(t) = M_X(t) := E[e^{tX}]$.
- When X is discrete, can write $M(t) = \sum_x e^{tx} p_X(x)$. So $M(t)$ is a weighted average of countably many exponential functions.
- When X is continuous, can write $M(t) = \int_{-\infty}^{\infty} e^{tx} f(x) dx$. So $M(t)$ is a weighted average of a continuum of exponential functions.

Recall: moment generating functions

- Let X be a random variable.
- The **moment generating function** of X is defined by $M(t) = M_X(t) := E[e^{tX}]$.
- When X is discrete, can write $M(t) = \sum_x e^{tx} p_X(x)$. So $M(t)$ is a weighted average of countably many exponential functions.
- When X is continuous, can write $M(t) = \int_{-\infty}^{\infty} e^{tx} f(x) dx$. So $M(t)$ is a weighted average of a continuum of exponential functions.
- We always have $M(0) = 1$.

Recall: moment generating functions

- Let X be a random variable.
- The **moment generating function** of X is defined by $M(t) = M_X(t) := E[e^{tX}]$.
- When X is discrete, can write $M(t) = \sum_x e^{tx} p_X(x)$. So $M(t)$ is a weighted average of countably many exponential functions.
- When X is continuous, can write $M(t) = \int_{-\infty}^{\infty} e^{tx} f(x) dx$. So $M(t)$ is a weighted average of a continuum of exponential functions.
- We always have $M(0) = 1$.
- If $b > 0$ and $t > 0$ then $E[e^{tX}] \geq E[e^{t\min\{X,b\}}] \geq P\{X \geq b\} e^{tb}$.

Recall: moment generating functions

- Let X be a random variable.
- The **moment generating function** of X is defined by $M(t) = M_X(t) := E[e^{tX}]$.
- When X is discrete, can write $M(t) = \sum_x e^{tx} p_X(x)$. So $M(t)$ is a weighted average of countably many exponential functions.
- When X is continuous, can write $M(t) = \int_{-\infty}^{\infty} e^{tx} f(x) dx$. So $M(t)$ is a weighted average of a continuum of exponential functions.
- We always have $M(0) = 1$.
- If $b > 0$ and $t > 0$ then $E[e^{tX}] \geq E[e^{t \min\{X,b\}}] \geq P\{X \geq b\} e^{tb}$.
- If X takes both positive and negative values with positive probability then $M(t)$ grows at least exponentially fast in $|t|$ as $|t| \to \infty$.

Recall: moment generating functions for i.i.d. sums

- We showed that if $Z = X + Y$ and X and Y are independent, then $M_Z(t) = M_X(t)M_Y(t)$

Recall: moment generating functions for i.i.d. sums

- We showed that if $Z = X + Y$ and X and Y are independent, then $M_Z(t) = M_X(t)M_Y(t)$
- If $X_1 \ldots X_n$ are i.i.d. copies of X and $Z = X_1 + \ldots + X_n$ then what is M_Z?

Recall: moment generating functions for i.i.d. sums

- We showed that if $Z = X + Y$ and X and Y are independent, then $M_Z(t) = M_X(t)M_Y(t)$
- If $X_1 \ldots X_n$ are i.i.d. copies of X and $Z = X_1 + \ldots + X_n$ then what is M_Z?
- Answer: M_X^n.

Large deviations

- Consider i.i.d. random variables X_i. Can we show that $P(S_n \geq na) \to 0$ exponentially fast when $a > E[X_i]$?

Large deviations

- Consider i.i.d. random variables X_i. Can we show that $P(S_n \geq na) \to 0$ exponentially fast when $a > E[X_i]$?
- Kind of a quantitative form of the weak law of large numbers. The empirical average A_n is *very* unlikely to ϵ away from its expected value (where "very" means with probability less than some exponentially decaying function of n).

General large deviation principle

- More general framework: a *large deviation principle* describes limiting behavior as $n \to \infty$ of family $\{\mu_n\}$ of measures on measure space $(\mathcal{X}, \mathcal{B})$ in terms of a *rate function* I.

General large deviation principle

- More general framework: a *large deviation principle* describes limiting behavior as $n \to \infty$ of family $\{\mu_n\}$ of measures on measure space $(\mathcal{X}, \mathcal{B})$ in terms of a *rate function* I.
- The **rate function** is a lower-semicontinuous map $I : \mathcal{X} \to [0, \infty]$. (The sets $\{x : I(x) \leq a\}$ are closed — rate function called "good" if these sets are compact.)

