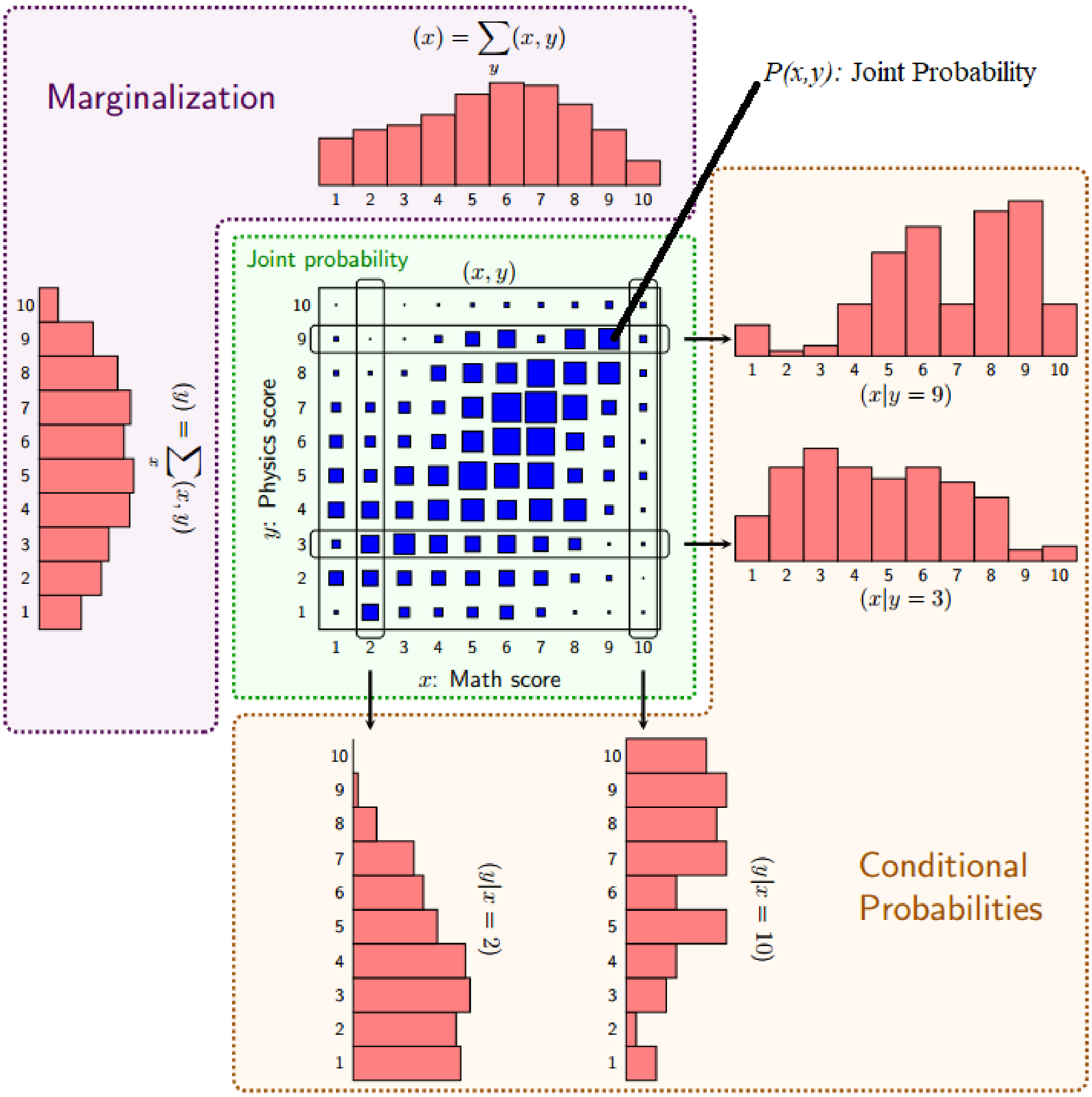
Probability

Random Variables

- Random Variable is the represented as the outcome of the experiment, for example, flip the coin (*discrete*) or measure the temperature of each day in one month (*continuous*)
- The set of all outcomes can be used to build probability distribution by function $p(x)$
- Contrary with discrete Random Variables which probability of particular outcome is the number $0 < p(X = x) < 1$, probability of continuous RV of particular outcome is 0
- Instead of calculating probability of particular outcome, continuous RV calculates probability of certain range, and it is called *probability density function (pdf)*: $\int p(x) = 1$

Joint Probability

- Joint Probability ($p(x, y)$): the probability such that $X = x$ *and* $Y = y$ simultaneously occur
- x *and* y may be both continuous or discrete or one is continuous, one is discrete
    - If both continuous: $\int p(x, y)\, dxdy = 1$
    - If both discrete: $\sum_{x,y} p(x, y) = 1$
    - If one is continuous, one is discrete: $\sum_x \int p(x, y)\, dy = \int \sum_x p(x, y)\, dy = 1$



- When we have joint probability of ≥ 3 RV, e.g. *p(x,y,z)* is equivalent with $p(\mathbf{x}), \mathbf{x} = \begin{bmatrix} x & y & z \end{bmatrix}^T$

