S&DS 242/542: Theory of Statistics

Lecture 2: Probability review I

Teaching staff and office hours

Course manager: Bella Bao (bella.bao@yale.edu)

Course TAs:

Johanna Dammann
Xinyang Hu
Langchen Liu
Linghai Liu
Neil Mathew
Selma Mazioud
Max Lovig
Matthew Ross
Ivan Sinyavin
Arjun Verma
Brian Xiang
Grant Zhang
Bronson Zhou

My office hours are Mondays 4-5PM in KT1101.

TA office hours will start next week, with times/locations posted to the course webpage.

S&DS DSAC groupme

The S&DS DSAC has created a new groupme to use as a hub for communication for S&DS majors, certificates, and interested students. Join here for news about merch handouts, student events, and class/bluebooking advice:

`https://groupme.com/join_group/103331993/SFgZGZMT`

Random variables and distributions

Throughout our course, we will model data using random variables.

- **Discrete random variables**
  - Can take a finite or countably infinite number of values
  - Describe categorical data, e.g. outcome of a dice roll

$$X \in \{1, 2, 3, 4, 5, 6\}$$

  and count data, e.g. number of students in S&DS 242

$$X \in \{0, 1, 2, 3, \ldots\}$$

- **Continuous random variables**
  - Can take a continuum of values on the real line, e.g.

$$X \in \mathbb{R} \text{ or } X \in (0, \infty) \text{ or } X \in (0, 1)$$

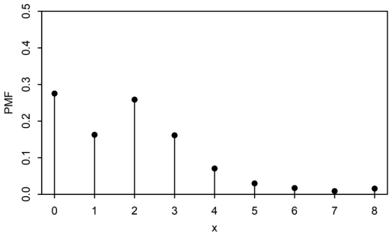
  - Describe continuous data, e.g. height or weight of a person

The *distribution* or *law* of X describes $\mathbb{P}[X \in A]$ for any set $A \subseteq \mathbb{R}$.

Probability mass functions

For discrete X, its distribution may be specified by its **probability mass function (PMF)**: For each possible value x that X can take,

$$f(x) = \mathbb{P}[X = x]$$



Then for any set of values A, $\mathbb{P}[X \in A] = \sum_{x \in A} f(x)$.

If $\mathcal{X}$ is the space of all possible values for X, then $\sum_{x \in \mathcal{X}} f(x) = 1$.

All figures are from *Introduction to Probability* by Blitzstein and Hwang

Bernoulli and Binomial distributions

Example: A **Bernoulli** random variable $X \sim \text{Bernoulli}(p)$ takes two possible values $\{0, 1\}$. Its PMF is

$$f(x) = \begin{cases} p & \text{if } x = 1 \\ 1 - p & \text{if } x = 0 \end{cases}$$

Example: A **Binomial** random variable $X \sim \text{Binomial}(n, p)$ takes values in $\{0, 1, 2, \ldots, n\}$. Its PMF is

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x} \text{ for each } x \in \{0, 1, 2, \ldots, n\}$$

If $X_1, \ldots, X_n \overset{IID}{\sim} \text{Bernoulli}(p)$, then

$$X_1 + \ldots + X_n \sim \text{Binomial}(n, p)$$

This representation is often more useful than the PMF.

Poisson and Negative Binomial distributions

Example: A **Poisson** random variable $X \sim \text{Poisson}(\lambda)$ takes nonnegative integer values. Its PMF is

$$f(x) = \frac{e^{-\lambda}\lambda^x}{x!} \text{ for each } x \in \{0, 1, 2, \ldots\}$$

Example: A **Negative Binomial** random variable $X \sim \text{NegBin}(r, p)$ also takes nonnegative integer values. Its PMF is

$$f(x) = \binom{x + r - 1}{r - 1} p^r (1 - p)^x \text{ for each } x \in \{0, 1, 2, \ldots\}$$

This represents the number of failures before the r^{th} success in a sequence of independent Bernoulli(p) trials.