General large deviation principle

- More general framework: a *large deviation principle* describes limiting behavior as $n \to \infty$ of family $\{\mu_n\}$ of measures on measure space $(\mathcal{X}, \mathcal{B})$ in terms of a *rate function* I.
- The **rate function** is a lower-semicontinuous map $I : \mathcal{X} \to [0, \infty]$. (The sets $\{x : I(x) \le a\}$ are closed — rate function called "good" if these sets are compact.)
- **DEFINITION:** $\{\mu_n\}$ satisfy LDP with rate function I and speed n if for all $\Gamma \in \mathcal{B}$,

$$-\inf_{x \in \Gamma^0} I(x) \le \liminf_{n\to\infty} \frac{1}{n} \log \mu_n(\Gamma) \le \limsup_{n\to\infty} \frac{1}{n} \log \mu_n(\Gamma) \le -\inf_{x \in \overline{\Gamma}} I(x).$$

General large deviation principle

- More general framework: a *large deviation principle* describes limiting behavior as $n \to \infty$ of family $\{\mu_n\}$ of measures on measure space $(\mathcal{X}, \mathcal{B})$ in terms of a *rate function* I.
- The **rate function** is a lower-semicontinuous map $I : \mathcal{X} \to [0, \infty]$. (The sets $\{x : I(x) \le a\}$ are closed — rate function called "good" if these sets are compact.)
- **DEFINITION:** $\{\mu_n\}$ satisfy LDP with rate function I and speed n if for all $\Gamma \in \mathcal{B}$,

$$-\inf_{x \in \Gamma^0} I(x) \le \liminf_{n \to \infty} \frac{1}{n} \log \mu_n(\Gamma) \le \limsup_{n \to \infty} \frac{1}{n} \log \mu_n(\Gamma) \le -\inf_{x \in \overline{\Gamma}} I(x).$$

- **INTUITION:** when "near x" the probability density function for μ_n is tending to zero like $e^{-I(x)n}$, as $n \to \infty$.

General large deviation principle

- More general framework: a *large deviation principle* describes limiting behavior as $n \to \infty$ of family $\{\mu_n\}$ of measures on measure space $(\mathcal{X}, \mathcal{B})$ in terms of a *rate function* I.
- The **rate function** is a lower-semicontinuous map $I : \mathcal{X} \to [0, \infty]$. (The sets $\{x : I(x) \leq a\}$ are closed — rate function called "good" if these sets are compact.)
- **DEFINITION:** $\{\mu_n\}$ satisfy LDP with rate function I and speed n if for all $\Gamma \in \mathcal{B}$,

$$-\inf_{x \in \Gamma^0} I(x) \leq \liminf_{n \to \infty} \frac{1}{n} \log \mu_n(\Gamma) \leq \limsup_{n \to \infty} \frac{1}{n} \log \mu_n(\Gamma) \leq -\inf_{x \in \overline{\Gamma}} I(x).$$

- **INTUITION:** when "near x" the probability density function for μ_n is tending to zero like $e^{-I(x)n}$, as $n \to \infty$.
- **Simple case:** I is continuous, Γ is closure of its interior.

General large deviation principle

- More general framework: a *large deviation principle* describes limiting behavior as $n \to \infty$ of family $\{\mu_n\}$ of measures on measure space $(\mathcal{X}, \mathcal{B})$ in terms of a *rate function* I.
- The **rate function** is a lower-semicontinuous map $I : \mathcal{X} \to [0, \infty]$. (The sets $\{x : I(x) \leq a\}$ are closed — rate function called "good" if these sets are compact.)
- **DEFINITION:** $\{\mu_n\}$ satisfy LDP with rate function I and speed n if for all $\Gamma \in \mathcal{B}$,

$$-\inf_{x \in \Gamma^0} I(x) \leq \liminf_{n \to \infty} \frac{1}{n} \log \mu_n(\Gamma) \leq \limsup_{n \to \infty} \frac{1}{n} \log \mu_n(\Gamma) \leq -\inf_{x \in \overline{\Gamma}} I(x).$$

- **INTUITION:** when "near x" the probability density function for μ_n is tending to zero like $e^{-I(x)n}$, as $n \to \infty$.
- **Simple case:** I is continuous, Γ is closure of its interior.
- **Question:** How would I change if we replaced the measures μ_n by weighted measures $e^{(\lambda n, \cdot)} \mu_n$?

