Stochastic Processes

Problems on Convergence of Sequences of Random Variables

Karthik P. N.
Assistant Professor, Department of AI
Email: pnkarthik@ai.iith.ac.in

07/08 February 2025

Problems

- Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of independent random variables with

$$\mathbb{P}\left(X_n = \frac{1}{2}\left(1 - \frac{1}{n}\right)\right) = \mathbb{P}\left(X_n = \frac{1}{2}\left(1 + \frac{1}{n}\right)\right) = \frac{1}{2}.$$

  - Determine whether $\{X_n\}_{n=1}^{\infty}$ converges in mean-squared sense.
  - Determine whether $\{X_n\}_{n=1}^{\infty}$ converges in almost-sure sense.

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Problems

- Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of independent random variables with

$$\mathbb{P}\left(X_n = \frac{1}{2}\left(1 - \frac{1}{n}\right)\right) = \mathbb{P}\left(X_n = \frac{1}{2}\left(1 + \frac{1}{n}\right)\right) = \frac{1}{2}\left(1 - \frac{1}{n}\right), \qquad \mathbb{P}(X_n = 1) = \frac{1}{n}.$$

  - Determine whether $\{X_n\}_{n=1}^{\infty}$ converges in distribution.
  - Determine whether $\{X_n\}_{n=1}^{\infty}$ converges in almost-sure sense.

Problems

- For each $n \in \mathbb{N}$, let
$$X_n \sim \mathcal{N}\left(0, \frac{1}{n}\right).$$
Determine the limit and the forms of convergence.

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Problems

- Let $X_1, X_2, \cdots \overset{\text{i.i.d.}}{\sim} \mathrm{Exp}(1)$.
  For each $n \in \mathbb{N}$, let
  $$Y_n = \max\{X_1, \ldots, X_n\}.$$
  - Compute the CDF of Y_n.
  - Show that
    $$\lim_{n\to\infty} \mathbb{P}(Y_n \le a \log n) = \begin{cases} 0, & a < 1, \\ 1, & a > 1. \end{cases}$$
    What can you conclude from this result about the sequence $\left\{ \frac{Y_n}{\log n} \right\}_{n=1}^{\infty}$?

Problems

- Let $W_1, W_2, \cdots \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$ for some fixed $\sigma > 0$. Let $X_0 = 0$, and for each $n \in \mathbb{N}$, let

$$X_{n+1} = \frac{X_n + W_{n+1}}{2}.$$

Prove that $\{X_n\}_{n=1}^{\infty}$ converges in distribution.

Empirical Risk Minimisation in Computer Vision

In a computer vision application, you have trained a neural network to detect objects in images by predicting bounding boxes. Each bounding box is represented by (x, y) coordinates for its center (along with width and height). After deployment, you suspect there might be a systematic bias in the predicted x-coordinate of the bounding box centers. For instance, the model might be consistently shifting bounding boxes slightly to the left or right. Let

$$X_i = (\text{predicted } x_i) - (\text{ground truth } x_i) \quad \text{for } i\text{th image.}$$

Empirical Risk Minimisation in Computer Vision

In a computer vision application, you have trained a neural network to detect objects in images by predicting bounding boxes. Each bounding box is represented by (x, y) coordinates for its center (along with width and height). After deployment, you suspect there might be a systematic bias in the predicted x-coordinate of the bounding box centers. For instance, the model might be consistently shifting bounding boxes slightly to the left or right. Let

$$X_i = (\text{predicted } x_i) - (\text{ground truth } x_i) \quad \text{for } i\text{th image.}$$

A reasonable model: $X_1, X_2, \cdots \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, 1)$, where μ is unknown model parameter

Empirical Risk Minimisation in Computer Vision

In a computer vision application, you have trained a neural network to detect objects in images by predicting bounding boxes. Each bounding box is represented by (x, y) coordinates for its center (along with width and height). After deployment, you suspect there might be a systematic bias in the predicted x-coordinate of the bounding box centers. For instance, the model might be consistently shifting bounding boxes slightly to the left or right. Let

$$X_i = (\text{predicted } x_i) - (\text{ground truth } x_i) \quad \text{for } i\text{th image.}$$

A reasonable model: $X_1, X_2, \cdots \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, 1)$, where μ is unknown model parameter

- Suppose that model parameter is θ. Define loss under θ as

$$\ell(X_1, \ldots, X_n; \theta) = \frac{1}{n} \sum_{i=1}^{n} (X_i - \theta)^2.$$

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Empirical Risk Minimisation in Computer Vision

In a computer vision application, you have trained a neural network to detect objects in images by predicting bounding boxes. Each bounding box is represented by (x, y) coordinates for its center (along with width and height). After deployment, you suspect there might be a systematic bias in the predicted x-coordinate of the bounding box centers. For instance, the model might be consistently shifting bounding boxes slightly to the left or right. Let

$$X_i = (\text{predicted } x_i) - (\text{ground truth } x_i) \quad \text{for } i\text{th image.}$$

A reasonable model: $X_1, X_2, \cdots \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, 1)$, where μ is unknown model parameter

- Suppose that model parameter is θ. Define loss under θ as

$$\ell(X_1, \ldots, X_n; \theta) = \frac{1}{n} \sum_{i=1}^{n} (X_i - \theta)^2.$$

Determine the model parameter $\widehat{\theta}_n$ that minimizes the loss and hence fits the data best.