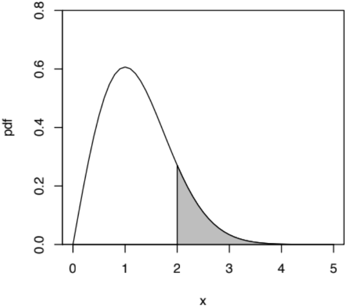
Both are common models for count data. NegBin(r, p) has two parameters, allowing for more flexible modeling of mean/variance.

Probability density functions

For continuous X, its distribution may be specified by its **probability density function (PDF)**: a function $f(x)$ such that for any set $A \subseteq \mathbb{R}$,

$$\mathbb{P}[X \in A] = \int_A f(x)dx$$

The integral over the whole real line is $\int_{-\infty}^{\infty} f(x)dx = 1$.



Normal and Gamma distributions

Example: A **Normal** (or Gaussian) random variable $X \sim \mathcal{N}(\mu, \sigma^2)$ takes any real value. Its PDF is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

The normal distribution appears ubiquitously throughout statistics, due to the Central Limit Theorem.

Example: A **Gamma** random variable $X \sim \text{Gamma}(\alpha, \beta)$ takes positive real values. Its PDF is

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \text{ for } x > 0$$

Here $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx$. (This extends the factorial function to all positive reals, with $\Gamma(n) = (n-1)!$ for positive integers n.)

Chi-squared distribution

Example: A **chi-squared** random variable $X \sim \chi^2(n)$ is a special case of a Gamma random variable, Gamma$(n/2, 1/2)$. Its PDF is

$$f(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} e^{-x/2} \text{ for } x > 0$$

The parameter n is called the "degrees of freedom".

If $X_1, \ldots, X_n \overset{IID}{\sim} \mathcal{N}(0, 1)$, then

$$X_1^2 + \ldots + X_n^2 \sim \chi^2(n)$$

(We'll show this next class.)

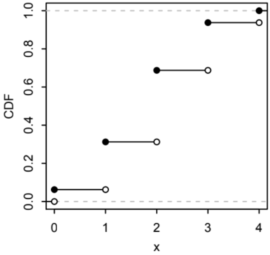
These representations are more useful than the PDF.

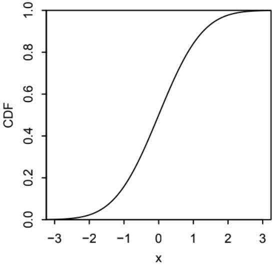
Cumulative distribution functions

The distribution of X can also be specified by its **cumulative distribution function (CDF)**

$$F(x) = \mathbb{P}[X \leq x]$$

Discrete: $F(x) = \sum_{y:\, y \leq x} f(y)$ Continuous: $F(x) = \int_{-\infty}^{x} f(y)dy$





When X is continuous, the derivative of $F(x)$ is the PDF $f(x)$.

Quantile functions

By definition, the CDF $F(x)$ is non-decreasing:

$$F(x) \leq F(y) \text{ for all } x \leq y$$

If $F : \mathbb{R} \to (0,1)$ is continuous and *strictly* increasing, then it has a continuous inverse function $F^{-1} : (0,1) \to \mathbb{R}$, which satisfies

$$F(x) = t \iff F^{-1}(t) = x$$

F^{-1} is called the **quantile function** of X. For any $t \in (0,1)$, $x = F^{-1}(t)$ is the t^{th} **quantile** of the distribution of X, satisfying

$$\mathbb{P}[X \leq x] = t$$

$F^{-1}(0.5)$ is the median, $F^{-1}(0.25)$ and $F^{-1}(0.75)$ are the first and third quartiles.

Expectation

The **expectation** or **mean** of X is its "average value".

