M0a. Background on Probability

Manikandan Narayanan

Week 1 (Jul 28-)

PRML Jul-Nov 2025 (Grads Section)

Acknowledgment of Sources

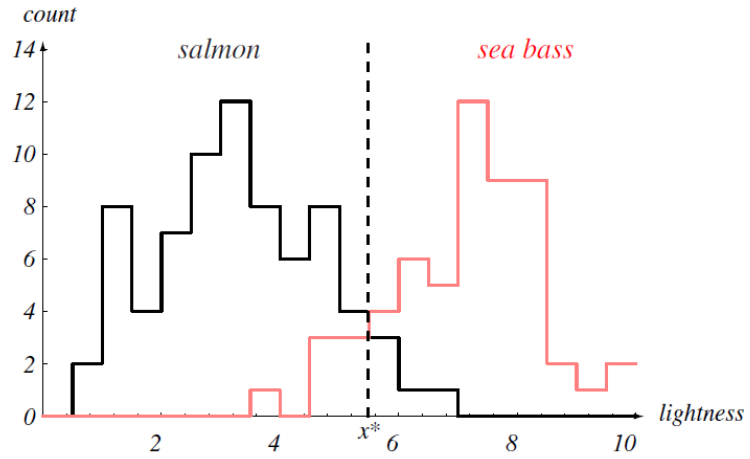
- Slides based on content from related
  - Courses:
    - IITM – Profs. Arun/**Harish**[HR]/Chandra[CC]/Prashanth's PRML offerings (slides, quizzes, notes, etc.), Prof. Ravi's "Intro to ML" slides – cited (e.g., **[HR]/[HG]**) in the bottom right of a slide.
    - India – NPTEL PR course by IISc Prof. PS. Sastry (slides, etc.) – cited as [PSS] in the bottom right of a slide.
  - Books:
    - PRML by **Bishop**. (content, figures, slides, etc.) – cited as **[CMB]**
    - Pattern Classification by Duda, Hart and Stork. (content, figures, etc.) – [DHS]
    - Mathematics for ML by Deisenroth, Faisal and Ong. (content, figures, etc.) – [DFO]
    - Foundations of ML by Mohri, Rostamizadeh, and Talwalkar (content, figures, slides by Mohri, etc.). – [MRT]

Outline of Module M0a

- M0a. **Background on Probability**
  - **M0a.1 Intuitive notion of joint, marginal, conditional probabilities, and Bayes rule**
  - **M0a.2 Intuitive notion of discrete and continuous random variables (r.v.s)**
  - M0a.3 Formal axiomatic introduction to probability (self review; brief tutorial exercises)
  - Appendix

[Why Probability?] Where we left our fish example last time: feature variation within/across classes

- Feature variation within a class, and separating it from noise is the challenge.
    - Holds for fish as well as other examples (e.g., wide variability in handwriting of digits)



- Language of probability needed to work with above uncertainty.

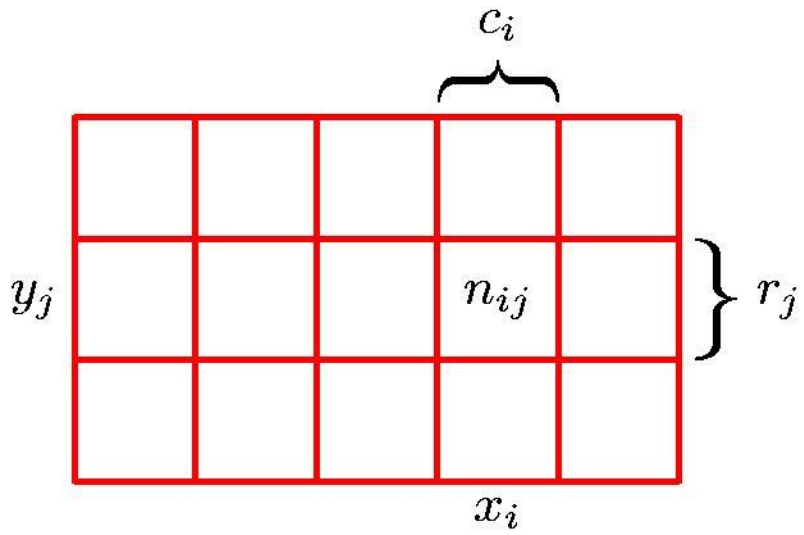
[DHS]

Interpretations of probability

- Frequentist (frequencies of events)
  - Events observed in a population (large popn. comprising all individuals), or a sample (of N individuals chosen randomly from the same population, for large enough N)
- Bayesian (subjective degrees of belief)
  - Degrees of belief in propositions can be mapped onto probabilities if they satisfy simple consistency rules known as the Cox axioms, which also leads to the rules of probability theory [see Section 1.2.3 “Bayesian probabilities” of CMB book, or Section 2.2 “The meaning of probability” of David JC MacKay’s book available at https://www.inference.org.uk/itila/].

- Axiomatic defn. same for both views

Probability Theory – Frequentist viewpoint



Sum Rule:

$$p(X = x_i) = \frac{c_i}{N} = \frac{1}{N}\sum_{j=1}^{L} n_{ij}$$

$$= \sum_{j=1}^{L} p(X = x_i, Y = y_j)$$

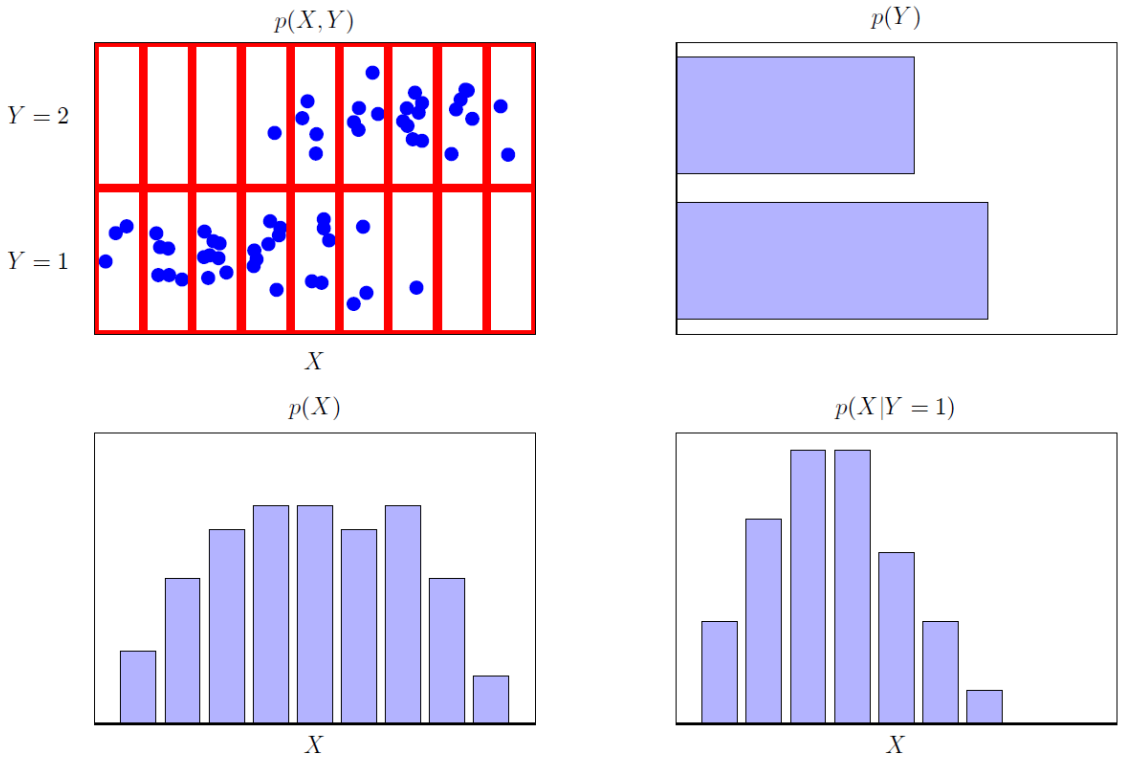
Product Rule:

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i} \cdot \frac{c_i}{N}$$

$$= p(Y = y_j | X = x_i) p(X = x_i)$$

[CMB]

An example



(example inspired by the fish PR problem)

[CMB]

The Rules of Probability

- Sum Rule $$p(X) = \sum_{Y} p(X, Y)$$
- Product Rule $$p(X, Y) = p(Y|X)p(X)$$

$= P(X|Y)\,P(Y)$

Note: Joint Distbn. $P(X,Y)$ captures everything.

[CMB]

Bayes' Theorem

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)} \qquad p(X) = \sum_Y p(X|Y)p(Y)$$

$$p(Y|X) \propto p(X|Y)p(Y)$$

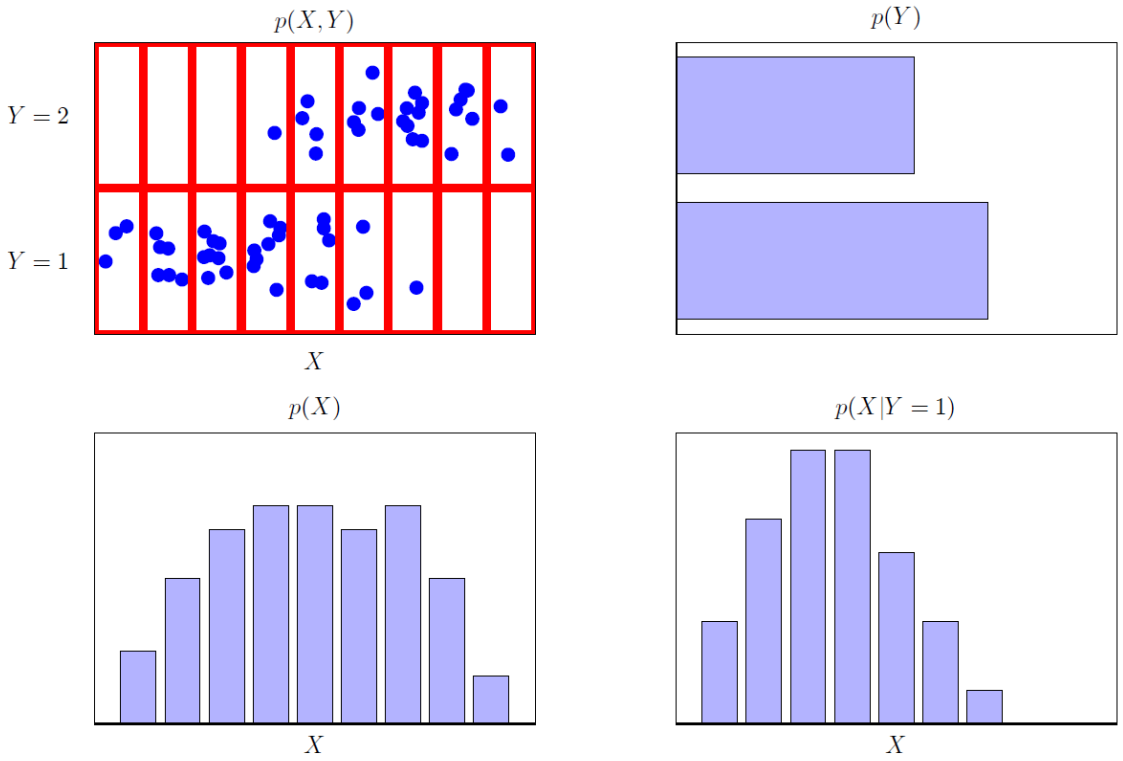
posterior $\propto$ likelihood × prior
(Bayesian terminology)

posterior $\propto$ class conditional × class prior
(supervised ML terminology for these distbns.)

$$P(Y=y \mid X=x) = \frac{P(X=x|Y=y)P(Y=y)}{P(X=x)}$$

[CMB]

An example (contd.)



(example inspired by the fish PR problem)

Tutorial exercise: Find Pr(Y|X=9) and Pr(Y|X=5).

(direct calculation using defn. of cdtnl. probab. is easier here;
Bayes' theorem can also be used as shown in next example)

[CMB]

Another example (tutorial exercise)

A basket contains solid objects of 3 colours (Red, Green and Blue) and 2 shapes (Sphere and Cube). The number of objects of different colors and shapes are given below.

Red Spheres: 20, Green Spheres: 10, Blue Spheres: 15

Red Cubes: 10, Green Cubes: 20, Blue Cubes: 25

Let the random variable X represent the color and the random variable Y represent the shape.

