National Ph.D. Program in *Artificial Intelligence for Society*

Statistics for Machine Learning

Lesson 04 - Functions of random variables. Distances between distributions. Simulation.

Andrea Pugnana, Salvatore Ruggieri

Department of Computer Science
University of Pisa, Italy
andrea.pugnana@di.unipi.it salvatore.ruggieri@unipi.it

Functions of two or more random variables: expectation

- $V = \pi HR^2$ be the volume of a vase of height H and radius R
- $g(H, R) = \pi HR^2$ is a random variable (function of random variables)
- $P_V(V = 3) = P_{HR}(\pi HR^2 = 3)$
- How to calculate $E[V]$?

Two-dimensional change-of-variable formula. Let X and Y be random variables, and let $g : \mathbb{R}^2 \to \mathbb{R}$ be a function.
If X and Y are *discrete* random variables with values $a_1, a_2, \ldots$ and $b_1, b_2, \ldots$, respectively, then

$$\mathrm{E}[g(X,Y)] = \sum_i \sum_j g(a_i, b_j)\mathrm{P}(X = a_i, Y = b_j)\,.$$

If X and Y are *continuous* random variables with joint probability density function f, then

$$\mathrm{E}[g(X,Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y) f(x,y)\,\mathrm{d}x\,\mathrm{d}y.$$

If $H \perp\!\!\!\perp R$:

$$E[V] = E[\pi HR^2] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \pi h r^2 f_H(h) f_R(r) dh dr$$

Linearity of expectations

Theorem. For X and Y random variables, and $s, t \in \mathbb{R}$:

$$E[rX + sY + t] = rE[X] + sE[Y] + t$$

Proof. (discrete case)

$$E[rX + Ys + t] = \sum_a \sum_b (ra + sb + t)P(X = a, Y = b)$$

$$= \left(r \sum_a \sum_b aP(X = a, Y = b) \right) + \left(s \sum_a \sum_b bP(X = a, Y = b) \right) + \left(t \sum_a \sum_b P(X = a, Y = b) \right)$$

$$= \left(r \sum_a aP(X = a) \right) + \left(s \sum_b bP(Y = b) \right) + t = rE[X] + sE[Y] + t$$

Corollary. $E[a_0 + \sum_{i=1}^n a_i X_i] = a_o + \sum_{i=1}^n a_i E[X_i]$

Corollary. $X \leq Y$ implies $E[X] \leq E[Y]$
Proof. $Z = Y - X \geq 0$ implies $E[Z] = E[Y] - E[X] \geq 0$, i.e., $E[Y] \geq E[X]$.

Applications

- Expectation of some discrete distributions
  - $X \sim Ber(p) \quad E[X] = p$
  - $X \sim Bin(n, p) \quad E[X] = n \cdot p$
    - Because $X = \sum_{i=1}^{n} X_i$ for $X_1, \ldots, X_n \sim Ber(p)$
  - $X \sim Geo(p) \quad E[X] = \frac{1}{p}$
  - $X \sim NBin(n, p) \quad E[X] = \frac{n \cdot (1-p)}{p}$
    - Because $X = \sum_{i=1}^{n} X_i - n$ for $X_1, \ldots, X_n \sim Geo(p)$
- Expectation of some continuous distributions
  - $X \sim Exp(\lambda) \quad E[X] = 1/\lambda$
  - $X \sim Erl(n, \lambda) \quad E[X] = \frac{n}{\lambda}$
    - Because $X = \sum_{i=1}^{n} X_i$ for $X_1, \ldots, X_n \sim Exp(\lambda)$

Expectation of product and quotients

Theorem. For $X \perp\!\!\!\perp Y$, we have: $E[XY] = E[X]E[Y]$ **Prove it!**

PROPAGATION OF INDEPENDENCE. Let $X_1, X_2, \ldots, X_n$ be independent random variables. For each i, let $h_i : \mathbb{R} \to \mathbb{R}$ be a function and define the random variable

$$Y_i = h_i(X_i).$$

Then $Y_1, Y_2, \ldots, Y_n$ are also independent.