If X is discrete with PMF $f(x)$, then

$$\mathbb{E}[X] = \sum_{x \in \mathcal{X}} x \cdot f(x)$$

If X is continuous with PDF $f(x)$, then analogously,

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x \cdot f(x) dx$$

More generally, for any function $g : \mathbb{R} \to \mathbb{R}$, the mean of $g(X)$ is

$$\mathbb{E}[g(X)] = \sum_{x \in \mathcal{X}} g(x) \cdot f(x) \quad \text{or} \quad \mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) \cdot f(x) dx$$

Poisson expectation

Example: Let $X \sim \text{Poisson}(\lambda)$.

$$\mathbb{E}[X] = \sum_{x=0}^{\infty} x \cdot \frac{e^{-\lambda}\lambda^x}{x!}$$

$$= \sum_{x=1}^{\infty} \frac{e^{-\lambda}\lambda^x}{(x-1)!} \qquad (y = x-1)$$

$$= \sum_{y=0}^{\infty} \frac{e^{-\lambda}\lambda^{y+1}}{y!} = e^{-\lambda}\lambda \cdot \underbrace{\sum_{y=0}^{\infty} \frac{\lambda^y}{y!}}_{= e^{\lambda}} = \lambda$$

Gamma expectation

Example: Let $X \sim \text{Gamma}(\alpha, \beta)$.

$$\mathbb{E}[X] = \int_0^\infty x \cdot \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}\, dx \qquad (y = \beta x)$$

$$= \int_0^\infty \frac{\beta^\alpha}{\Gamma(\alpha)} \left(\frac{y}{\beta}\right)^\alpha e^{-y} \frac{1}{\beta} dy$$

$$= \frac{1}{\beta \Gamma(\alpha)} \underbrace{\int_0^\infty \underbrace{y^\alpha}_{u} \underbrace{e^{-y} dy}_{dv}}$$

$$= -y^\alpha e^{-y} \Big|_0^\infty + \int_0^\infty e^{-y} \cdot \alpha y^{\alpha-1} dy$$

$$= 0 + \alpha \Gamma(\alpha)$$

$$= \alpha / \beta.$$

Linearity of expectation

A very important property of expectation is that it is *linear*: For any random variables $X_1, \ldots, X_n$ (not necessarily independent),

$$\mathbb{E}[X_1 + \ldots + X_n] = \mathbb{E}[X_1] + \ldots + \mathbb{E}[X_n]$$

For any constant $c \in \mathbb{R}$,

$$\mathbb{E}[cX] = c\,\mathbb{E}[X]$$

Consequently, also

$$\mathbb{E}[c_1 X_1 + \ldots + c_n X_n] = c_1 \mathbb{E}[X_1] + \ldots + c_n \mathbb{E}[X_n]$$

Example: Let $X \sim \text{Binomial}(n, p)$. Recalling $X = X_1 + \ldots + X_n$ where $X_i \overset{IID}{\sim} \text{Bernoulli}(p)$, we may compute $\mathbb{E}[X]$ as

$$\mathbb{E}[X] = \mathbb{E}[X_1] + \ldots + \mathbb{E}[X_n] = np$$

Variance and standard deviation

The **variance** of X is defined by the two equivalent expressions

$$\mathrm{Var}[X] = \mathbb{E}\big[(X - \mathbb{E}X)^2\big] = \mathbb{E}[X^2] - (\mathbb{E}X)^2$$

If X is centered such that $\mathbb{E}X = 0$, then $\mathrm{Var}[X] = \mathbb{E}[X^2]$.

Variance is translation-invariant: For any constant $c \in \mathbb{R}$,

$$\mathrm{Var}[X + c] = \mathrm{Var}[X]$$

Also variance scales quadratically: For any constant $c \in \mathbb{R}$,

$$\mathrm{Var}[cX] = c^2\,\mathrm{Var}[X]$$

The **standard deviation** of X is $\sqrt{\mathrm{Var}[X]}$, which is interpretable on the scale of X rather than X^2.