(a) Give the marginal distributions $P(X)$ and $P(Y)$

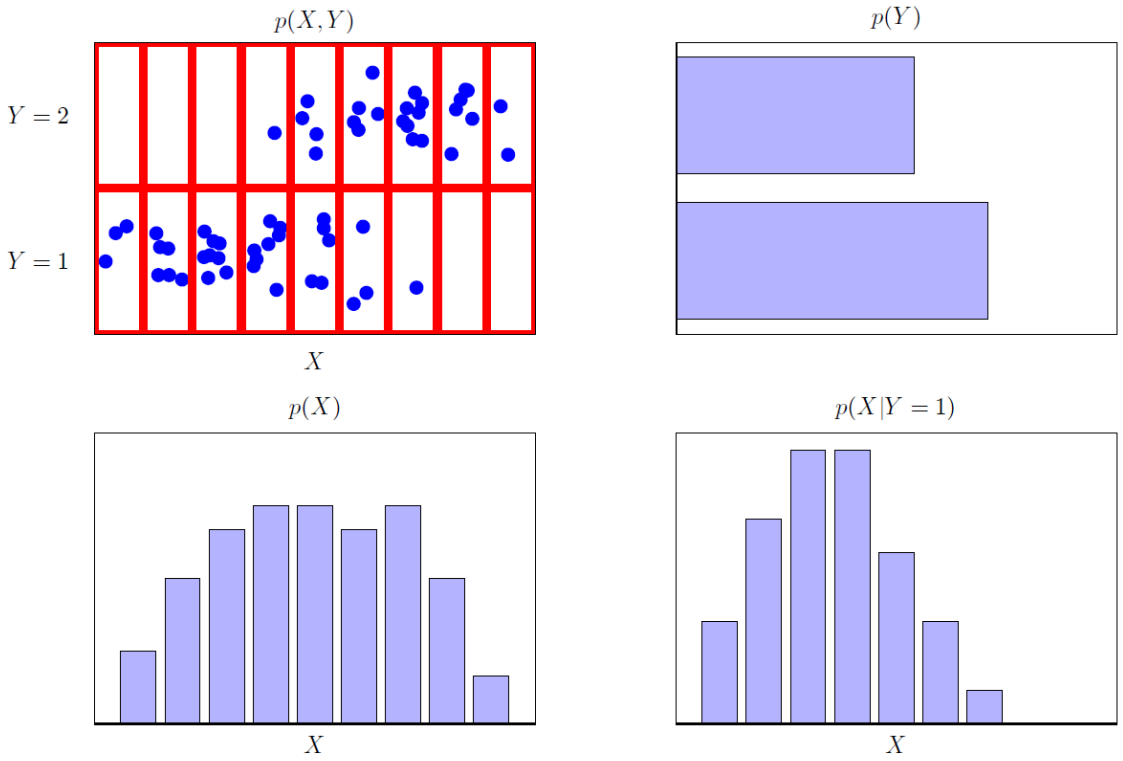
(b) Give the conditional distributions $P(X/Y = \text{Cubes})$ and $P(Y/X = \text{Blue})$

(c) Compute $P(Y = \text{Cube}/X = \text{Blue})$ using the Bayes' theorem

[CC]

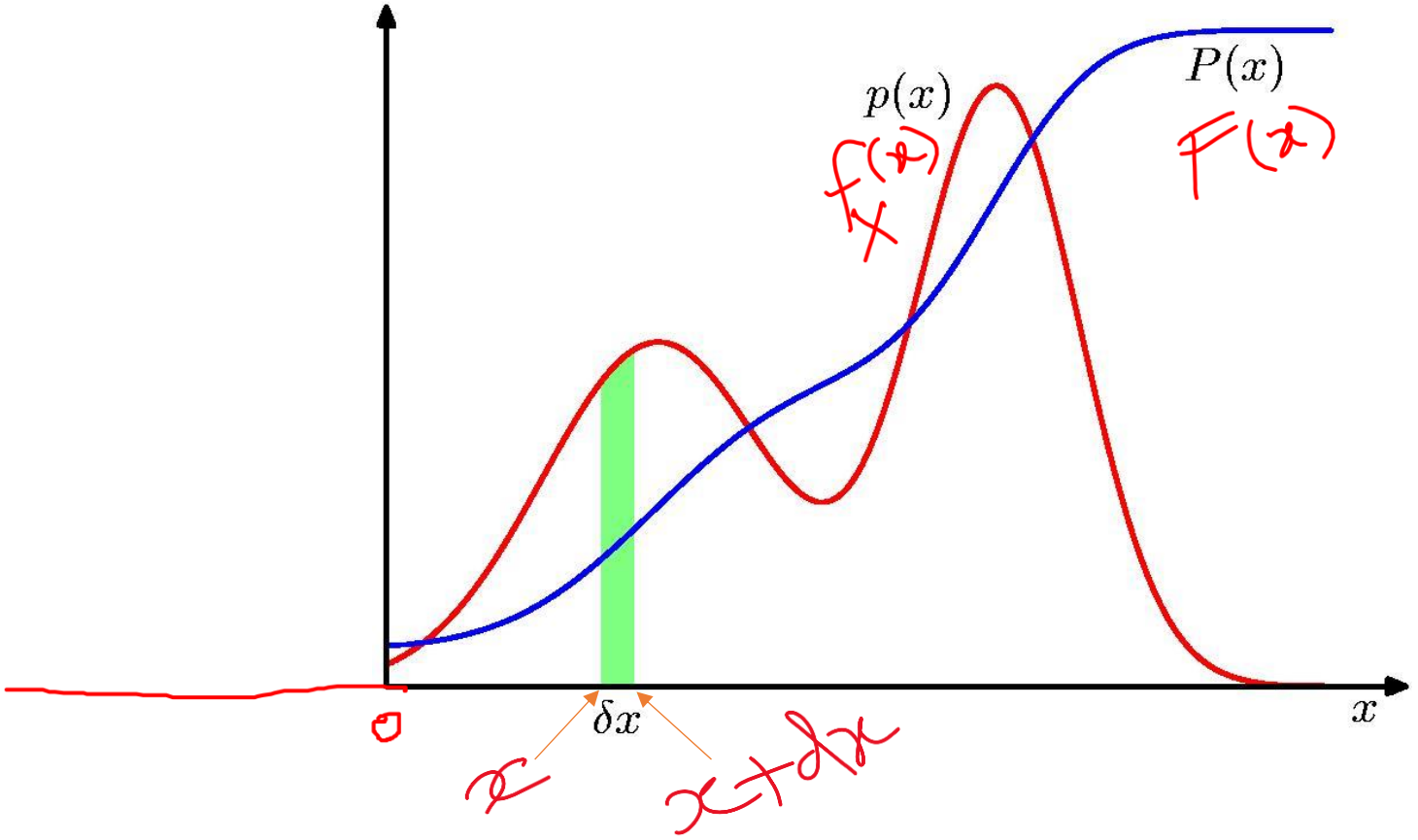
From discrete to continuous random variables (rv's)

Intuition: divide into smaller & smaller bins…



[CMB]

Probability densities: probability density function (pdf) and cumulative distribution function (cdf)



$$p(x) \geqslant 0 \qquad \int_{-\infty}^{\infty} p(x)\,\mathrm{d}x = 1$$

$$P(z) = \int_{-\infty}^{z} p(x)\,\mathrm{d}x$$

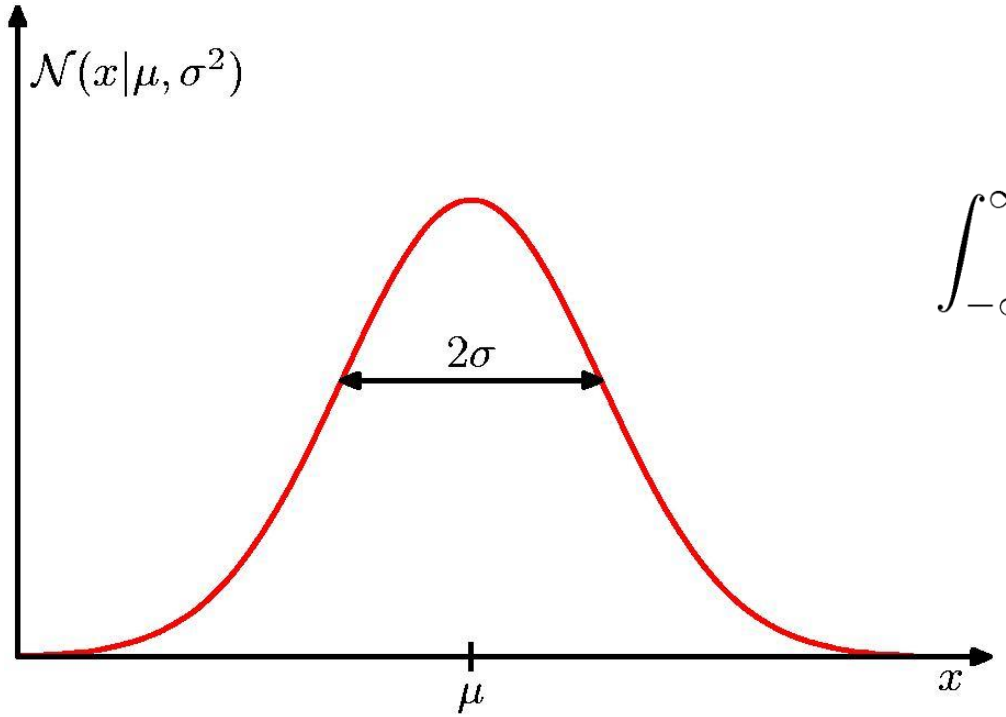
$$P_r[X \in [x, x+dx]] \approx p(x)dx \qquad p(x \in (a,b)) = \int_a^b p(x)\,\mathrm{d}x$$

[CMB]

Example: Gaussian Distribution

pdf

$p(x) := f_X(x) := \mathcal{N}\left(x|\mu,\sigma^2\right) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$



$$\mathcal{N}(x|\mu,\sigma^2) > 0$$

$$\int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right)\,\mathrm{d}x = 1$$

[CMB]

Expectations: Discrete rv X vs. Cont. rv X

	Discrete rv X	Cont. rv X
Expectation	$E[X] = \sum_x p(x)x$	$E[X] = \int_x p(x)\, x\, dx$
Expectation of fn. (law of the unconscious statistician)	$\mathbb{E}[f] = \sum_x p(x)f(x)$ $E[f(X)]$	$\mathbb{E}[f] = \int p(x)f(x)\,\mathrm{d}x$
Conditional Expectation	$E_{x\|y=y}[x] = \sum_x p(x\|y)x$ $\mathbb{E}_x[f\|y] = \sum_x p(x\|y)f(x)$ $E_{x\|y=y}[f(x)]$	$E_{x\|y=y}[x] = \int_x p(x\|y)\, x\, dx$ $E_{x\|y=y}[f(x)] = \int_x p(x\|y)f(x)dx$ Cdtnl. density $\frac{p(x,y)}{p(y)}$
Approximate Expectation (discrete and continuous)	$E[x] \approx \frac{1}{N}\sum_{n=1}^{N} x_n$ $\mathbb{E}[f] \simeq \frac{1}{N}\sum_{n=1}^{N} f(x_n)$ x_n iid sample	

[CMB]

Variance (in terms of expectation)

$$Var[X] = E[(X - E[X])^2] = E[X^2] - E[X]^2$$

$$\mathrm{var}[f] = \mathbb{E}\left[(f(x) - \mathbb{E}[f(x)])^2\right] = \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2$$

[CMB]

Example: Gaussian Mean and Variance

$$\mathbb{E}[x] = \int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right) x \,\mathrm{d}x = \mu$$

$$\mathbb{E}[x^2] = \int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right) x^2 \,\mathrm{d}x = \mu^2 + \sigma^2$$

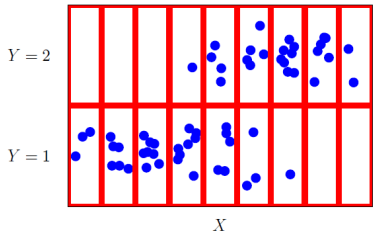
$$\mathrm{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \sigma^2$$

[CMB]

Conditional distbn./expectation: Example rvs

What is $p(X|Y=2)$ and $E[X|Y=2]$ for each of these cases?