Corollary. For $X \perp\!\!\!\perp Y$ and $Y > 0$, we have: $E[X/Y] \geq E[X]/E[Y]$
Proof. $X \perp\!\!\!\perp Y$ implies $X \perp\!\!\!\perp 1/Y$. By theorem above:

$$E[X/Y] = E[X \cdot 1/Y] = E[X]E[1/Y] \geq E[X]/E[Y]$$

because by Jensen's inequality $E[1/Y] \geq 1/E[Y]$ since $1/y$ is convex for y 0. □

Exercise at home. Show that $E[X/Y] = E[X]/E[Y]$ is a false claim.

Law of iterated/total expectation

Conditional expectation

$$E[X|Y=b] = \sum_i a_i p(a_i|b) \qquad E[X|Y=y] = \int_{-\infty}^{\infty} x f(x|y)dx$$

Theorem. (Law of iterated/total expectation)

$$E_Y[E[X|Y]] = E[X]$$

Proof. (for X, Y discrete random variables)

$$E_Y[E[X|Y]] = \sum_j \sum_i a_i p_{X|Y}(a_i|b_j) p_Y(b_j) = \sum_j \sum_i a_i p_{XY}(a_i, b_j) = \sum_i a_i p_X(a_i) = E[X]$$

Example (cfr the example from Lesson 1 on the Law of total probability)

- Factory 1's light bulbs working hours $\sim Exp(1/1000)$
- Factory 2's light bulbs working hours $\sim Exp(1/2000)$
- Factory 1 supplies 60% of the total bulbs on the market and Factory 2 supplies 40% of it.
- *What is the average work hour of a light bulb on the market?*

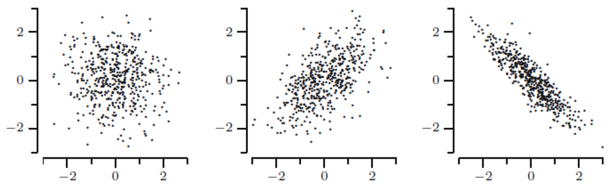
Variance of the sum and covariance

$$
\begin{aligned}
& Var(X+Y) = E[(X+Y-E[X+Y])^2] = E[((X-E[X])+(Y-E[Y]))^2] \\
= \;& E[(X-E[X])^2] + E[(Y-E[Y])^2] + 2E[(X-E[X])(Y-E[Y])] \\
= \;& Var(X) + Var(Y) + 2Cov(X,Y)
\end{aligned}
$$

Covariance

The *covariance* $Cov(X,Y)$ of two random variables X and Y is the number:

$$Cov(X,Y) = E[(X-E[X])(Y-E[Y])]$$



Uncorrelated **Positively correlated** **Negatively correlated**

Covariance

Theorem. $Cov(X, Y) = E[XY] - E[X]E[Y]$ **Prove it!**

- If X and Y are independent ($X \perp\!\!\!\perp Y$):

$$Cov(X, Y) = 0 \qquad Var(X + Y) = Var(X) + Var(Y)$$

- But there are X and Y uncorrelated (ie., $Cov(X, Y) = 0$) that are dependent!
- Variances of some discrete distributions
  - $X \sim Ber(p)$ $\quad Var(X) = p(1-p)$
  - $X \sim Bin(n, p)$ $\quad Var(X) = np(1-p)$
    - Because $X = \sum_{i=1}^{n} X_i$ for $X_1, \ldots, X_n \sim Ber(p)$ and independent
  - $X \sim Geo(p)$ $\quad Var(X) = \frac{1-p}{p^2}$
  - $X \sim NBin(n, p)$ $\quad Var(X) = n\frac{1-p}{p^2}$
    - Because $X = \sum_{i=1}^{n} X_i - n$ for $X_1, \ldots, X_n \sim Geo(p)$ and independent
- Variances of some continuous distributions
  - $X \sim Exp(\lambda)$ $\quad Var(X) = 1/\lambda^2$
  - $X \sim Erl(n, \lambda)$ $\quad Var(X) = \frac{n}{\lambda^2}$
    - Because $X = \sum_{i=1}^{n} X_i$ for $X_1, \ldots, X_n \sim Exp(\lambda)$ and independent

Covariance and covariance matrix

COVARIANCE UNDER CHANGE OF UNITS. Let X and Y be two random variables. Then

$$\mathrm{Cov}(rX + s, tY + u) = rt\,\mathrm{Cov}(X, Y)$$

for all numbers r, s, t, and u.