Variance of independent sums

If $X_1, \ldots, X_n$ are *independent* (or more generally, pairwise uncorrelated), then

$$\text{Var}[X_1 + \ldots + X_n] = \text{Var}[X_1] + \ldots + \text{Var}[X_n]$$

Example: Let $X \sim \text{Binomial}(n, p)$. Recalling $X = X_1 + \ldots + X_n$ where $X_i \overset{IID}{\sim} \text{Bernoulli}(p)$,

$\text{Var}[X_1] = E[X_1^2] - (EX_1)^2 = p - p^2$

$$\begin{aligned}\text{Var}[X] &= \text{Var}[X_1] + \ldots + \text{Var}[X_n] \\ &= (p - p^2) + \ldots + (p - p^2) = np(1-p)\end{aligned}$$

$X_1 = \begin{cases} 1 & \text{w/ prob } p \\ 0 & \text{w/ prob } 1-p \end{cases}$

If $X_1, \ldots, X_n$ are *correlated*, then this is not true. For example,

$$\text{Var}[X_1 + X_2] = \text{Var}[X_1] + \text{Var}[X_2] + 2\,\text{Cov}[X_1, X_2]$$

and we will see later a more general expression.

Chi-squared expectation and variance

Example: Let $X \sim \chi^2(n)$. Recall that $X = X_1^2 + \ldots + X_n^2$ where $X_1, \ldots, X_n \overset{IID}{\sim} \mathcal{N}(0,1)$.

$\mathbb{E}[X] = \mathbb{E}[X_1^2 + \ldots + X_n^2] = \mathbb{E}[X_1^2] + \ldots + \mathbb{E}[X_n^2] = n$

$\text{Var}[X] = \text{Var}[X_1^2 + \ldots + X_n^2] = \text{Var}[X_1^2] + \ldots + \text{Var}[X_n^2] = n \cdot \text{Var}[X_1^2]$

$$\text{Var}[X_1^2] = \mathbb{E}[X_1^4] - (\mathbb{E}[X_1^2])^2 = 3 - 1 = 2$$

$$\Rightarrow \text{Var}[X] = 2n$$

$$\left(\mathbb{E}[X_1^4] = \int_{-\infty}^{\infty} x^4 \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \right.$$

$$= \int_{-\infty}^{\infty} \underbrace{x^3}_{u} \cdot \underbrace{\frac{1}{\sqrt{2\pi}} x e^{-\frac{x^2}{2}} dx}_{dv} = -x^3 \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \Big|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} 3x^2 \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$$

$$\left. = 0 + 3 \cdot \mathbb{E}[X_1^2] = 3 \right)$$

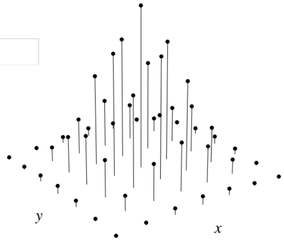
Joint distributions

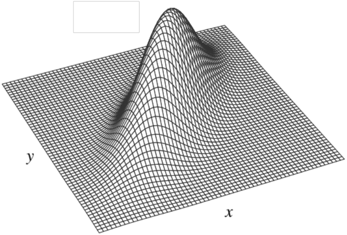
The *joint distribution* of k random variables $(X_1, \ldots, X_k)$ may be specified, in the discrete case, by the **joint PMF**

$$f(x_1, \ldots, x_k) = \mathbb{P}[X_1 = x_1, \ldots, X_k = x_k]$$

and in the continuous case, by the **joint PDF** $f(x_1, \ldots, x_k)$ which satisfies, for any $A \subseteq \mathbb{R}^k$,

$$\mathbb{P}[(X_1, \ldots, X_k) \in A] = \int \ldots \int_A f(x_1, \ldots, x_k) dx_1 \ldots dx_k.$$





Multinomial distribution

Example: The **multinomial** distribution generalizes the binomial to $k > 2$ outcomes: For n total samples, each independently belonging to outcomes $1, \ldots, k$ with probabilities $p_1, \ldots, p_k$, the total number of samples for each outcome is

$$(X_1, \ldots, X_k) \sim \text{Multinomial}\left(n, (p_1, \ldots, p_k)\right)$$

E.g., if we roll a standard six-sided die 100 times and $(X_1, \ldots, X_6)$ are the numbers of rolls 1 to 6, then

$$(X_1, \ldots, X_6) \sim \text{Multinomial}\left(100, (\tfrac{1}{6}, \tfrac{1}{6}, \tfrac{1}{6}, \tfrac{1}{6}, \tfrac{1}{6}, \tfrac{1}{6})\right).$$

The joint PMF is

$$f(x_1, \ldots, x_k) = \binom{n}{x_1, \ldots, x_k} p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k}$$

for all $x_1, \ldots, x_k \geq 0$ such that $x_1 + \ldots + x_k = n$

Marginal distributions

Given a joint distribution of (X, Y), the **marginal distribution** of X is its individual distribution ignoring Y.