- X discr., Y discr.: Already seen example →



- X cont., Y discr.: Let $p(X,Y) = p(Y)p(X|Y) = 0.5 \times \mathcal{N}(X \mid 2Y, \sigma^2)$

pmf pdf assuming uniform prior μ can be any fn. of y (here it is $2y$)

- X cont., Y cont.: Let W_1, W_2 be two indept. Gaussian rvs, i.e.,

$p(W_1) = \mathcal{N}(W_1 \mid \mu_1, \sigma^2)$, $p(W_2) = \mathcal{N}(W_2 \mid \mu_2, \sigma^2)$, & $p(W_1, W_2) = p(W_1)p(W_2)$.

  - Independent rvs: Let $X = W_1$, and $Y = W_2$
  - Dependent rvs: Let $X = W_1 + W_2$, and $Y = W_2$

Outline of Module M0a

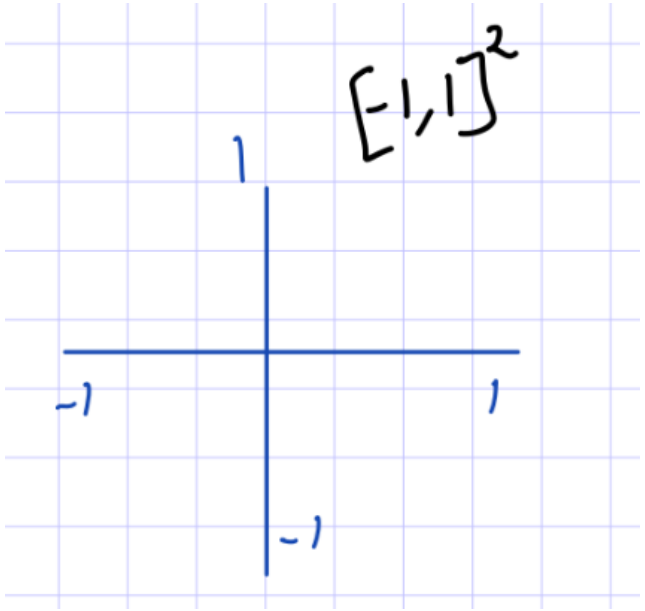
- M0a. **Background on Probability**
  - M0a.1 Intuitive notion of joint, marginal, conditional probabilities, and Bayes rule
  - M0a.2 Intuitive notion of discrete and continuous random variables (r.v.s)
  - **M0a.3 Formal axiomatic introduction to probability (self review; brief tutorial exercises)**
  - **Appendix**

Probability Theory – a more detailed review

- Switch to Harish's notes (N4_Probability.pdf)

A recap from [HG]Notes

Sets



$[d] = \{1, 2, \cdots, d\}$

Sets

$\mathbb{R}$ → Set of real numbers

$\mathbb{Z}$ → Set of Integers

$\mathbb{R}^d$ → Set of d-dimensional vectors

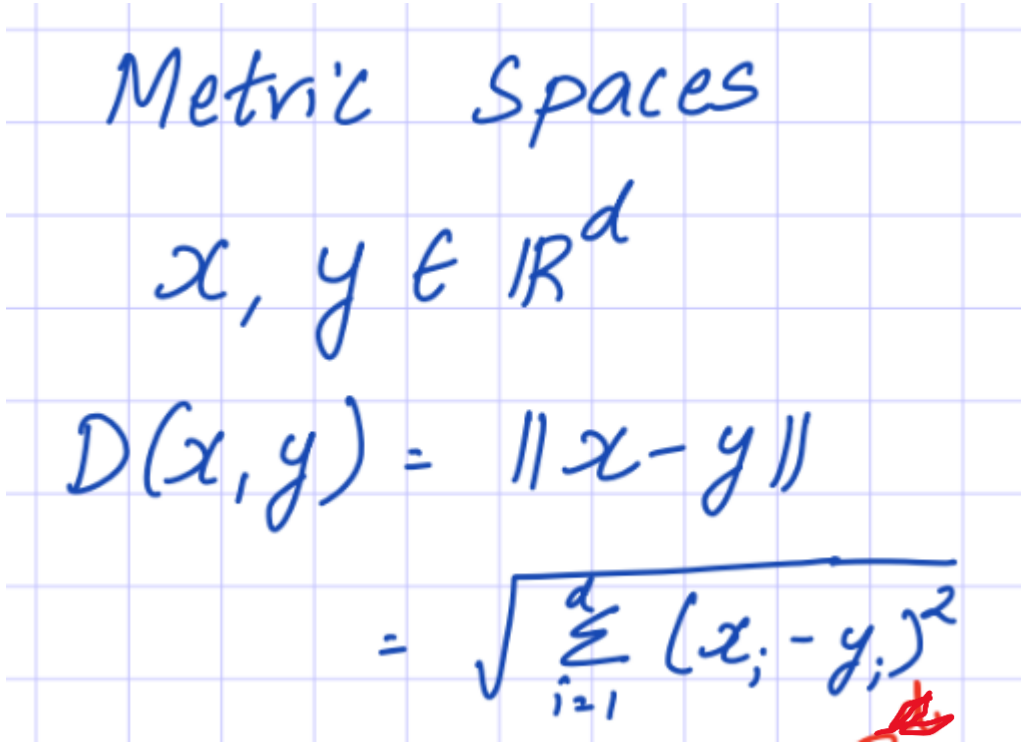
$[a,b] = \{x \in \mathbb{R} : a \leq x \leq b\}$

$(a,b) = \{x \in \mathbb{R} : a < x < b\}$

$[a,b]^d = \{x \in \mathbb{R}^d : a \leq x_i \leq b \quad \forall i \in [d]\}$

[HG]

Metric spaces



[HG]

Probability space

Each experiment

Probability space = $(\Omega, \mathcal{F}, P)$

Ω → Sample space

$\mathcal{F}$ → Collection of subsets of Ω ($\subseteq 2^{\Omega}$) (Power set)

P is a mapping from $\mathcal{F}$ to $[0,1]$

What properties must $\mathcal{F}$ satisfy: $\mathcal{F} \subseteq 2^{\Omega}$

(i) $\emptyset \in \mathcal{F}$

(ii) $\Omega \in \mathcal{F}$

(iii) $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$ ($\Omega \setminus A$)

(iv) $A, B \in \mathcal{F} \Rightarrow A \cup B \in \mathcal{F}$

What properties must P satisfy?

$P: \mathcal{F} \rightarrow [0,1]$

(i) $P(\emptyset) = 0$

(ii) $P(\Omega) = 1$

(iii) $P(A \cup B) = P(A) + P(B)$ if A & B are disjoint

[HG]

Examples

Example 1

$\Omega = \{HH, HT, TH, TT\}$

$\mathcal{F} = \{\emptyset,$
$\{HH\}, \{HT\}, \{TH\}, \{TT\},$
$\{HH, HT\}, \{HH, TH\}, \{HH, TT\}, \{HT, TH\}, \{HT, TT\}, \{TH, TT\},$
$\{HH, HT, TH\}, \{HH, HT, TT\}, \{HH, TH, TT\}, \{HT, TH, TT\},$
$\{HH, HT, TH, TT\}\}$

$P(A) = \frac{1}{4}|A|$

Example 2:

Ω and $\mathcal{F}$ same as above:

$P(A) = 0.81\, 1(HH \in A) + 0.01\, 1(TT \in A) +$
$0.09\, 1(HT \in A) + 0.09\, 1(TH \in A)$

[HG]

Events, etc.

Events

$A \subseteq \Omega$ is an event. (Technically $A \in \mathcal{F}$)

Two events A & B are independent if

$$P(A \cap B) = P(A) \cdot P(B)$$

Conditioning:

Probability of event A, given event B has happened is

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Bayes Rule

$$P(A \cap B) = P(A|B)\,P(B) = P(B|A)\,P(A)$$

[HG]

Independence of events

A,B indept. if

$P(A|B) = P(A)$, or equivalently $P(B|A) = P(B)$, or equivalently

$P(A \cap B) = P(A)\,P(B)$

Examples: Physical vs. Logical independence in dice experiment.

From events to rvs

Random Variables (rvs)

- **Discrete rv**
- Continuous rv

pmf properties:

non-negative,

adds up to 1.

$X : \Omega \to \mathbb{R}$

Eg: Let X be the number of heads. Then X expressed as a function is given by:

$X(HH) = 2$; $X(TH) = 1$

$X(HT) = 1$; $X(TT) = 0$

Probability Mass function example:

$f_X(x) = P(X = x)$

$= P(\{\omega \in \Omega : X(\omega) = x\})$

[HG]

Expectation (mean) and variance of discr. rv

$$EX = \sum x f_X(x) \quad \text{(captures the "average" value of } X\text{)}$$

$$\text{Var}[X] = E[(X-EX)^2] = E[X^2] - (EX)^2 \quad \text{(captures "variation")}$$

Linearity of expectation:

$$E[aX + bY] = aEX + bEY$$

[HG]

pmf examples

Discrete Distribution Examples:

(i) Bernoulli (θ) ; $X \in \{0,1\}$

$P(X=0) = 1-\theta$

$P(X=1) = \theta$

(ii) Binomial (n,θ) ; $X \in \{0,1,\cdots,n\}$

$P(X=m) = \binom{n}{m} \theta^m \cdot (1-\theta)^{n-m}$

(iii) Geometric (θ) ; $X \in \{1,2,\cdots\}$

$P(X=k) = (1-\theta)^{k-1} \cdot \theta$

Properties:

$X_1, \cdots, X_n \sim$ Bernoulli (θ), & independent

$X_1 + \cdots + X_n = Y \sim$ Binomial (n,θ)

Exercise:

i) What is the mean of Bernoulli(θ)?

ii) Use linearity of expectation to derive mean of Binomial(θ, n) distbn.

iii) What is the mean of Geometric(θ) ditbn.?

[HG]

Random Variables (rvs)

- Discrete rv
- **Continuous rv**

Continuous Random Variables:

$X : \Omega \to \mathbb{R}$ and Range (X) is uncountable

$F_X(x) = P(X \leq x)$ CDF or cumulative Distribution Function

Properties

$F_X(\infty) = 1$

$F_X(-\infty) = 0$

F_X is an increasing function.

$f_X(x) \geq \frac{P(X \in [x, x+dx])}{dx}$ (Note the difference to PMF)

Properties

$f_X(x) \geq 0$

$\int_{-\infty}^{\infty} f_X(x)\,dx = 1$

$\int_{-\infty}^{u} f_X(x)\,dx = F(u)$ f_X is the derivative of F_X & F_X is the anti-derivative of f_X

$F'(x) = f(x)$

[HG]

Expectation (mean) and variance of cont. rv

Expectation & Variance of a continuous RV

$$EX = \int_{x=-\infty}^{\infty} x\, f_x(x)\, dx$$

$$\mathrm{Var}[X] = E\left[(X-EX)^2\right] = E[X^2] - (EX)^2$$

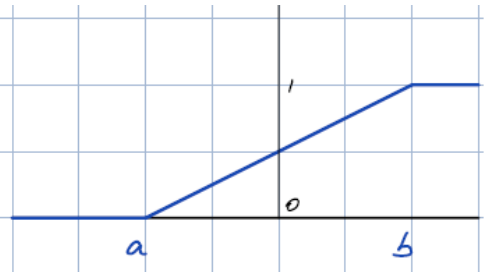
[HG]

pdf examples

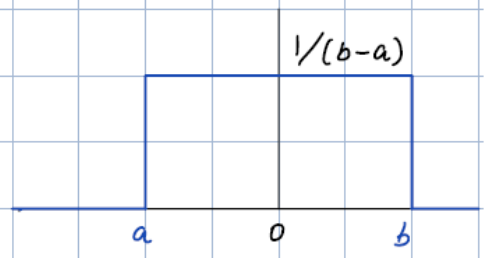
Uniform rv/pdf/distbn.