- Hence, $Var(rX + sY + t) = r^2 Var(X) + s^2 Var(Y) + 2rsCov(X, Y)$
- **Bivariate** Normal/Gaussian distribution:

$$(X, Y) \sim \mathcal{N}((\mu_X, \mu_X), \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{pmatrix})$$

  - where marginals are $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$, $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$, and $Cov(X, Y) = \sigma_{XY}$
  - Covariance matrix $\mathbf{\Sigma}_{ij} = Cov(X_i, X_j)$ for a vector $\mathbf{X} = (X_1, \ldots, X_n)$ of r.v.'s

- Covariance depends on the unit of measure!

Correlation coefficient

DEFINITION. Let X and Y be two random variables. The *correlation coefficient* $\rho(X,Y)$ is defined to be 0 if $\mathrm{Var}(X) = 0$ or $\mathrm{Var}(Y) = 0$, and otherwise

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}}.$$

- Correlation coefficient is *dimensionless* (not affected by change of units)
  - E.g., if X and Y are in Km, then $Cov(X,Y)$, $Var(X)$ and $Var(Y)$ are in Km^2
- Moreover: $-1 \leq \rho(X,Y) \leq 1$
  - The bounds are derived from the **Cauchy–Schwarz's inequality**:

$$E[|XY|] \leq \sqrt{E[X^2]}\sqrt{E[Y^2]}$$

Proof. For any $u, w \in \mathbb{R}$, we have $2|uw| \leq u^2 + w^2$. Therefore, $2|UW| \leq U^2 + W^2$ for r.v.'s U and V. By defining $U = X/\sqrt{E[X^2]}$ and $W = Y/\sqrt{E[Y^2]}$ (*), we have $2 \cdot |XY|/\sqrt{E[X^2]}\sqrt{E[Y^2]} \leq X^2/E[X^2] + Y^2/E[Y^2]$. Taking the expectations, we conclude: $2 \cdot E[|XY|]/\sqrt{E[X^2]}\sqrt{E[Y^2]} \leq 2$.

(*) The case $E[X^2] = 0$ or $E[Y^2] = 0$ is left as an exercise. □

Bivariate Normal/Gaussian distribution

$$(X, Y) \sim \mathcal{N}((\mu_X, \mu_Y), \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{pmatrix})$$

where marginals are $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$, $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$, and $Cov(X, Y) = \sigma_{XY}$

- Since $\sigma_{XY} = \rho(X, Y) \cdot \sigma_X \cdot \sigma_Y$:

$$(X, Y) \sim \mathcal{N}((\mu_X, \mu_Y), \begin{pmatrix} \sigma_X^2 & \rho(X, Y) \cdot \sigma_X \cdot \sigma_Y \\ \rho(X, Y) \cdot \sigma_X \cdot \sigma_Y & \sigma_Y^2 \end{pmatrix})$$

- Density of $\mathcal{N}((0, 0), (1, \sigma_{XY}, \sigma_{XY}, 1))$:

$$f(x, y) = \frac{1}{2\pi\sqrt{1 - \sigma_{XY}^2}} e^{-\frac{1}{2(1-\sigma_{XY}^2)}(x^2 + y^2 - 2xy\sigma_{XY})}$$

- Useful facts for (X, Y) bivariate Normal:
  - for (X, Y) bivariate Normal: $\rho(X, Y) = 0$ iff $X \perp\!\!\!\perp Y$, i.e., uncorrelation equals independence
  - (X, Y) bivariate Normal iff $aX + bY$ is Normal for any $a, b \in \mathbb{R}$

Sum of independent Normal random variables

- See Lesson 04 and Lesson 08 for convolution formulas

> ADDING TWO INDEPENDENT CONTINUOUS RANDOM VARIABLES. Let X and Y be two independent continuous random variables, with probability density functions f_X and f_Y. Then the probability density function f_Z of $Z = X + Y$ is given by
>
> $$f_Z(z) = \int_{-\infty}^{\infty} f_X(z-y) f_Y(y)\,\mathrm{d}y$$
>
> for $-\infty < z < \infty$.