If (X, Y) are discrete with joint PMF $f_{XY}(x, y)$, then the marginal PMF of X is

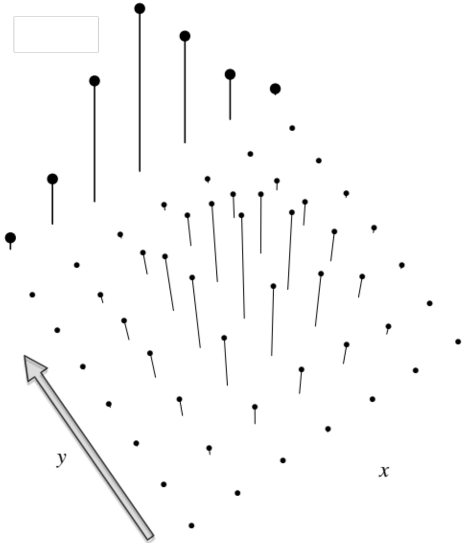
$$f_X(x) = \sum_{y \in \mathcal{Y}} f_{XY}(x, y)$$

where the sum is over all possible values of $\mathcal{Y}$.

If (X, Y) are continuous with joint PDF $f_{XY}(x, y)$, then the marginal PDF of X is

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y) dy$$

Marginal distributions



Conditional distributions

Given a joint distribution of (X, Y), the **conditional distribution** of Y given $X = x$ is its distribution after observing $X = x$.

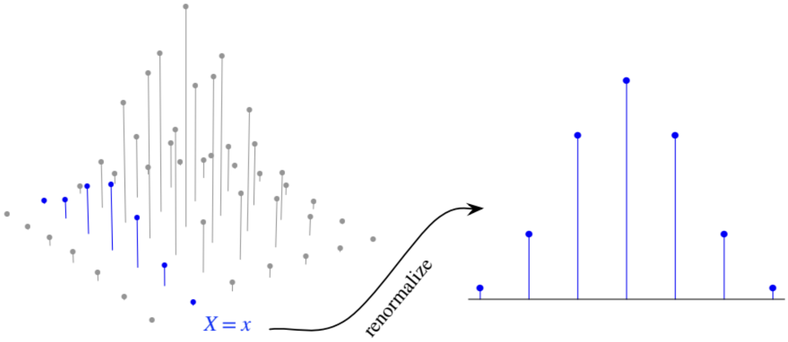
If (X, Y) are discrete with joint PMF $f_{XY}(x, y)$ and X has marginal PMF $f_X(x)$, the conditional PMF of Y given $X = x$ is

$$f_{Y|X}(y|x) = \mathbb{P}[Y = y \mid X = x] = \frac{\mathbb{P}[X = x, Y = y]}{\mathbb{P}[X = x]} = \frac{f_{XY}(x, y)}{f_X(x)}$$

If (X, Y) are continuous with joint PDF $f_{XY}(x, y)$ and X has marginal PDF $f_X(x)$, the conditional PDF of Y given $X = x$ is also

$$f_{Y|X}(y|x) = \frac{f_{XY}(x, y)}{f_X(x)}$$

Conditional distributions



Independence of random variables

Random variables $X_1, \ldots, X_n$ are **independent** when their PMFs or PDFs satisfy

$$f(x_1, \ldots, x_n) = f(x_1) \times \ldots \times f(x_n)$$

Thus their joint distribution is fully specified by the marginal distributions of the individual variables $X_1, \ldots, X_n$.