$X \sim \text{Unif}([a,b])$

$$F_X(x) = \begin{cases} 0 & \text{if } x \le a \\ \frac{x-a}{b-a} & \text{if } x \in [a,b] \\ 1 & \text{if } x \ge b \end{cases}$$



$$f_X(x) = \begin{cases} 0 & \text{if } x \le a \\ \frac{1}{b-a} & \text{if } x \in (a,b) \\ 0 & \text{if } x \ge b \end{cases}$$



P.T: $EX = \frac{a+b}{2}$

$\text{Var}[X] = \frac{(b-a)^2}{12}$

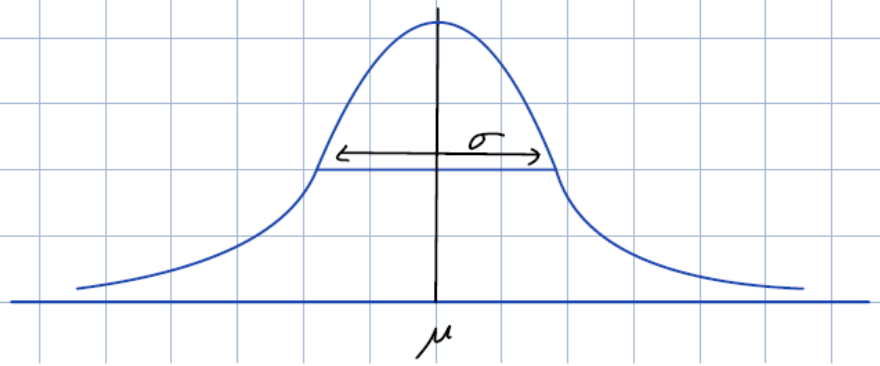
Normal rv/pdf/distbn.

$X \sim N(\mu, \sigma^2)$

$$\Downarrow$$

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-1}{2\sigma^2}(x-\mu)^2\right)$$

$E[X] = \mu$, $E[(X-\mu)^2] = \sigma^2$

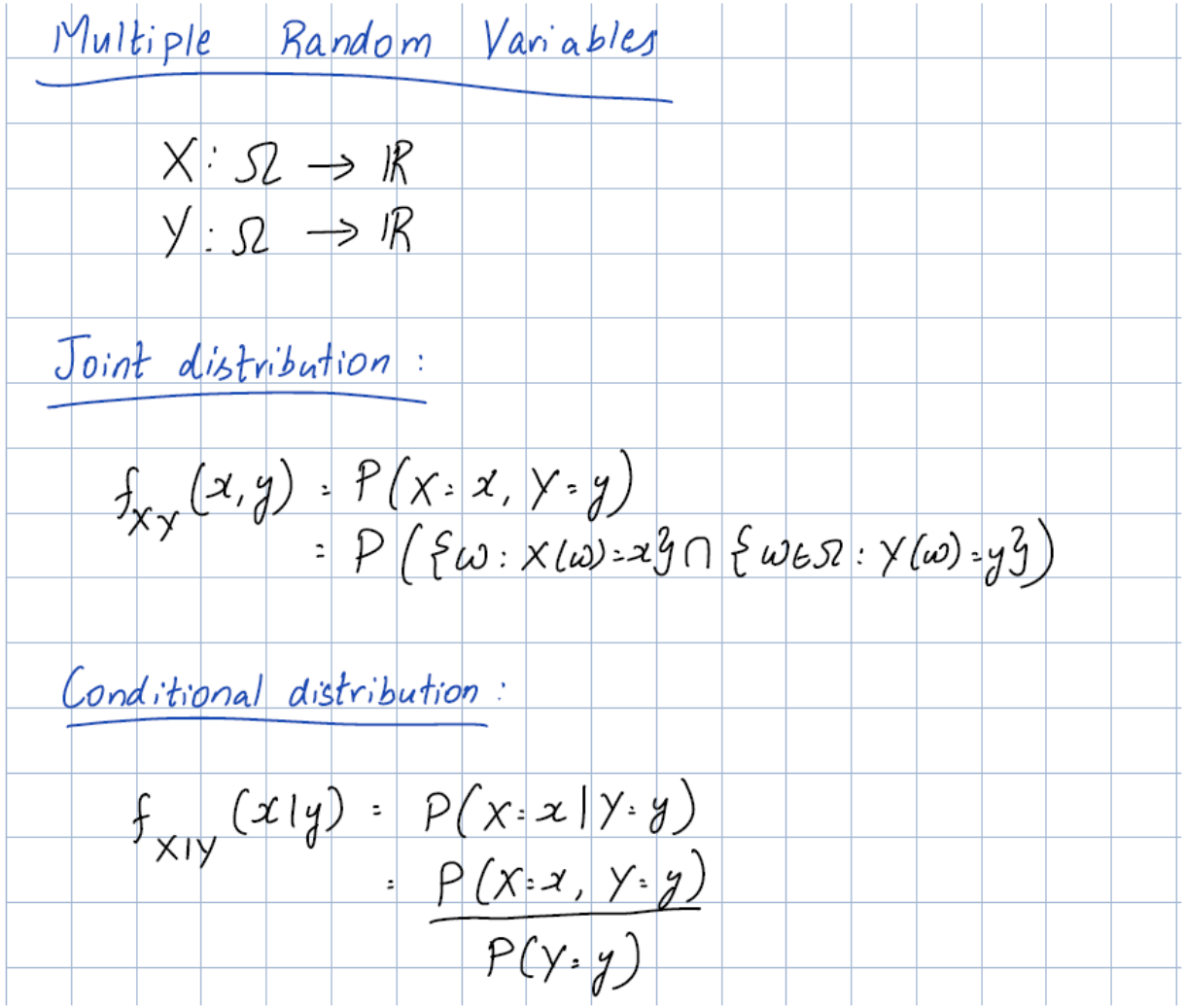


[HG]

From single to multiple (joint) rvs

Joint distbn.

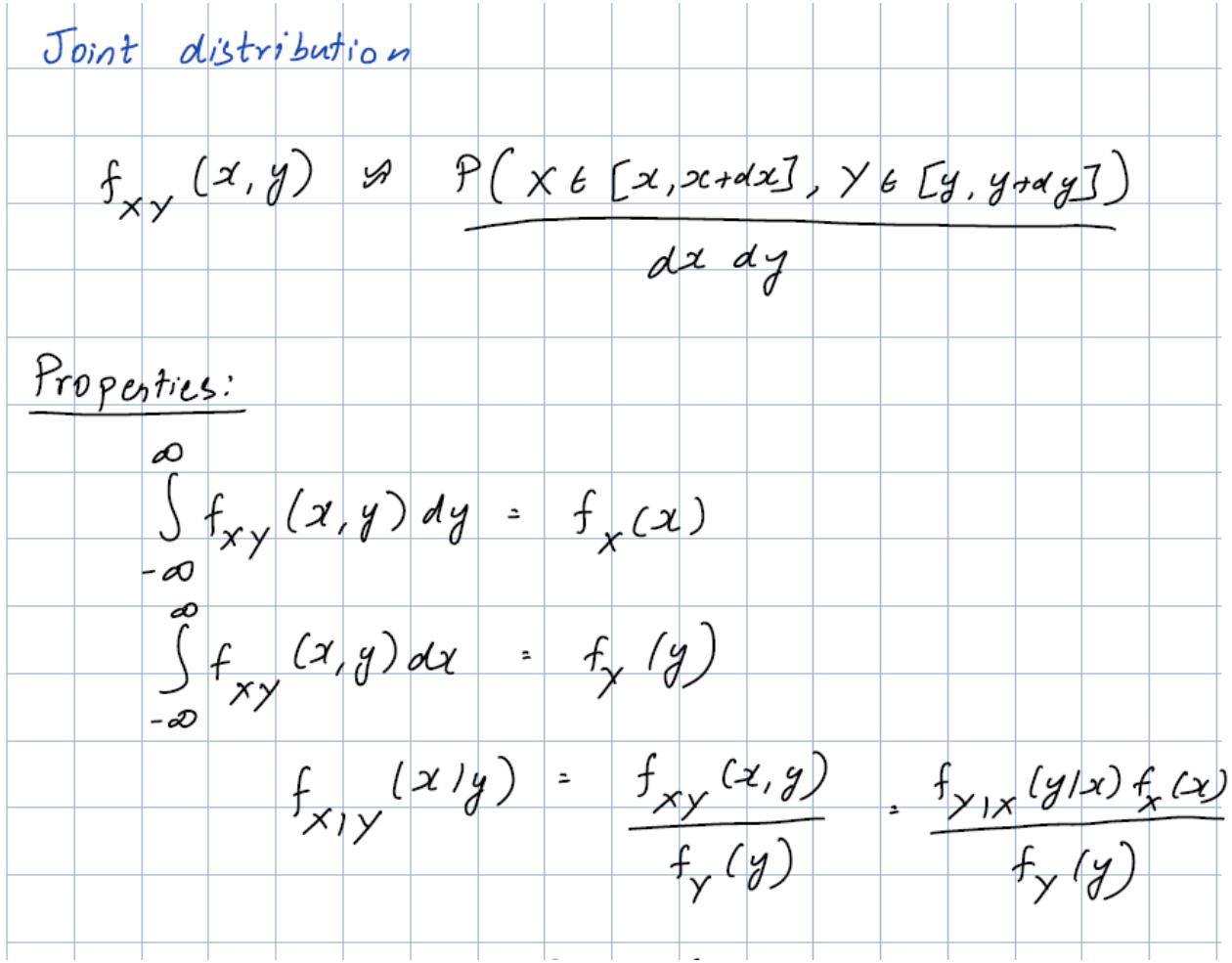
- **Discrete**
- Continuous



[HG]

Joint distbn.

- Discrete
- **Continuous**
  - **joint,**
  - **marginal, &**
  - **conditional distbns./densities**



[HG]

Properties

$E[aX + bY] = aE[X] + bE[Y]$ for any joint rvs X,Y

$Var[aX + bY] = a^2\,Var[X] + b^2\,Var[Y] + 2abCov(X,Y)$ for any joint rvs X, Y

$Var[aX + bY] = a^2\,Var[X] + b^2\,Var[Y]$ if X, Y uncorrelated (i.e., if $Cov(X,Y) = 0$)

Note: X, Y indept. ⇒ X, Y uncorrelated

Definition:

$Cov(X,Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\,]$ (or $\mathbb{E}[(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])(\boldsymbol{Y} - \mathbb{E}[\boldsymbol{Y}])^T]$ for random vectors)

⇒ $Var[X] \coloneqq E[(X - E[X])^2] = Cov[X,X]$ (≥ 0)

Independence of two rvs (joint rvs, i.e., defined on the same probab. space)

X,Y indept. if

$\forall \boldsymbol{x}, \boldsymbol{y}$

$f_{X|Y}(x|y) = f_X(x)$, or equivalently $f_{Y|X}(y|x) = f_Y(y)$, or equivalently

$f_{X,Y}(x,y) = f_X(x)\, f_Y(y)$

Note1: Above applies for both discrete and continuous rvs (because $f_{\cdot}(.)$ above denote pmf for discrete, and pdf for cont. rv).