Theorem. If $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ and $X \perp\!\!\!\perp Y$, then:

$$Z = X + Y \sim \mathcal{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$$

Proof. See [T, Sect. 11.2] □

- In general: $Z = rX + sY + t \sim \mathcal{N}(r\mu_X + s\mu_Y + t, r^2\sigma_X^2 + s^2\sigma_Y^2)$
- The converse of the theorem also holds: [**Lévy-Cramér theorem**]
  - If $X \perp\!\!\!\perp Y$ and $Z = X + Y$ is normally distributed, then X and Y follow a normal distribution.

Extremes of independent random variables

THE DISTRIBUTION OF THE MAXIMUM. Let $X_1, X_2, \ldots, X_n$ be n independent random variables with the same distribution function F, and let $Z = \max\{X_1, X_2, \ldots, X_n\}$. Then

$$F_Z(a) = (F(a))^n.$$

- $P(Z \leq a) = P(X_1 \leq a, \ldots, X_n \leq a) = \prod_{i=1}^{n} P(X_i \leq a) = ((F(a))^n$
- Example: maximum water level over 365 days assuming water level on a day is $U(0, 1)$
- Example: maximum of two rolls of **a die with 4 sides**

THE DISTRIBUTION OF THE MINIMUM. Let $X_1, X_2, \ldots, X_n$ be n independent random variables with the same distribution function F, and let $V = \min\{X_1, X_2, \ldots, X_n\}$. Then

$$F_V(a) = 1 - (1 - F(a))^n.$$

- $P(V \leq a) = 1 - P(X_1 > a, \ldots, X_n > a) = 1 - \prod_{i=1}^{n}(1 - P(X_i \leq a) = 1 - ((1 - F(a))^n$

Product and quotient of independent random variables

Product of independent continuous random variables. Let X and Y be two independent continuous random variables with probability densities f_X and f_Y. Then the probability density function f_Z of $Z = XY$ is given by

$$f_Z(z) = \int_{-\infty}^{\infty} f_Y\left(\frac{z}{x}\right) f_X(x) \frac{1}{|x|}\,\mathrm{d}x$$

for $-\infty < z < \infty$.

Quotient of independent continuous random variables. Let X and Y be two independent continuous random variables with probability densities f_X and f_Y. Then the probability density function f_Z of $Z = X/Y$ is given by

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(zx) f_Y(x) |x|\,\mathrm{d}x$$

for $-\infty < z < \infty$.

- $X, Y \sim \mathcal{N}(0,1)$ independent, $Z = X/Y \sim Cau(0,1)$ where:

$$f_Z(x) = \frac{1}{\pi(1+x^2)}$$

Distances and Metrics

A numerical measurement of how far apart two objects are.

Distances and Metrics

A distance over a set $\mathcal{A}$ is a function $d : \mathcal{A} \times \mathcal{A} \to \mathbb{R}$ such that:

- $d(x, y) \geq 0$ *non-negativity*
- $d(x, y) = 0$ iff $x = y$ *identity of indiscernibles*
- $d(x, y) = d(y, x)$ *symmetry*

Moreover, d is called a metric if in addition:

- $d(x, z) \leq d(x, y) + d(y, z)$ *triangle inequality*

Examples over $\mathcal{A} = \mathbb{R}^n$:

- Manhattan or L_1 distance $d_1(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_1 = \sum_{i=1}^{n} |\mathbf{x}_i - \mathbf{y}_i|$
- Euclidian or L_2 distance $d_2(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_2 = \sqrt{\sum_{i=1}^{n} (\mathbf{x}_i - \mathbf{y}_i)^2}$
- Chebyshev or L_∞ distance $d_\infty(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_\infty = \max_{i=1}^{n} |\mathbf{x}_i - \mathbf{y}_i|$

We aim at defining distances and metrics over probability distributions, i.e., when

$$\mathcal{A} = \{F \mid F : \mathbb{R} \to [0, 1] \text{ is a CDF}\}$$

Distances over probability distributions

A numerical measurement of how far apart two probability distributions are.