General large deviation principle

- More general framework: a *large deviation principle* describes limiting behavior as $n \to \infty$ of family $\{\mu_n\}$ of measures on measure space $(\mathcal{X}, \mathcal{B})$ in terms of a *rate function* I.
- The **rate function** is a lower-semicontinuous map $I : \mathcal{X} \to [0, \infty]$. (The sets $\{x : I(x) \leq a\}$ are closed — rate function called "good" if these sets are compact.)
- **DEFINITION:** $\{\mu_n\}$ satisfy LDP with rate function I and speed n if for all $\Gamma \in \mathcal{B}$,

$$-\inf_{x \in \Gamma^0} I(x) \leq \liminf_{n \to \infty} \frac{1}{n} \log \mu_n(\Gamma) \leq \limsup_{n \to \infty} \frac{1}{n} \log \mu_n(\Gamma) \leq -\inf_{x \in \overline{\Gamma}} I(x).$$

- **INTUITION:** when "near x" the probability density function for μ_n is tending to zero like $e^{-I(x)n}$, as $n \to \infty$.
- **Simple case:** I is continuous, Γ is closure of its interior.
- **Question:** How would I change if we replaced the measures μ_n by weighted measures $e^{(\lambda n, \cdot)} \mu_n$?
- Replace $I(x)$ by $I(x) - (\lambda, x)$? What is $\inf_x I(x) - (\lambda, x)$?

Cramer's theorem

- Let μ_n be law of empirical mean $A_n = \frac{1}{n}\sum_{j=1}^n X_j$ for i.i.d. vectors $X_1, X_2, \ldots, X_n$ in $\mathbb{R}^d$ with same law as X.

Cramer's theorem

- Let μ_n be law of empirical mean $A_n = \frac{1}{n}\sum_{j=1}^n X_j$ for i.i.d. vectors $X_1, X_2, \ldots, X_n$ in $\mathbb{R}^d$ with same law as X.
- Define **log moment generating function** of X by

$$\Lambda(\lambda) = \Lambda_X(\lambda) = \log M_X(\lambda) = \log \mathbb{E} e^{(\lambda, X)},$$

where $(\cdot, \cdot)$ is inner product on $\mathbb{R}^d$.

Cramer's theorem

- Let μ_n be law of empirical mean $A_n = \frac{1}{n}\sum_{j=1}^n X_j$ for i.i.d. vectors $X_1, X_2, \ldots, X_n$ in $\mathbb{R}^d$ with same law as X.
- Define **log moment generating function** of X by

$$\Lambda(\lambda) = \Lambda_X(\lambda) = \log M_X(\lambda) = \log \mathbb{E} e^{(\lambda, X)},$$

  where $(\cdot,\cdot)$ is inner product on $\mathbb{R}^d$.
- Define **Legendre transform** of Λ by

$$\Lambda^*(x) = \sup_{\lambda \in \mathbb{R}^d} \{(\lambda, x) - \Lambda(\lambda)\}.$$

Cramer's theorem

- Let μ_n be law of empirical mean $A_n = \frac{1}{n}\sum_{j=1}^n X_j$ for i.i.d. vectors $X_1, X_2, \ldots, X_n$ in $\mathbb{R}^d$ with same law as X.
- Define **log moment generating function** of X by

$$\Lambda(\lambda) = \Lambda_X(\lambda) = \log M_X(\lambda) = \log \mathbb{E} e^{(\lambda, X)},$$

  where $(\cdot,\cdot)$ is inner product on $\mathbb{R}^d$.
- Define **Legendre transform** of Λ by

$$\Lambda^*(x) = \sup_{\lambda \in \mathbb{R}^d} \{(\lambda, x) - \Lambda(\lambda)\}.$$

- **CRAMER'S THEOREM:** μ_n satisfy LDP with convex rate function Λ^*.

Thinking about Cramer's theorem

- Let μ_n be law of empirical mean $A_n = \frac{1}{n}\sum_{j=1}^n X_j$.

Thinking about Cramer's theorem

- Let μ_n be law of empirical mean $A_n = \frac{1}{n}\sum_{j=1}^n X_j$.
- **CRAMER'S THEOREM:** μ_n satisfy LDP with convex rate function

$$I(x) = \Lambda^*(x) = \sup_{\lambda \in \mathbb{R}^d} \{(\lambda, x) - \Lambda(\lambda)\},$$

where $\Lambda(\lambda) = \log M(\lambda) = \mathbb{E}e^{(\lambda, X_1)}$.