If $X_1, \ldots, X_n$ are independent, then for any $A_1, \ldots, A_n \subseteq \mathbb{R}$,

$$\mathbb{P}[X_1 \in A_1, \, \ldots, \, X_n \in A_n] = \mathbb{P}[X_1 \in A_1] \times \ldots \times \mathbb{P}[X_n \in A_n]$$

Furthermore, for any functions $g_1, \ldots, g_n : \mathbb{R} \to \mathbb{R}$,

$$\mathbb{E}[g_1(X_1) \ldots g_n(X_n)] = \mathbb{E}[g_1(X_1)] \times \ldots \times \mathbb{E}[g_n(X_n)].$$

Covariance

The **covariance** between two random variables X and Y is defined by the two equivalent expressions

$$\mathrm{Cov}[X, Y] = \mathbb{E}\left[(X - \mathbb{E}X)(Y - \mathbb{E}Y)\right] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

In particular, $\mathrm{Cov}[X, X] = \mathrm{Var}[X]$. If X and Y are centered so that $\mathbb{E}X = 0$ and $\mathbb{E}Y = 0$, then $\mathrm{Cov}[X, Y] = \mathbb{E}[XY]$.

If X, Y are independent, then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ so

$$\mathrm{Cov}[X, Y] = 0$$

However, the converse is not true: $\mathrm{Cov}[X, Y] = 0$ does not imply that X, Y are independent.

Bilinearity of covariance

Covariance is translation invariant: For any constants $a, b \in \mathbb{R}$,

$$\mathrm{Cov}[X + a, Y + b] = \mathrm{Cov}[X, Y]$$

Furthermore, covariance is *bilinear*: For any random variables $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ (not necessarily independent),

$$\mathrm{Cov}[X_1 + \ldots + X_n,\ Y_1 + \ldots + Y_m] = \sum_{i=1}^{n} \sum_{j=1}^{m} \mathrm{Cov}[X_i, Y_j]$$

For any constants $a, b \in \mathbb{R}$, $\mathrm{Cov}[aX, bY] = ab\,\mathrm{Cov}[X, Y]$.

Consequently, also

$$\mathrm{Cov}[a_1X_1 + \ldots + a_nX_n,\ b_1Y_1 + \ldots + b_mY_m] = \sum_{i=1}^{n} \sum_{j=1}^{m} a_i b_j\,\mathrm{Cov}[X_i, Y_j]$$

Bilinearity of covariance

Example: This allows us to derive a general expression for $\mathrm{Var}[X_1 + \ldots + X_n]$ when $X_1, \ldots, X_n$ may be dependent:

$$\mathrm{Var}[X_1 + \ldots + X_n] = \mathrm{Cov}[X_1 + \ldots + X_n, X_1 + \ldots + X_n]$$

$$= \sum_{i,j=1}^{n} \mathrm{Cov}[X_i, X_j] \quad \text{(by bilinearity)}$$

$$= \sum_{i=1}^{n} \mathrm{Cov}[X_i, X_i] + 2 \sum_{i<j} \mathrm{Cov}[X_i, X_j]$$

$$= \sum_{i=1}^{n} \mathrm{Var}[X_i] + \underbrace{2 \cdot \sum_{i<j} \mathrm{Cov}[X_i, X_j]}_{= 0 \text{ if } X_1, \ldots, X_n \text{ independent}}$$

Correlation

The **correlation** between (X, Y) is their covariance normalized by the product of standard deviations:

$$\mathrm{corr}(X, Y) = \frac{\mathrm{Cov}[X, Y]}{\sqrt{\mathrm{Var}[X]}\sqrt{\mathrm{Var}[Y]}}$$

Correlation is both translation and scale invariant: For any $a, b \in \mathbb{R}$ and $c, d > 0$,

$$\mathrm{corr}(aX + b, cY + d) = \mathrm{corr}(X, Y)$$

The **Cauchy-Schwarz inequality** says that for any (X, Y),

$$\mathrm{Cov}[X, Y]^2 \leq \mathrm{Var}[X]\,\mathrm{Var}[Y]$$

Consequently, we always have $\mathrm{corr}(X, Y) \in [-1, 1]$.