- ML/DM models are supposed to be applied on the same distribution as the training set:
  - How far is the test data distribution from the one of the training data? *[Transfer learning*
  - Is the data changing over time, thus my model is inadequate? *[Dataset shift]*
- ML/DM algorithms are supposed to choose the best hypothesis:
  - What is the split in a DT which best distinguish the distribution of classes?
  - Is my model separating positive and negatives as much as possible?
  - Is my clustering separating groups with different distributions?
- Data preprocessing looks at feature distribution:
  - Are these two features conveying the same information?
  - Can this feature be predictive to the class feature?
- ... and many other applications in Data Science

Total variation distance and KS distance

Let X, Y be random variables:

- Total Variation (TV) distance (discrete and continuous case):

$$d_{TV}(X,Y) = \frac{1}{2}\sum_i |p_X(a_i) - p_Y(a_i)| \qquad d_{TV}(X,Y) = \frac{1}{2}\int |f_X(x) - f_Y(x)|dx$$

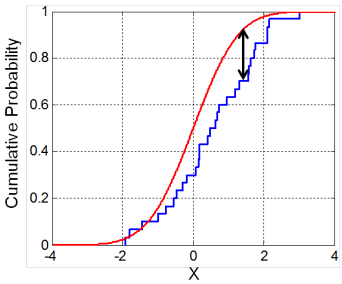
  - d_{TV} is a metric with $d_{TV}(X,Y) \in [0,1]$

- Kolmogorov-Smirnov (KS) distance:

$$d_{KS}(X,Y) = \sup_x |F_X(x) - F_Y(x)|$$

  - d_{KS} is a metric with $d_{KS}(X,Y) \in [0,1]$

- d_{TV} and d_{KS} have no closed forms in general
- d_{KS} can be estimated from samples of the distributions



Entropy $H(X)$ of a random variable X

- The **Shannon's information entropy** is the average level of "information" (or "surprise", "uncertainty", "unpredictability") inherent to the variable's possible outcomes
  - Information is inversely proportional to probability $\frac{1}{p(a_i)}$
    - Highly likely/unlikely events carry less/more new information
  - Information content $ic()$ of two independent events should sum up $\log \frac{1}{p(a_i)}$
    - $ic(p(A \cap B)) = ic(p(A)) + ic(p(B)) = ic(p(A)p(B))$
    - $ic(p(\Omega)) = ic(1) = 0$
    - $ic(p(A)) \geq 0$
- $H(X) = E[-\log p(X)]$ (*discrete*) $\qquad$ $H(X) = E[-\log f(X)]$ (*continuous*)

$$H(X) = -\sum_i p(a_i) \log p(a_i) \qquad H(X) = -\int_{-\infty}^{\infty} f(x) \log f(x) dx$$

  - For X discrete, $H(X) \geq 0$ since $-\log p(X) = \log {}^1\!/_{p(X)} \geq 0$
    - zero reached when $p(a_1) = 1$ and $p(a_i) = 0$ for $i \neq 1$
  - For $X \sim Ber(p)$, $H(X) = -p \log p - (1-p) \log (1-p)$ *[binary entropy function]*
    - for $X \sim Ber(0.5) : H(X) = -2 \cdot {}^1\!/_2 \log {}^1\!/_2 = 1$ *[unit of entropy is called a* **bit***]*

Cross entropy

- X, Y discrete random variables with p.m.f. p_X and p_Y:
- Cross entropy of X w.r.t. Y: $H(X;Y) = E_X[-\log p(Y)]$

$$H(X;Y) = -\sum_i p_X(a_i) \log p_Y(a_i)$$

$$\text{with } p_X(a_i)\log p_Y(a_i) = \begin{cases} 0 & \text{if } p_X(a_i) = 0 \\ -\infty & \text{if } p_X(a_i) > 0 \wedge p_Y(a_i) = 0 \end{cases}$$

- $H(X;Y)$ is the "information" or "uncertainty" or "loss" when using Y to encode X
- The closer p_X and p_Y, the lower is $H(X;Y)$
- The lower bound is for $Y = X$, for which $H(X;Y) = H(X)$