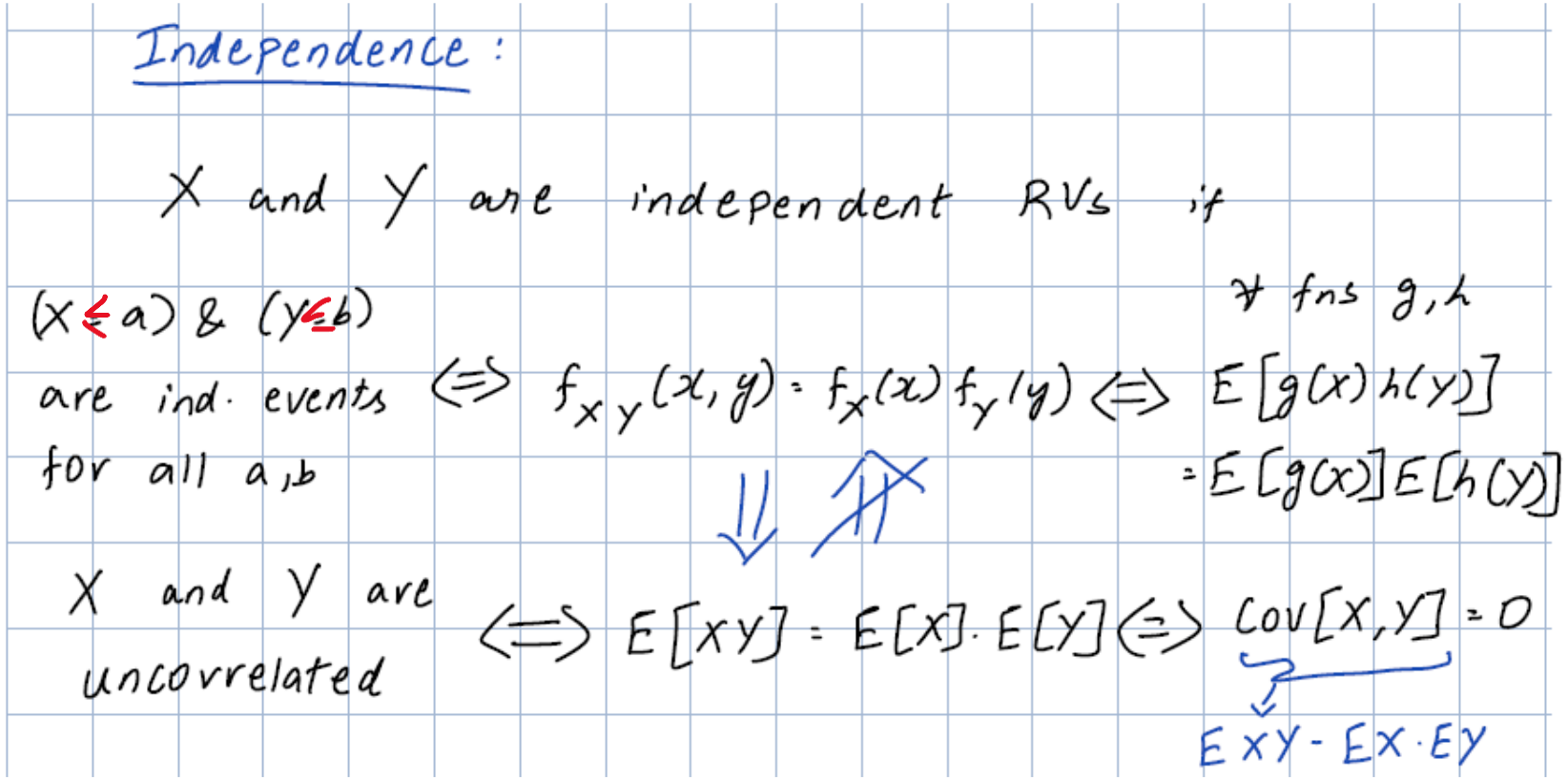
Note2: Compare with defn. of indep. of events, and also with alternate defn. of indep. of rvs below:

$\forall \boldsymbol{x}, \boldsymbol{y}$

$\{X = x\}$ indept. of $\{Y = y\}$ for discr. rvs.

$\{X \leq x\}$ indept. of $\{Y \leq y\}$ for cont. rvs

More on indep. of rvs



[HG]

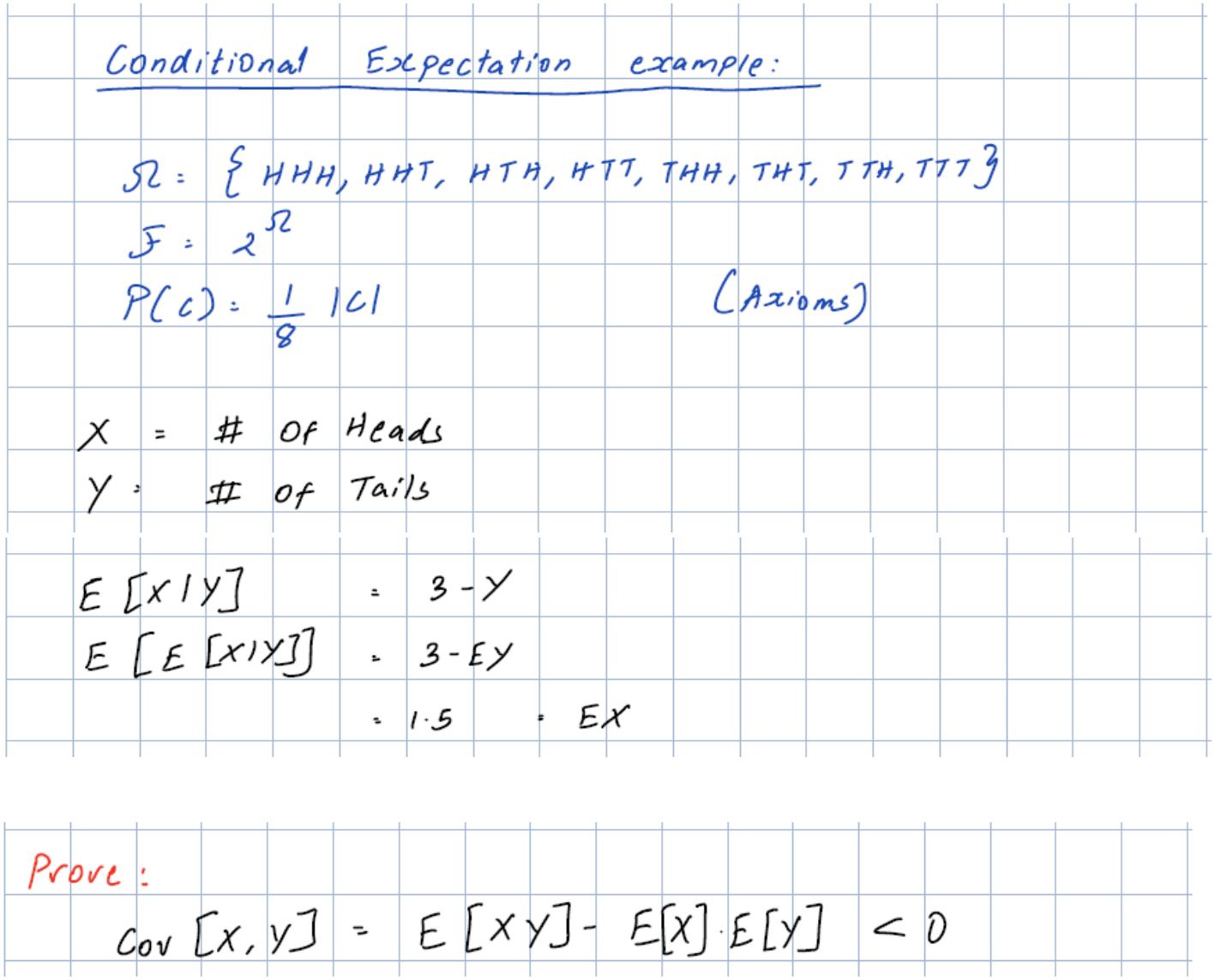
Conditional expectation…

…is simply the expectation of the conditional distribution.

- X Discr. : $E[X|Y=y] = \sum_x x\, f_{X|Y=y}(x)$
- X Cont. : $E[X|Y=y] = \int_x x\, f_{X|Y=y}(x)\, dx$

Being a function of y, it is a rv itself, i.e.,
$E[X|Y=y] = g(y)$, also denoted as $E[X|Y] = g(Y)$.

Conditional expectation example



[HG]

Property of conditional expectation

(Law of Total Expectation):

$$\mathbf{E}\big[\boldsymbol{E}[\boldsymbol{X}|\boldsymbol{Y}]\big] = \boldsymbol{E}[\boldsymbol{X}]$$

Proof sketch:

$$\mathrm{E}\big[E[X|Y]\big] \coloneqq \mathrm{E}_{\mathrm{Y}}\left[E_{X|Y}[X]\right]$$

$$= E_{X,Y}[X]$$

$$= E_X[X]$$

$$\coloneqq E[X]$$

Exercise: If $X \perp Y$, what is $E[X|Y]$?

Example Joint Density

Univariate → Bivariate → Multivariate Gaussian/Normal

Univariate Standard(/General) Normal

$Z \sim N(0,1)$ (Standard Normal)

$f_Z(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right)$

$E[Z] = 0$; $E[Z^2] = 1$

Let, $X = \sigma Z + \mu \Rightarrow EX = \mu$, $E[(X-\mu)^2] = \sigma^2$

$\Downarrow$

$X \sim N(\mu, \sigma^2)$

[HG]

Bivariate (standard/general) normal

Let $z = [z_1, z_2]$

$z_1 \sim N(0,1)$

$z_2 \sim N(0,1)$

Then

$$f_z(z) = \frac{1}{2\pi} \exp\left(-\frac{1}{2}(z_1^2 + z_2^2)\right)$$

$$z \sim N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\right)$$

$$E[z] = \begin{bmatrix}0\\0\end{bmatrix}, \quad E[zz^T] = \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$$

Let

$$X = WZ + \mu$$

$$E[X] = \mu \quad ;$$

$$\begin{aligned} E[(X-\mu)(X-\mu)^T] &= E[(WZ)(WZ)^T] \\ &= E[WZZ^TW^T] \\ &= W\,E[ZZ^T]\,W^T \\ &= WW^T \\ &= \Sigma \end{aligned}$$

$$X \sim N(\mu, \Sigma)$$

[HG]

Multivariate Gaussian (MVG), and **its density**!

What about density?

Let $Z \sim N(0_d, I_d)$

Let $X = WZ + \mu$

Then $X \sim N(\mu, \Sigma)$ where $\Sigma = WW^T$

& $f_X(x) = \frac{1}{\sqrt{|\Sigma|}(2\pi)^{d/2}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$

[HG]

How to derive conditional and marginal from joint MVG density?

Joint MVG density:

$$X \sim N(\mu, \Sigma) \qquad \mu \in \mathbb{R}^d, \quad \Sigma \in \mathbb{R}^{d \times d}$$

$$f_X(x) = \frac{1}{(2\pi)^{d/2}\sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$$

$$EX = \mu, \quad E[(X-\mu)(X-\mu)^T] = \Sigma$$

Let's look at Bivariate Gaussian first!

[HG]

A deeper look into bivariate normal

Let $d=2$, $\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$, $\Sigma^{-1} = \frac{1}{1-\rho^2}\begin{bmatrix} 1 & -\rho \\ -\rho & 1 \end{bmatrix}$

Recall earlier $X = WZ$ and $\Sigma = WW^T$

Let $\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$

- What all ρ are legal?
- What about $\begin{bmatrix} a & b \\ c & d \end{bmatrix} = \Sigma$, what all a,b,c,d are legal?
- In general for $\Sigma \in \mathbb{R}^{d \times d}$, when is it legal?
- Qn: why are other values "illegal"?

$$f_X(x) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2}\cdot\frac{1}{1-\rho^2}\cdot[x_1\ x_2]\begin{bmatrix} 1 & -\rho \\ -\rho & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right)$$

$$= \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(\frac{-1}{2}\cdot\frac{1}{1-\rho^2}\left(x_1^2 + x_2^2 - 2\rho x_1 x_2\right)\right)$$

$$= \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(\frac{-1}{2}\cdot\frac{1}{1-\rho^2}\left[(x_1-\rho x_2)^2 + (1-\rho^2)x_2^2\right]\right)$$

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}(x_1-\rho x_2)^2\right)\cdot\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}x_2^2\right)$$