Thinking about Cramer's theorem

- Let μ_n be law of empirical mean $A_n = \frac{1}{n}\sum_{j=1}^n X_j$.
- **CRAMER'S THEOREM:** μ_n satisfy LDP with convex rate function
$$I(x) = \Lambda^*(x) = \sup_{\lambda \in \mathbb{R}^d} \{(\lambda, x) - \Lambda(\lambda)\},$$
where $\Lambda(\lambda) = \log M(\lambda) = \mathbb{E}e^{(\lambda, X_1)}$.
- This means that for all $\Gamma \in \mathcal{B}$ we have this **asymptotic lower bound** on probabilities $\mu_n(\Gamma)$
$$-\inf_{x \in \Gamma^0} I(x) \le \liminf_{n\to\infty} \frac{1}{n} \log \mu_n(\Gamma),$$
so (up to sub-exponential error) $\mu_n(\Gamma) \ge e^{-n \inf_{x\in\Gamma^0} I(x)}$.

Thinking about Cramer's theorem

- Let μ_n be law of empirical mean $A_n = \frac{1}{n}\sum_{j=1}^n X_j$.
- **CRAMER'S THEOREM:** μ_n satisfy LDP with convex rate function

$$I(x) = \Lambda^*(x) = \sup_{\lambda \in \mathbb{R}^d} \{(\lambda, x) - \Lambda(\lambda)\},$$

  where $\Lambda(\lambda) = \log M(\lambda) = \mathbb{E}e^{(\lambda, X_1)}$.
- This means that for all $\Gamma \in \mathcal{B}$ we have this **asymptotic lower bound** on probabilities $\mu_n(\Gamma)$

$$-\inf_{x \in \Gamma^0} I(x) \leq \liminf_{n \to \infty} \frac{1}{n} \log \mu_n(\Gamma),$$

  so (up to sub-exponential error) $\mu_n(\Gamma) \geq e^{-n \inf_{x \in \Gamma^0} I(x)}$.
- and this **asymptotic upper bound** on the probabilities $\mu_n(\Gamma)$

$$\limsup_{n \to \infty} \frac{1}{n} \log \mu_n(\Gamma) \leq -\inf_{x \in \overline{\Gamma}} I(x),$$

  which says (up to subexponential error) $\mu_n(\Gamma) \leq e^{-n \inf_{x \in \overline{\Gamma}} I(x)}$.

Proving Cramer upper bound

- Recall that $I(x) = \Lambda^*(x) = \sup_{\lambda \in \mathbb{R}^d}\{(\lambda, x) - \Lambda(\lambda)\}$.

Proving Cramer upper bound

- Recall that $I(x) = \Lambda^*(x) = \sup_{\lambda \in \mathbb{R}^d}\{(\lambda, x) - \Lambda(\lambda)\}$.
- For simplicity, assume that Λ is defined for all x (which implies that X has moments of all orders and Λ and Λ^* are strictly convex, and the derivatives of Λ and Λ' are inverses of each other). It is also enough to consider the case X has mean zero, which implies that $\Lambda(0) = 0$ is a minimum of Λ, and $\Lambda^*(0) = 0$ is a minimum of Λ^*.

Proving Cramer upper bound

- Recall that $I(x) = \Lambda^*(x) = \sup_{\lambda \in \mathbb{R}^d}\{(\lambda, x) - \Lambda(\lambda)\}$.
- For simplicity, assume that Λ is defined for all x (which implies that X has moments of all orders and Λ and Λ^* are strictly convex, and the derivatives of Λ and Λ' are inverses of each other). It is also enough to consider the case X has mean zero, which implies that $\Lambda(0) = 0$ is a minimum of Λ, and $\Lambda^*(0) = 0$ is a minimum of Λ^*.
- We aim to show (up to subexponential error) that $\mu_n(\Gamma) \leq e^{-n \inf_{x \in \bar{\Gamma}} I(x)}$.

Proving Cramer upper bound

- Recall that $I(x) = \Lambda^*(x) = \sup_{\lambda \in \mathbb{R}^d}\{(\lambda, x) - \Lambda(\lambda)\}$.
- For simplicity, assume that Λ is defined for all x (which implies that X has moments of all orders and Λ and Λ^* are strictly convex, and the derivatives of Λ and Λ' are inverses of each other). It is also enough to consider the case X has mean zero, which implies that $\Lambda(0) = 0$ is a minimum of Λ, and $\Lambda^*(0) = 0$ is a minimum of Λ^*.
- We aim to show (up to subexponential error) that $\mu_n(\Gamma) \leq e^{-n \inf_{x \in \bar{\Gamma}} I(x)}$.
- If Γ were singleton set $\{x\}$ we could find the λ corresponding to x, so $\Lambda^*(x) = (x, \lambda) - \Lambda(\lambda)$. Note then that
$$\mathbb{E}e^{(n\lambda, A_n)} = \mathbb{E}e^{(\lambda, S_n)} = M_X^n(\lambda) = e^{n\Lambda(\lambda)},$$
and also $\mathbb{E}e^{(n\lambda, A_n)} \geq e^{n(\lambda, x)}\mu_n\{x\}$. Taking logs and dividing by n gives $\Lambda(\lambda) \geq \frac{1}{n}\log \mu_n + (\lambda, x)$, so that $\frac{1}{n}\log \mu_n(\Gamma) \leq -\Lambda^*(x)$, as desired.