Marginal Probability

- Based on the above plot, marginal probability is
  - If *x* and *y* are discrete $\begin{cases} p(x) = \sum_y p(x,y) \\ p(y) = \sum_x p(x,y) \end{cases}$
  - If *x* and *y* is continuous $\begin{cases} p(x) = \int p(x,y)dy \\ p(y) = \int p(x,y)dx \end{cases}$
  - If we have > 2 RV: $\begin{cases} p(x) = \sum_{y,z,w} p(x,y,z,w) \\ p(x,y) = \sum_{z,w} p(x,y,z,w) \end{cases}$
- The way we $p(X=x) = \sum_y p(X=x, Y=y)$ or $p(X=x) = \int p(X=x, Y=y)dy$ is called *marginalization* and its distribution is called *marginal probability*
- From now on, regardless of continuous or discrete, we always use $\sum$. If discrete, $\sum$ means $\sum$, otherwise $\int$

Conditional Probability

- For example, we want to find out $p(x=1 \mid y=9) = \frac{p(x=1, y=9)}{\sum_x p(x, y=9)} = \frac{p(x=1, y=9)}{p(y=9)} = \frac{Joint\ Proba}{Marginal\ Proba}$
- We have > 2 RV:

$$
\begin{aligned}
p(x,y,z,w) = p(x,y,z \mid w)\,p(w) \\
&= p(x,y \mid z,w)\,p(z,w) = p(x,y \mid z,w)\,p(z \mid w)\,p(w) \\
&= p(x \mid y,z,w)\,p(y \mid z,w)\,p(z \mid w)\,p(w)
\end{aligned}
$$

Bayes Theorem

Write later

Independent RV

- 2 RV are called *independent* if they don't have any connection with each other. E.g. The height and Math score of student are 2 independent RV
- When 2 RV *x* and *y* are independent, $\begin{cases} p(x,y) = p(x) \\ p(x,y) = p(y) \end{cases} \rightarrow p(x,y) = p(x \mid y)\,p(y) = p(x)\,p(y)$

Expectation and Covariance Matrix

- Expectation of Random Variable X: $\begin{cases} E[X] = \sum_x xp(x) & \textit{for discrete} \\ E[X] = \int xp(x)dx & \textit{for continuous} \end{cases}$
- Expectation Properties
  - If α = const, E[α] = α

- Linear Property: $\begin{cases} E[\alpha X] = \alpha E[X] \\ E[X+Y] = E[X] + E[Y] \end{cases}$
- If 2 Random variables are independent: $E[XY] = E[X]E[Y]$

- We have N values $x_1, x_2, \ldots x_N$, $\begin{cases} Expectation: \bar{x} = \mu = \frac{1}{N}\sum_{n=1}^{N} x_n \\ Variance: \sigma^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n - \bar{x})^2 \end{cases}$, so if Variance is large, the outcome is more spread compared to the mean outcome
- We have N points represented as N vector $\mathrm{x}_1, \mathrm{x}_2, \ldots \mathrm{x}_N$, $\begin{cases} Expectation: \bar{\mathrm{x}} = \frac{1}{N}\sum_{n=1}^{N} \mathrm{x}_n \\ Covariance: S = \frac{1}{N}\sum_{n=1}^{N}(\mathrm{x}_n - \bar{\mathrm{x}})(\mathrm{x}_n - \bar{\mathrm{x}})^T = \frac{1}{N}\hat{X}\hat{X}^T \end{cases}$

  E.g. We have matrix $\begin{bmatrix} 0.1 & 0.3 & 0.4 & 0.8 & 0.9 \\ 3.2 & 2.4 & 2.4 & 0.1 & 5.5 \\ 10. & 8.2 & 4.3 & 2.6 & 0.9 \end{bmatrix}$, its covariance matrix is $\begin{bmatrix} 0.115 & 0.0575 & -1.2325 \\ 0.0575 & 3.757 & -0.8775 \\ -1.2325 & -0.8775 & 14.525 \end{bmatrix}$
- Covariance Matrix Properties
  - Covariance Matrix is the symmetric matrix and PSD
  - All values in diagonal ≥ 0, they are also variance of each row, e.g. *0.115* is variance of $\begin{bmatrix} 0.1 & 0.3 & 0.4 & 0.8 & 0.9 \end{bmatrix}$
  - All values outside diagonal represent the correlation between row *i* and row *j*, e.g. 0.0575 is the correlation between row $\begin{bmatrix} 0.1 & 0.3 & 0.4 & 0.8 & 0.9 \end{bmatrix}$ and $\begin{bmatrix} 3.2 & 2.4 & 2.4 & 0.1 & 5.5 \end{bmatrix}$
  - Therefore, if covariance matrix is a diagonal matrix, the correlation between points (rows) are *0*

Common Distribution

Bernoulli Distribution

- Bernoulli distribution is discrete distribution with the output of 2 value *0* or *1*
- We can generalize these 2 values: *fraud transaction* or *common transaction*, *head* or *tail*
- Bernoulli has parameter $\lambda \in [0,1]$ which is the probability of *1*: $\begin{cases} p(x=1) = \lambda \\ p(x=0) = 1-\lambda \end{cases}$
- Symbol: $p(x) = Bern_x[\lambda] = \lambda^x(1-\lambda)^{1-x}$

Categorical Distribution

- Categorical Distribution is also discrete distribution but its output > *2*
- If we have K outputs, categorical distribution has $\lambda = [\lambda_1, \lambda_2, \ldots, \lambda_k]$ and sum of all value in λ is *1*
- Symbol: $p(x) = Cat_x[\lambda]$

Beta Distribution

Write later

Dirichlet Distribution

Write later