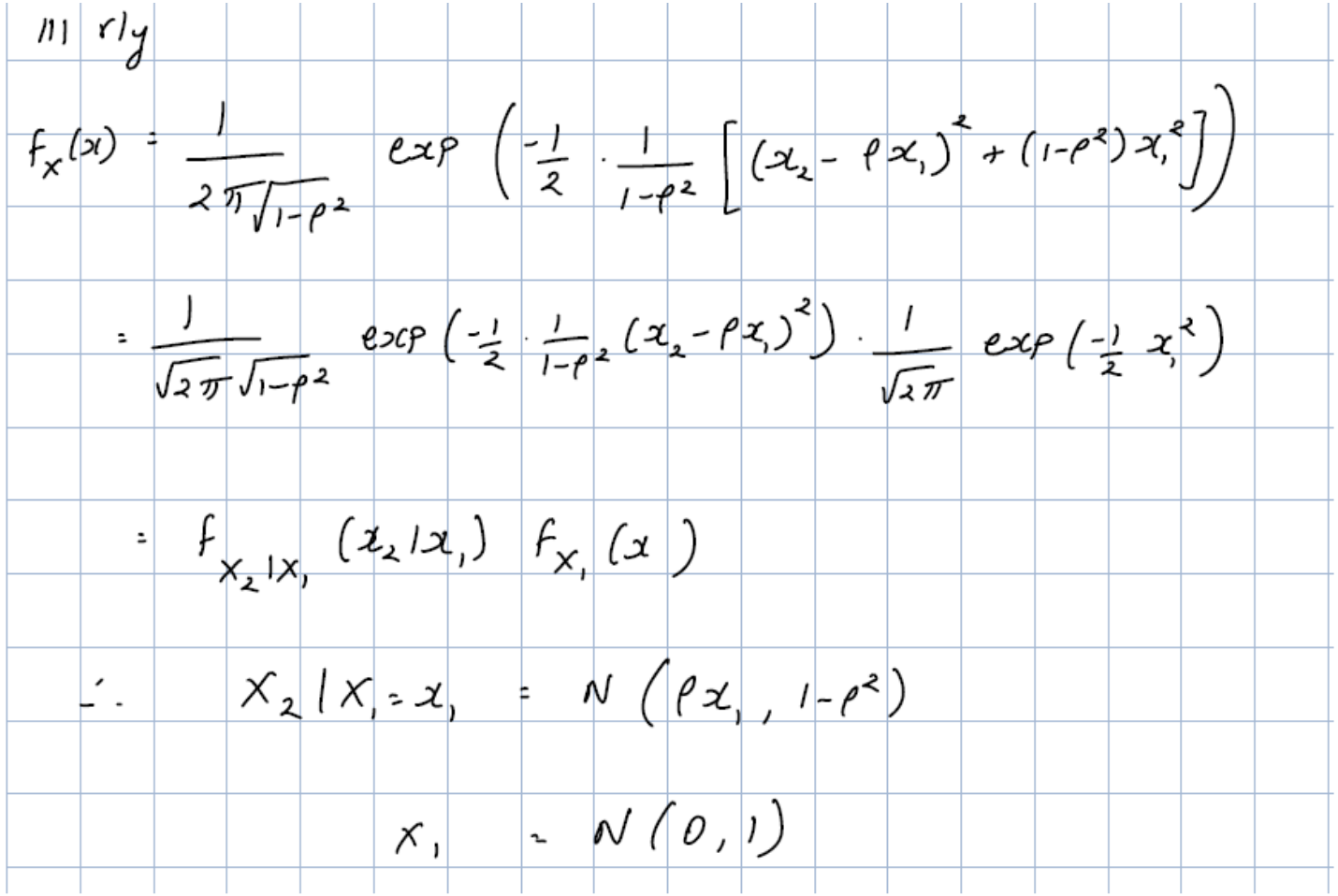
$$f_X(x) = f_{X_1|X_2}(x_1|x_2)\, f_{X_2}(x_2)$$

$\therefore$ $X_1|X_2 = x_2 \sim N(\rho x_2, 1-\rho^2)$

$X_2 \sim N(0,1)$

[HG]

A deeper look into bivariate normal (contd.)



[HG]

In summary, for bivariate normal:

Joint = Marginal x Conditional

If $\boldsymbol{X} = (X_1, X_2)$ & $\boldsymbol{x} = (x_1, x_2)$, then

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = f_{X_1}(x_1)\ f_{X_2|X_1}(x_2)$$

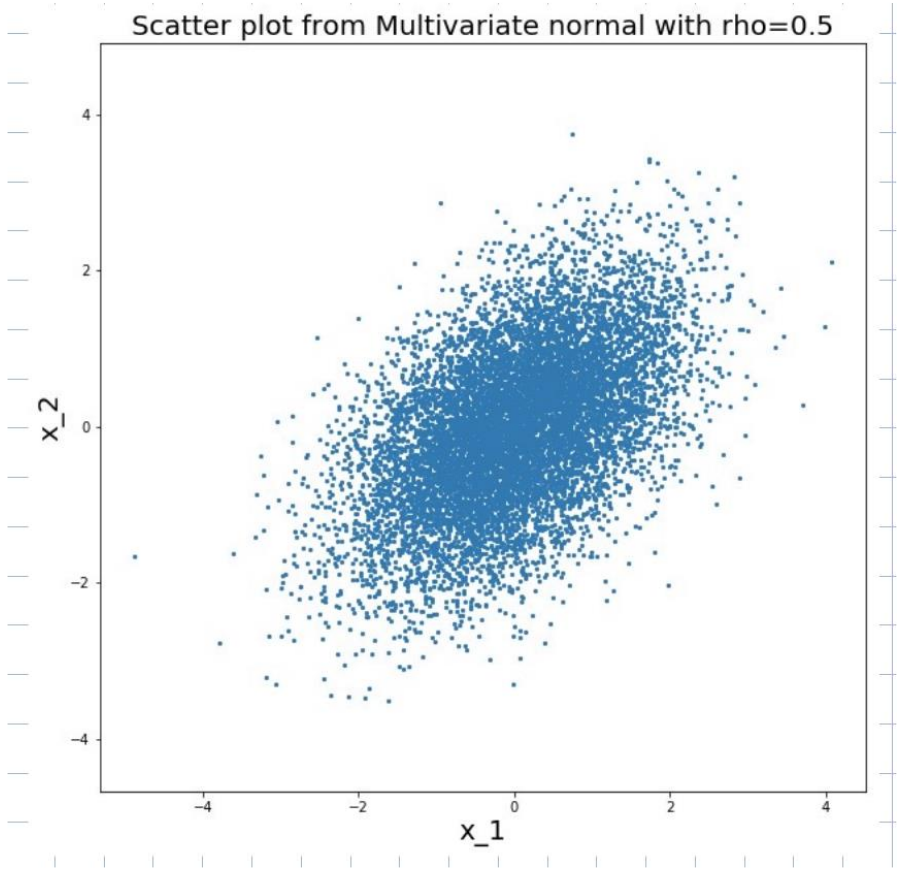
Joint: If $X \sim N([0\ 0\ ]^T, [1\ \rho; \rho\ 1])$, then

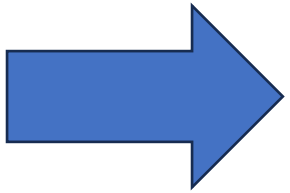
Marginal: $\boldsymbol{X_1 \sim N(0, 1)}$

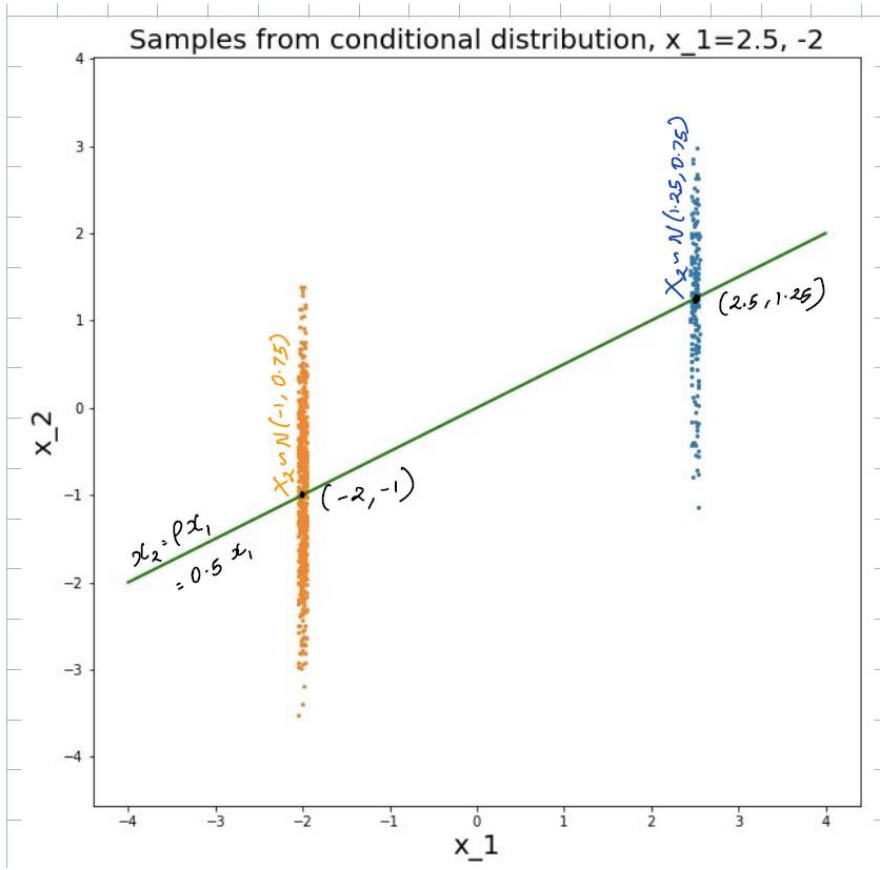
Conditional: $\boldsymbol{X_2 \mid X_1 \sim N(\rho x_1, 1 - \rho^2)}$

((results similar for conditioning on X_2))

Visualization of Conditional of Bivariate Gaussian

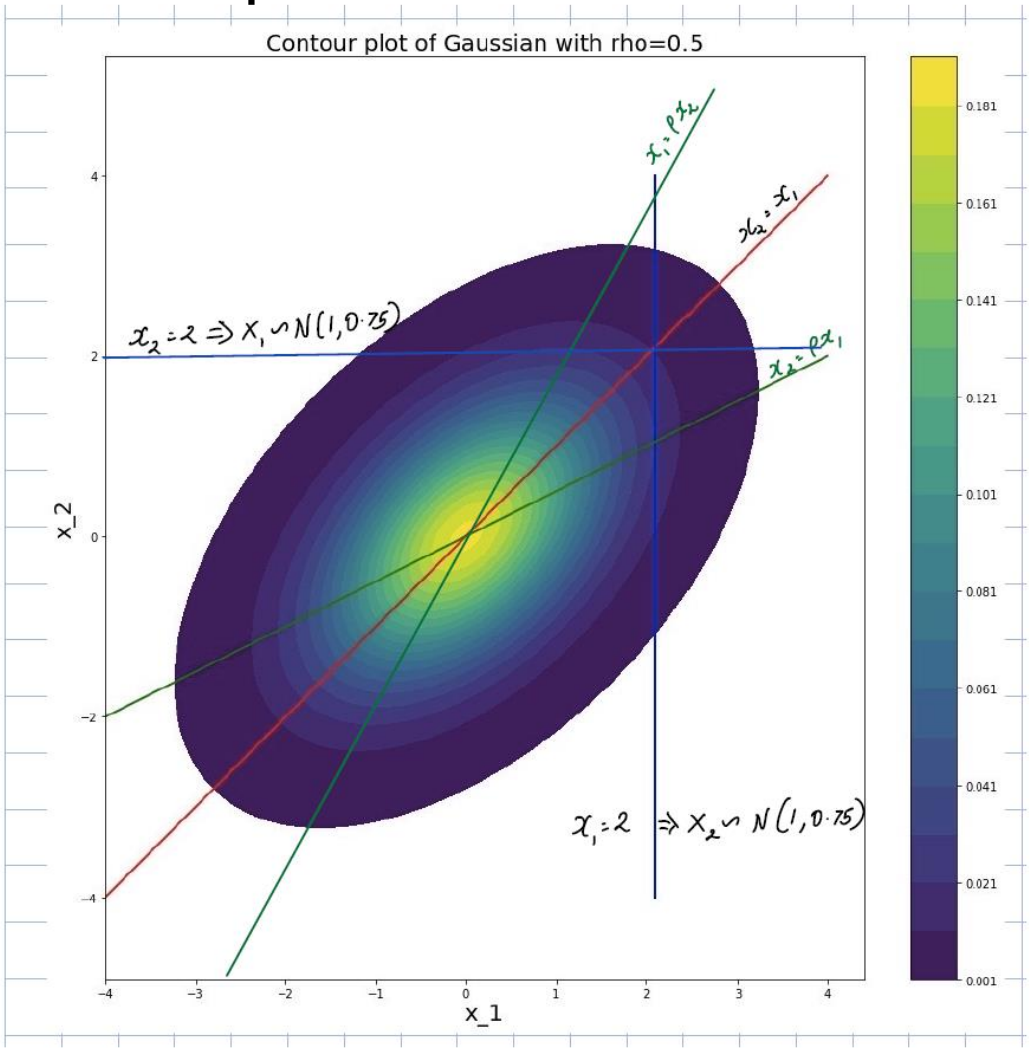






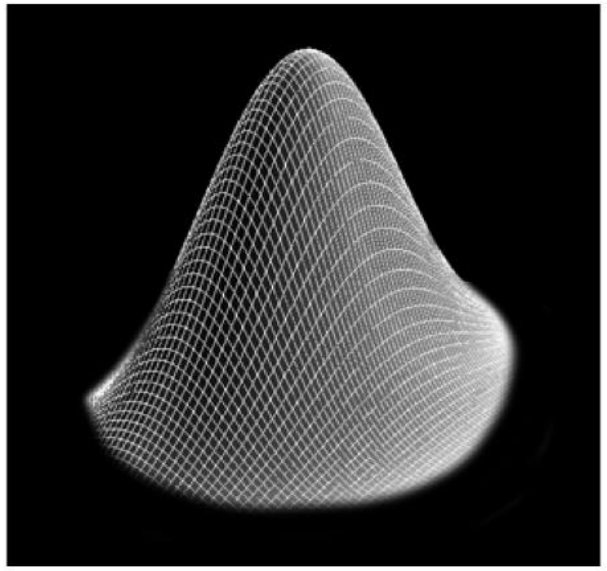
[HG]