Kullback-Leibler divergence

KL divergence

For X, Y discrete random variables with p.m.f. p_X and p_Y:

$$D_{KL}(X \parallel Y) = \sum_i p_X(a_i) \log \frac{p_X(a_i)}{p_Y(a_i)} = H(X; Y) - H(X)$$

- Measure how distribution of Y (model) can reconstruct the distribution of X (data)
  - Also called: *relative entropy* or *information gain* of X w.r.t. Y
- Properties
  - $D_{KL}(X \parallel Y) \geq 0$ **[Gibbs' inequality]**
  - $D_{KL}(X \parallel Y) = 0$ iff $F_X = F_Y$
  - $D_{KL}(X \parallel Y) \neq D_{KL}(Y \parallel X)$ *[not a distance!]*
- For X, Y continuous: $D_{KL}(X \parallel Y) = \int_{-\infty}^{\infty} f_X(x) \log \frac{f_X(x)}{f_Y(x)} dx$

Joint entropy

- X, Y discrete random variables with p.m.f. p_X and p_Y:
- Joint p.m.f. p_{XY}. Joint entropy of (X, Y):

$$H((X,Y)) = -\sum_{i,j} p_{XY}(a_i, a_j) \log p_{XY}(a_i, a_j)$$

- If $X \perp\!\!\!\perp Y$, then:

$$H((X,Y)) = -\sum_{i,j} p_X(a_i) p_Y(a_j)(\log p_X(a_i) + \log p_Y(a_j)) =$$

$$= -(\sum_i p_X(a_i))(\sum_j p_Y(a_j) \log p_Y(a_j)) - (\sum_j p_Y(a_j))(\sum_i p_X(a_i) \log p_X(a_i)) = H(X) + H(Y)$$

Mutual information

Mutual information

For X, Y discrete random variables with p.m.f. p_X and p_Y and joint p.m.f. p_{XY}:

$$I(X,Y) = D_{KL}(p_{XY} \parallel p_X p_Y) = \sum_{i,j} p_{XY}(a_i, a_j) \log \frac{p_{XY}(a_i, a_j)}{p_X(a_i) p_Y(a_j)} = H(X) + H(Y) - H((X,Y))$$

- MI measures how dependent two distributions are
  - Measure how product of marginals can reconstruct the joint distribution
- Properties
  - $I(X,Y) = I(Y,X)$, and $I(X,Y) \geq 0$
  - $I(X,Y) = 0$ iff $X \perp\!\!\!\perp Y$
  - $NMI = \frac{I(X,Y)}{\min\{H(X), H(Y)\}} \in [0,1]$ *[Normalized mutual information]*
- For X, Y continuous: $I(X,Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x,y) \log \frac{f_{XY}(x,y)}{f_X(x) f_Y(y)} dxdy$

The data processing inequality

- Let X be unknown, and assume to observe a noisy version Y of it
- Let $Z = f(Y)$ be a data processing to improve the "quality" of Y
- Z does not increase the information about X, i.e.: *[Data processing inequality]*

$$I(X, Y) \geq I(X, Z)$$

- If $I(X, Y) = I(X, Z)$ and Z is a summary of Y, we call it a *sufficient statistics*
  - Let $X \sim Ber(\theta)$ and $Y = (Y_1, \ldots, Y_n) \sim Ber(\theta)^n$ modelling i.i.d. observations
  - $Z = \sum_{i=1}^{n} Y_i \sim Binom(n, \theta)$ is a sufficient statistics
  - **Proof (sketch):** use $D_{KL}(p_{XY} \parallel p_X p_Y)$ and:

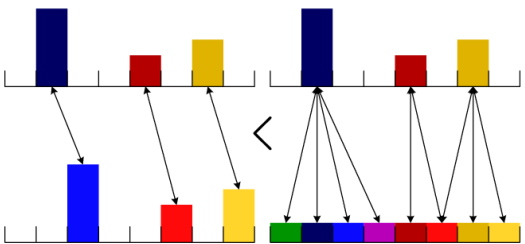
$$p(Y_1 = y1, \ldots, Y_n = y_n) = \prod_i \theta^{y_i}(1-\theta)^{(1-y_i)} = \theta^{\sum_i y_i}(1-\theta)^{n-\sum_i y_i} = p(Z = \sum_i y_i)$$