Empirical Risk Minimisation in Computer Vision

In a computer vision application, you have trained a neural network to detect objects in images by predicting bounding boxes. Each bounding box is represented by (x, y) coordinates for its center (along with width and height). After deployment, you suspect there might be a systematic bias in the predicted x-coordinate of the bounding box centers. For instance, the model might be consistently shifting bounding boxes slightly to the left or right. Let

$$X_i = (\text{predicted } x_i) - (\text{ground truth } x_i) \quad \text{for } i\text{th image.}$$

A reasonable model: $X_1, X_2, \cdots \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, 1)$, where μ is unknown model parameter

- Suppose that model parameter is θ. Define loss under θ as

$$\ell(X_1, \ldots, X_n; \theta) = \frac{1}{n} \sum_{i=1}^{n} (X_i - \theta)^2.$$

  Determine the model parameter $\widehat{\theta}_n$ that minimizes the loss and hence fits the data best.

- Show that $\widehat{\theta}_n \xrightarrow{\text{a.s.}} \mu$.

Linear Regression in LLMs

Consider a large language model that produces token embeddings at some intermediate layer (say L). Each token embedding is a vector in $\mathbb{R}^d$ for some large d. Focus on the first coordinate X of the vector embedding, and suppose that you observe, across many contexts, the value of X. Denote these values by $X_1, X_2, \ldots$.

Linear Regression in LLMs

Consider a large language model that produces token embeddings at some intermediate layer (say L). Each token embedding is a vector in $\mathbb{R}^d$ for some large d. Focus on the first coordinate X of the vector embedding, and suppose that you observe, across many contexts, the value of X. Denote these values by $X_1, X_2, \ldots$.

A plausible model: $X_1, X_2, \cdots \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$

Linear Regression in LLMs

Consider a large language model that produces token embeddings at some intermediate layer (say L). Each token embedding is a vector in $\mathbb{R}^d$ for some large d. Focus on the first coordinate X of the vector embedding, and suppose that you observe, across many contexts, the value of X. Denote these values by $X_1, X_2, \ldots$.

A plausible model: $X_1, X_2, \cdots \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$

Suppose that the first coordinate of the vector embedding of layer L has a strong bearing on some downstream performance metric Y (e.g., how “on-topic” the generated text is). We want to capture this dependence.

Linear Regression in LLMs

Consider a large language model that produces token embeddings at some intermediate layer (say L). Each token embedding is a vector in $\mathbb{R}^d$ for some large d. Focus on the first coordinate X of the vector embedding, and suppose that you observe, across many contexts, the value of X. Denote these values by $X_1, X_2, \ldots$.

A plausible model: $X_1, X_2, \cdots \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$

Suppose that the first coordinate of the vector embedding of layer L has a strong bearing on some downstream performance metric Y (e.g., how “on-topic” the generated text is). We want to capture this dependence.

We wish to model the relationship between X and Y is given by

$$Y_i = \beta X_i + \varepsilon_i,$$

where $\beta \in \mathbb{R}$ is unknown, and $\varepsilon_1, \varepsilon_2 \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$, independent of $X_1, X_2, \ldots,$ and σ^2 is known

Linear Regression in LLMs

$$Y_i = \beta X_i + \varepsilon_i,$$

$\beta \in \mathbb{R}$ unknown, $\quad \varepsilon_1, \varepsilon_2 \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$ (independent of $X_1, X_2, \ldots$), $\quad \sigma^2$ known

- Considering n data points $\{(X_i, Y_i)\}_{i=1}^{n}$, write down the expression for the ℓ_2 loss (a.k.a. squared loss) between input and output.

Linear Regression in LLMs

$$Y_i = \beta X_i + \varepsilon_i,$$

$\beta \in \mathbb{R}$ unknown, $\quad \varepsilon_1, \varepsilon_2 \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$ (independent of $X_1, X_2, \ldots$), $\quad \sigma^2$ known

- Considering n data points $\{(X_i, Y_i)\}_{i=1}^n$, write down the expression for the ℓ_2 loss (a.k.a. squared loss) between input and output.
- Determine the value of model parameter, say $\widehat{\beta}_n$, that minimizes the ℓ_2 loss.

Linear Regression in LLMs

$$Y_i = \beta X_i + \varepsilon_i,$$

$\beta \in \mathbb{R}$ unknown, $\quad \varepsilon_1, \varepsilon_2 \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$ (independent of $X_1, X_2, \ldots$), $\quad \sigma^2$ known

- Considering n data points $\{(X_i, Y_i)\}_{i=1}^n$, write down the expression for the ℓ_2 loss (a.k.a. squared loss) between input and output.
- Determine the value of model parameter, say $\widehat{\beta}_n$, that minimizes the ℓ_2 loss.
- Prove that

$$\widehat{\beta}_n \xrightarrow{\text{a.s.}} \beta.$$