Visualization of Conditional of Bivariate Gaussian (contd.) – contour plot



[HG]

Contour plot captures the multivariate density…



[From http://i.imgur.com/rrjJtoO.png, also http://www.oneweirdkerneltrick.com]

In a lighter-vein [joke]

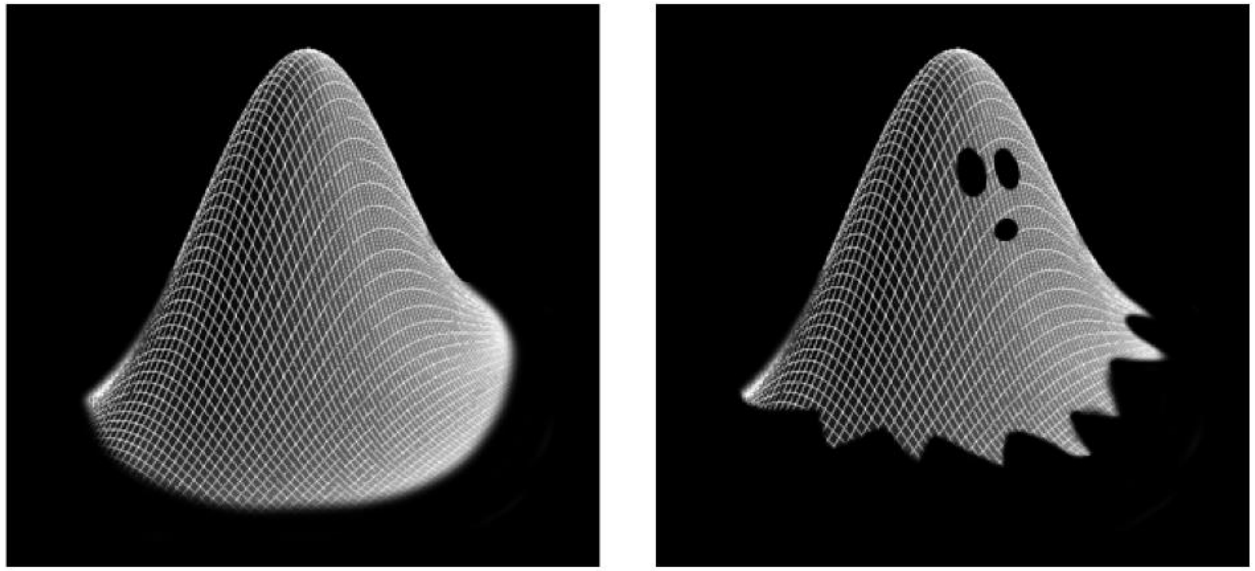
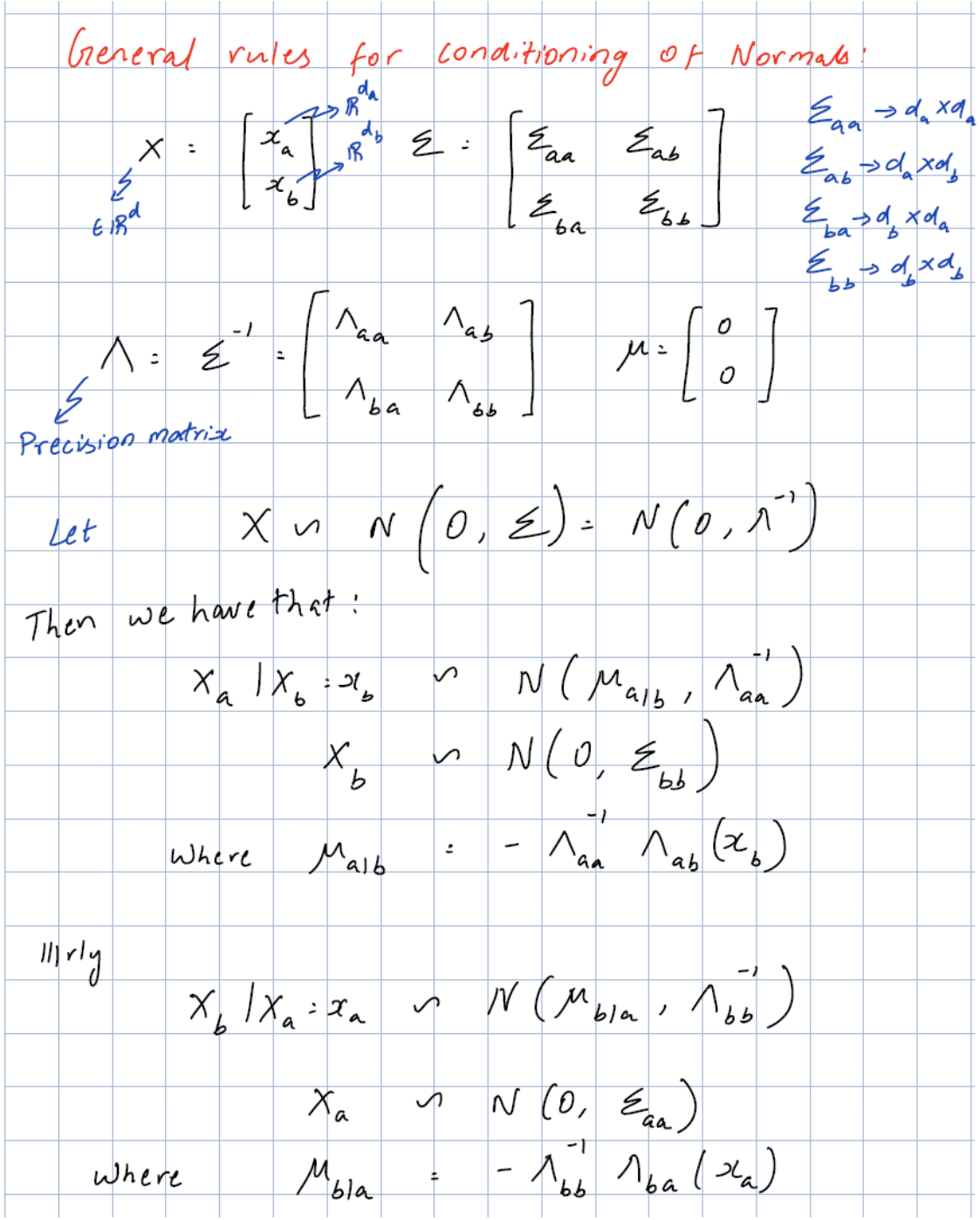


Fig. 4. The normal (left) and paranormal (right) distributions. [From http://i.imgur.com/rrjJtoO.png, also http://www.oneweirdkerneltrick.com]

From Bivariate to MVG:



[HG]

MVG Handy Results (cheat-sheet)

Partitioned Gaussians

Given a joint Gaussian distribution $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Lambda} \equiv \boldsymbol{\Sigma}^{-1}$ and

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix}, \qquad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix} \tag{2.94}$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix}, \qquad \boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{pmatrix}. \tag{2.95}$$

Conditional distribution:

$$p(\mathbf{x}_a|\mathbf{x}_b) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_{a|b}, \boldsymbol{\Lambda}_{aa}^{-1}) \tag{2.96}$$

$$\boldsymbol{\mu}_{a|b} = \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b). \tag{2.97}$$

Marginal distribution:

$$p(\mathbf{x}_a) = \mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_{aa}). \tag{2.98}$$

Marginal and Conditional Gaussians

Given a marginal Gaussian distribution for $\mathbf{x}$ and a conditional Gaussian distribution for $\mathbf{y}$ given $\mathbf{x}$ in the form

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) \tag{2.113}$$

$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1}) \tag{2.114}$$

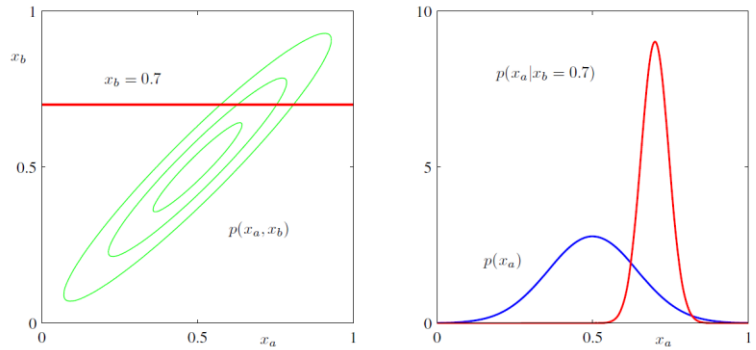
the marginal distribution of $\mathbf{y}$ and the conditional distribution of $\mathbf{x}$ given $\mathbf{y}$ are given by

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}) \tag{2.115}$$

$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\Sigma}\{\mathbf{A}^{\mathrm{T}}\mathbf{L}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\}, \boldsymbol{\Sigma}) \tag{2.116}$$

where

$$\boldsymbol{\Sigma} = (\boldsymbol{\Lambda} + \mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A})^{-1}. \tag{2.117}$$



[CMB: Bishop, Chapter 2]

Exercise: Use above formula to derive and therefore verify $p(X \mid Y = 2)$ in the example seen before, where $X = W_1 + W_2$, and $Y = W_2$ (and W_1, W_2 are indept. Gaussian rvs with mean μ_1, μ_2 respect. and variance σ^2).

(See Also) Appendix on …

- A) Transformed densities (pdf -- change-of-variables using Jacobian)
- B) Why Gaussian is a celebrated distbn.? (Central Limit Theorem (CLT), and related LLN)
- C) Conditional Expectation (more examples)

In summary, and next steps

- Probability theory gives a language to represent uncertainty or variability in data, and thereby forms a foundation of different ML problems.

- Next step: Let's see a concrete application in Decision theory (incl. Bayes classifier), and Density estimation (which captures the essence of what it means to learn from data!).

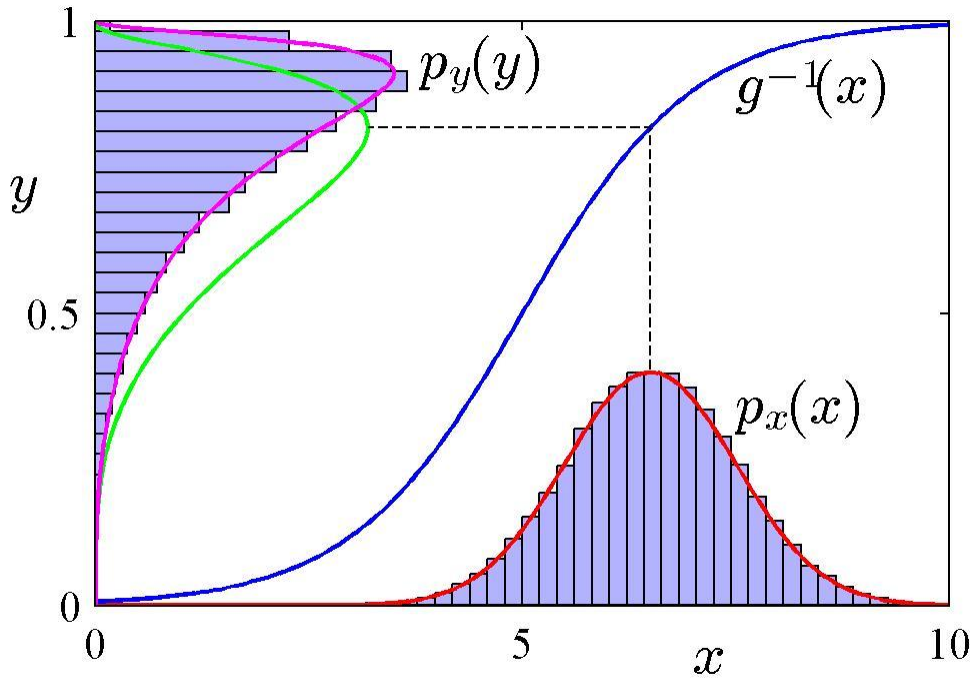
Thank you!