Proving Cramer upper bound

- Recall that $I(x) = \Lambda^*(x) = \sup_{\lambda \in \mathbb{R}^d}\{(\lambda, x) - \Lambda(\lambda)\}$.
- For simplicity, assume that Λ is defined for all x (which implies that X has moments of all orders and Λ and Λ^* are strictly convex, and the derivatives of Λ and Λ' are inverses of each other). It is also enough to consider the case X has mean zero, which implies that $\Lambda(0) = 0$ is a minimum of Λ, and $\Lambda^*(0) = 0$ is a minimum of Λ^*.
- We aim to show (up to subexponential error) that $\mu_n(\Gamma) \leq e^{-n \inf_{x \in \bar{\Gamma}} I(x)}$.
- If Γ were singleton set $\{x\}$ we could find the λ corresponding to x, so $\Lambda^*(x) = (x, \lambda) - \Lambda(\lambda)$. Note then that
$$\mathbb{E}e^{(n\lambda, A_n)} = \mathbb{E}e^{(\lambda, S_n)} = M_X^n(\lambda) = e^{n\Lambda(\lambda)},$$
and also $\mathbb{E}e^{(n\lambda, A_n)} \geq e^{n(\lambda, x)}\mu_n\{x\}$. Taking logs and dividing by n gives $\Lambda(\lambda) \geq \frac{1}{n}\log \mu_n + (\lambda, x)$, so that $\frac{1}{n}\log \mu_n(\Gamma) \leq -\Lambda^*(x)$, as desired.
- General Γ: cut into finitely many pieces, bound each piece?

Proving Cramer lower bound

- Recall that $I(x) = \Lambda^*(x) = \sup_{\lambda \in \mathbb{R}^d}\{(\lambda, x) - \Lambda(\lambda)\}$.

Proving Cramer lower bound

- Recall that $I(x) = \Lambda^*(x) = \sup_{\lambda \in \mathbb{R}^d}\{(\lambda, x) - \Lambda(\lambda)\}$.
- We aim to show that asymptotically $\mu_n(\Gamma) \geq e^{-n \inf_{x \in \Gamma^0} I(x)}$.

Proving Cramer lower bound

- Recall that $I(x) = \Lambda^*(x) = \sup_{\lambda \in \mathbb{R}^d}\{(\lambda, x) - \Lambda(\lambda)\}$.
- We aim to show that asymptotically $\mu_n(\Gamma) \geq e^{-n \inf_{x \in \Gamma^0} I(x)}$.
- It's enough to show that for each given $x \in \Gamma^0$, we have that asymptotically $\mu_n(\Gamma) \geq e^{-nI(x)}$.

Proving Cramer lower bound

- Recall that $I(x) = \Lambda^*(x) = \sup_{\lambda \in \mathbb{R}^d}\{(\lambda, x) - \Lambda(\lambda)\}$.
- We aim to show that asymptotically $\mu_n(\Gamma) \geq e^{-n \inf_{x \in \Gamma^0} I(x)}$.
- It's enough to show that for each given $x \in \Gamma^0$, we have that asymptotically $\mu_n(\Gamma) \geq e^{-nI(x)}$.
- Idea is to weight law of each X_i by $e^{(\lambda,x)}$ to get a new measure whose expectation is in the interior of x. In this new measure, A_n is "typically" in Γ for large Γ, so the probability is of order 1.

Proving Cramer lower bound

- Recall that $I(x) = \Lambda^*(x) = \sup_{\lambda \in \mathbb{R}^d}\{(\lambda, x) - \Lambda(\lambda)\}$.
- We aim to show that asymptotically $\mu_n(\Gamma) \geq e^{-n \inf_{x \in \Gamma^0} I(x)}$.
- It's enough to show that for each given $x \in \Gamma^0$, we have that asymptotically $\mu_n(\Gamma) \geq e^{-nI(x)}$.
- Idea is to weight law of each X_i by $e^{(\lambda,x)}$ to get a new measure whose expectation is in the interior of x. In this new measure, A_n is "typically" in Γ for large Γ, so the probability is of order 1.
- But by how much did we have to modify the measure to make this typical? Aren't we weighting the law of A_n by about $e^{-nI(x)}$ near x?