Earth mover's distance / Wasserstein metric

- The minimum cost to transform one distribution to another
- Cost = amount of mass to move $\times$ distance to move it
- X, Y discrete random variables:

$$EMD(X, Y) = \frac{\sum_{i,j} F_{i,j} \cdot |a_i - a_j|}{\sum_{i,j} F_{i,j}}$$

where F is the flow which minimizes the numerator (total cost) subject to **some constraints**.



Earth mover's distance / Wasserstein metric

- The minimum cost to transform one distribution to another
- Solution of the transportation problem for X, Y multivariate (version from **Ramdas et al. 2015**):

$$EMD(X, Y) = \int_0^1 \|F_X^{-1}(p) - F_Y^{-1}(p)\| \, dp$$

For X, Y univariate, this simplifies to:

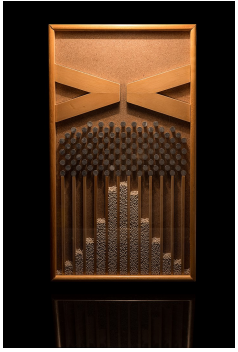
$$EMD(X, Y) = \sum_i |F_X(a_i) - F_Y(a_i)| \qquad EMD(X, Y) = \int_{-\infty}^{\infty} |F_X(x) - F_Y(x)| dx$$

- For empirical distributions from **ordered** samples $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$:

$$EMD(X, Y) = \frac{1}{n} \sum_i |x_i - y_i|$$

Simulation

- Not all problems can be solved with calculus!
- Complex interactions among random variables can be simulated
- Generated random values are called *realizations*
- Basic issue: *how to generate realizations?*
  - The **Galton Board**



Simulation

- Not all problems can be solved with calculus!
- Complex interactions among random variables can be simulated
- Generated random values are called *realizations*
- Basic issue: *how to generate realizations?*
- **Assumption**: we are only given $U(0, 1)$
- **Problem**: derive all the other random generators

Simulation: discrete distributions

Bernoulli random variables

Suppose U has a $U(0,1)$ distribution. To construct a $Ber(p)$ random variable for some $0 < p < 1$, we define

$$X = \begin{cases} 1 & \text{if } U < p, \\ 0 & \text{if } U \geq p \end{cases}$$

so that

$$\begin{aligned} \mathrm{P}(X = 1) &= \mathrm{P}(U < p) = p, \\ \mathrm{P}(X = 0) &= \mathrm{P}(U \geq p) = 1 - p. \end{aligned}$$

This random variable X has a Bernoulli distribution with parameter p.

- For $X_1, \ldots, X_n \sim Ber(p)$ i.i.d., we have: $\sum_{i=1}^{n} X_i \sim Binom(n, p)$

$X \sim Cat(\mathbf{p})$

DEFINITION. A discrete random variable X has a *Bernoulli distribution* with parameter p, where $0 \leq p \leq 1$, if its probability mass function is given by

$$p_X(1) = \mathrm{P}(X = 1) = p \quad \text{and} \quad p_X(0) = \mathrm{P}(X = 0) = 1 - p.$$

We denote this distribution by $Ber(p)$.

- Alternative definition: $p_X(a) = p^a \cdot (1-p)^{1-a}$ for $a \in \{0, 1\}$
- Categorical distribution generalizes to $n_C \geq 2$ possible values

$X \sim Cat(\mathbf{p})$

Categorical distribution

A discrete random variable X has a Categorical distribution with parameters $p_0, \ldots, p_{n_C-1}$ where $\sum_i p_i = 1$ and $p_i \in [0, 1]$ if its p.m.f. is given by:

$$p_X(i) = P(X = i) = p_i \quad \text{for } i = 0, \ldots, n_C - 1$$

- Alternative definition: $p_X(a) = \prod_i p_i^{\mathbb{1}_{a=i}}$ for $a \in \{0, \ldots, n_C - 1\}$

Notation. Indicator function: $\mathbb{1}_\varphi(x) = \begin{cases} 1 & \text{if } \varphi(x) \\ 0 & \text{otherwise} \end{cases}$

$X \sim Mult(n, \mathbf{p})$

- $X \sim Bin(n, p)$ models the number of successes in n Bernoulli trials
- **Intuition**: for $X_1, X_2, \ldots, X_n$ i.i.d. $X_i \sim Ber(p)$: $X = \sum_{i=1}^{n} X_i \sim Bin(n, p)$
- $X \sim Mult(n, \mathbf{p})$ models the number of categories in n Categorical trials
- **Intuition**: for $X_1, X_2, \ldots, X_n$ such that $X_i \sim Cat(\mathbf{p})$ and independent (**i.i.d.**), define:

$$Y_1 = \sum_{i=1}^{n} \mathbb{1}_{X_i=0} \sim Bin(n, p_0) \quad \ldots \quad Y_{n_C} = \sum_{i=1}^{n} \mathbb{1}_{X_i=n_C-1} \sim Bin(n, p_{n_C-1})$$

$$X = (Y_1, \ldots, Y_{n_C}) \sim Mult(n, \mathbf{p})$$

Multinomial distribution

A discrete random variable $X = (Y_1, \ldots, Y_{n_C})$ has a Multinomial distribution with parameters $p_0, \ldots, p_{n_C-1}$ where $\sum_i p_i = 1$ and $p_i \in [0, 1]$ if its p.m.f. is given by:

$$p_X(i_0, \ldots, i_{n_C-1}) = P(X = (i_0, \ldots, i_{n_C-1})) = \frac{n!}{i_0! i_1! \ldots i_{n_C-1}!} p_0^{i_0} p_1^{i_1} \cdots p_{(n_C-1)}^{i_{(n_C-1)}}$$

$X \sim Mult(n, \mathbf{p})$

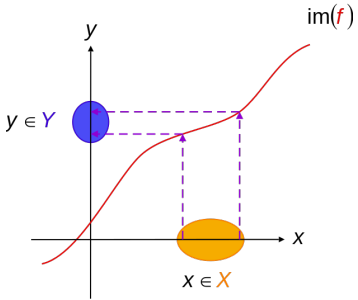
- Example: student selection from a population with $n_C = 3$:
  - $p_0 = 60\%$ undergraduates
  - $p_1 = 30\%$ graduate
  - $p_2 = 10\%$ PhD students
- Assume $n = 20$ students are randomly selected
- $X \sim (Y_1, Y_2, Y_3)$ where:
  - Y_1 number of undergraduate students selected
  - Y_2 number of graduate students selected
  - Y_3 number of PhD students selected
- $P(X = (10, 6, 4)) = \frac{20!}{10!6!4!}(0.6)^{10}(0.3)^6(0.1)^4 = 9.6\%$

Simulation: continuous distributions

- $F(x) = P_X(X \leq x)$
- $F : \mathbb{R} \to [0,1]$ invertible as $F^{-1} : [0,1] \to \mathbb{R}$
  - E.g., F *strictly* increasing
  - N.B., the textbook notation for F^{-1} is F^{inv}
- For $Y \sim U(0,1)$ and $0 \leq b \leq 1$
  $P_Y(Y \leq b) = b$

  then, for $b = F(x)$
  $P_Y(Y \leq F(x)) = F(x)$

  and then by inverting $X = F^{-1}(Y)$
  $P_X(X \leq x) = P_Y(F^{-1}(Y) \leq x) = F(x)$
- In summary:
  $X = F^{-1}(Y) \sim F$ for $Y \sim U(0,1)$
- Example: $F(x) = 1 - e^{-\lambda x}$ for $Exp(\lambda)$
  - $F^{-1}(y) = -1/\lambda \log(1-y)$



Reference book chapter for this lesson

Kevin P. Murphy (2022)

Probabilistic Machine Learning: An Introduction

Chapter 6: Information Theory

online book

Optional reference

William H. Press, Saul A. Teukolsky, William T. Vetterling, Brian P. Flannery (2007)
Numerical Recipes - The Art of Scientific Computing
Chapter 7: Random Numbers
online book