Appendix/Backup slides follow…

Appendix on…

- **A) Transformed densities (pdf -- change-of-variables using Jacobian)**
- B) Why Gaussian is a celebrated distbn.? (Central Limit Theorem (CLT), and related LLN)
- C) Conditional Expectation (examples)

Transformed Densities



$$
\begin{aligned}
p_y(y) &= p_x(x) \left| \frac{\mathrm{d}x}{\mathrm{d}y} \right| \\
&= p_x(g(y)) \, |g'(y)|
\end{aligned}
$$

[CMB]

Jacobian and change of variables

The Jacobian technique extends to higher dimensions. The transformation formula is a natural generalization of the two and three-dimensional cases:

$$f_{Y_1 Y_2 \cdots Y_n}(y_1, \ldots, y_n) = \frac{f_{X_1 \cdots X_n}(x_1, \ldots, x_n)}{|\partial(y_1, \ldots, y_n)/\partial(x_1, \ldots, x_n)|}$$

where

$$\frac{\partial(y_1, \ldots, y_n)}{\partial(x_1, \ldots, x_n)} = \begin{vmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_n} \\ & \vdots & \\ \frac{\partial y_n}{\partial x_1} & \cdots & \frac{\partial y_n}{\partial x_n} \end{vmatrix}.$$

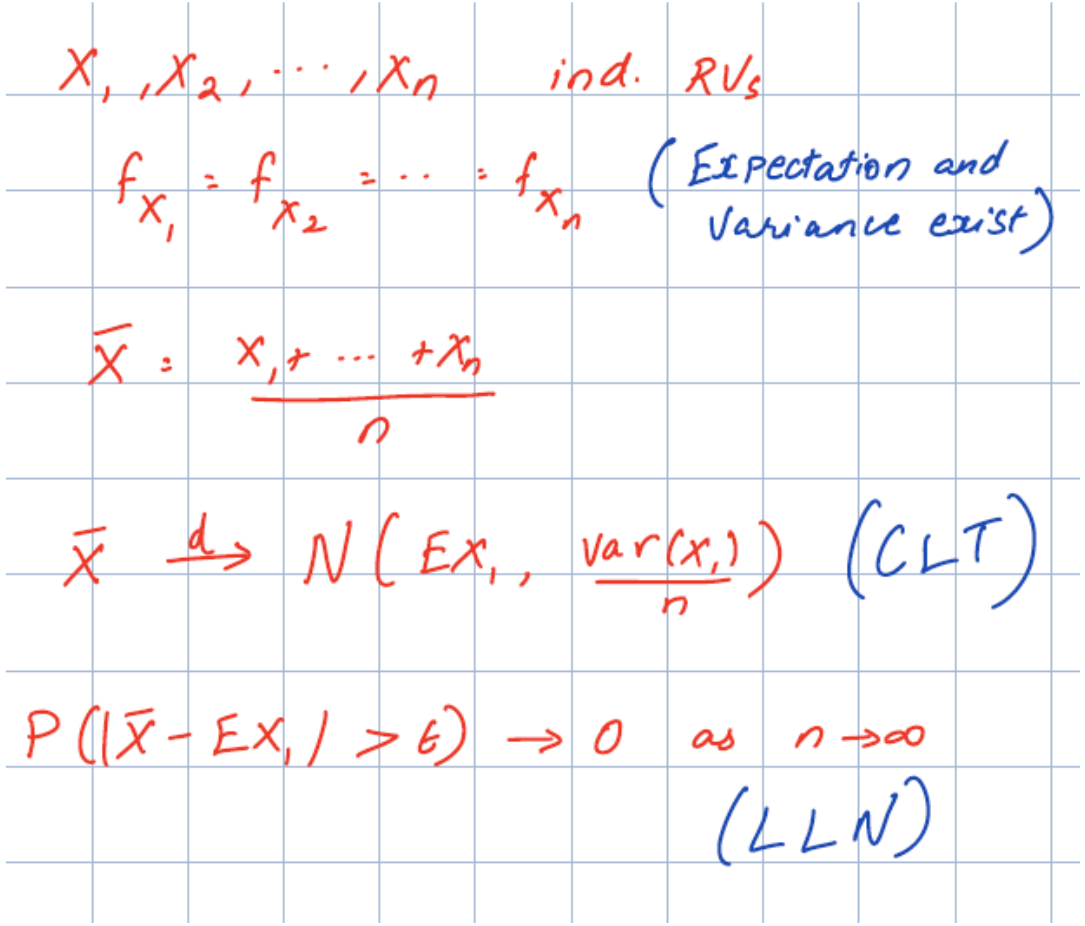
To help you remember the formula, think $f_Y(y)\,dy = f_X(x)dx$.

From https://faculty.math.illinois.edu/~r-ash/Stat/StatLec1-5.pdf

Appendix on…

- A) Transformed densities (pdf -- change-of-variables using Jacobian)
- **B) Why Gaussian is a celebrated distbn.? (Central Limit Theorem (CLT), and related LLN)**
- C) Conditional Expectation (examples)

Q: Why Gaussian is a preferred model for parametric density estimation? Brief Answer



[HR]

Q: Why Gaussian is a preferred model for parametric density estimation? Detailed Ans.

- A: Central Limit Theorem (CLT), along with tractability of the distribution and many known results.

- LLN (Law of Large Numbers) and CLT in next few slides (taken from Arun Rajkumar's course offering!)

Law of Large Numbers

Let $X_1, X_2, \ldots, X_n$ be i.i.d.r.v with finite mean μ and variance σ^2

Consider $S_n = \frac{X_1+X_2+\ldots+X_n}{n}$ Sample mean

$$E[S_n] = \frac{E[X_1+..+X_n]}{n} = \frac{n\mu}{n} = \mu$$

$$Var(S_n) = \frac{Var(X_1+\ldots+X_n)}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

$$E[S_n] = \mu \qquad\qquad Var(S_n) = \frac{\sigma^2}{n}$$

Applying Chebyshev inequality

for any $\epsilon > 0$

$$Pr(|S_n - \mu| \geq \epsilon) \leq \frac{\sigma^2/n}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \xrightarrow[n \to \infty]{} 0$$

Weak Law of Large Numbers

Let X_1, X_2,...., X_n be a sequence of independent and identically distributed random variables, each having a finite mean μ , Then for any $\varepsilon > 0$,

$$P\{|\frac{X_1+X_2+..+X_n}{n} - \mu| \geq \epsilon\} \to 0 \text{ as } n \to \infty$$

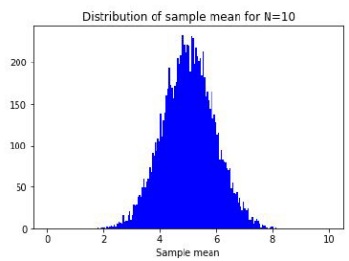
Strong Law of Large Numbers

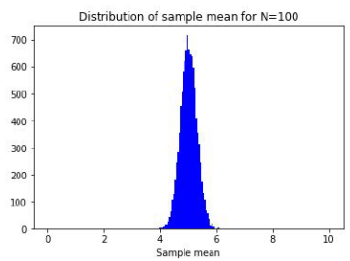
Let X_1, X_2,...., X_n be a sequence of independent and identically distributed random variables, each having a finite mean μ , Then with probability 1,

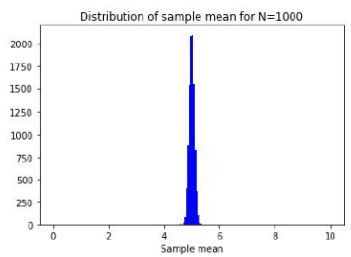
$$\frac{X_1+X_2+...+X_n}{n} \rightarrow \mu \quad \text{as} \quad n \rightarrow \infty$$

Interpreting Law of Large Numbers

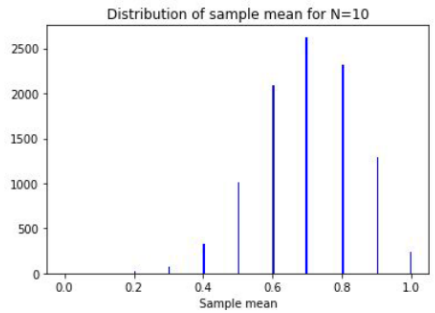
- Experiment to find an estimate of the mean of a distribution(X_i are samples from the distribution)
  - Eg: Uniform(0,10)

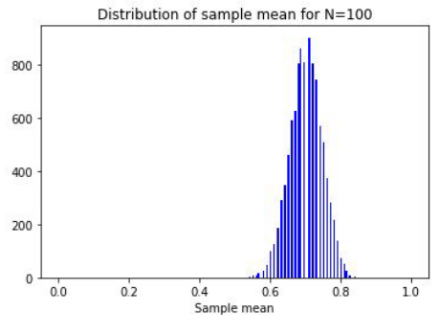


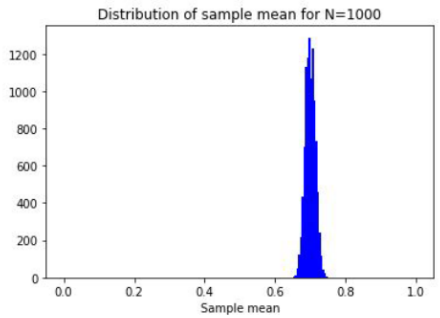




- Finding an estimate of probability of occurrence of an event E. (X_i as indicator random variable for the event.)
  - Eg: Finding probability of event (X=1) for Bernoulli(0.7)







Central Limit Theorem

Let $X_1, X_2, \ldots, X_n$ be i.i.d.r.v with finite mean μ and variance σ^2

$$S_n = X_1 + \ldots + X_n \qquad \text{Variance: } n\sigma^2$$

$$\frac{S_n}{n} = \frac{X_1 + \ldots + X_n}{n} \qquad \text{Variance: } \frac{\sigma^2}{n}$$

$$\frac{S_n}{\sqrt{n}} = \frac{X_1 + \ldots + X_n}{\sqrt{n}} \qquad \text{Variance: } \sigma^2$$

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{X_1 + \ldots + X_n - n\mu}{\sigma\sqrt{n}} \qquad E[Z_n] = 0, Var(Z_n) = 1$$

Central Limit Theorem

Let X_1, X_2,...., X_n be a sequence of independent and identically distributed random variables, each having a finite mean μ , and variance σ^2 , Then the distribution of

$$\frac{X_1+...+X_n-n\mu}{\sigma\sqrt{n}}$$

tends to the **standard normal** as $n \to \infty$. That is for $-\infty < a < \infty$,

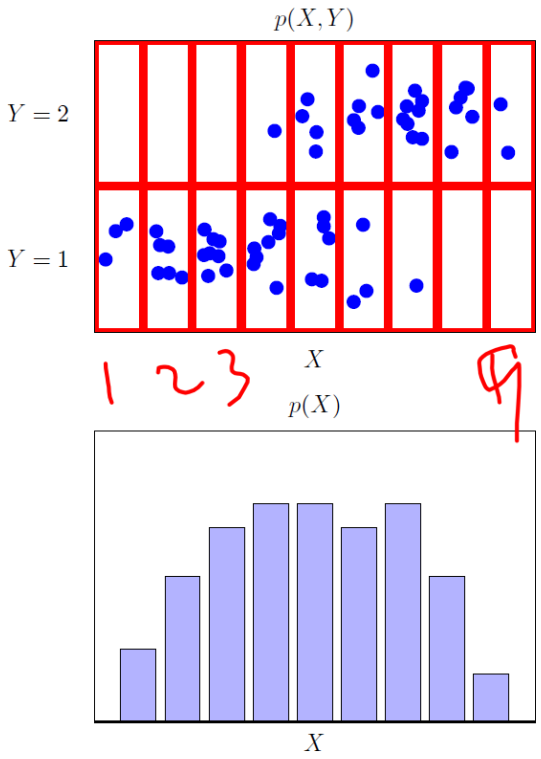
$$P\{\frac{X_1+...+X_n-n\mu}{\sigma\sqrt{n}} \leq a\} \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-x^2/2} dx \quad \text{as} \quad n \to \infty$$

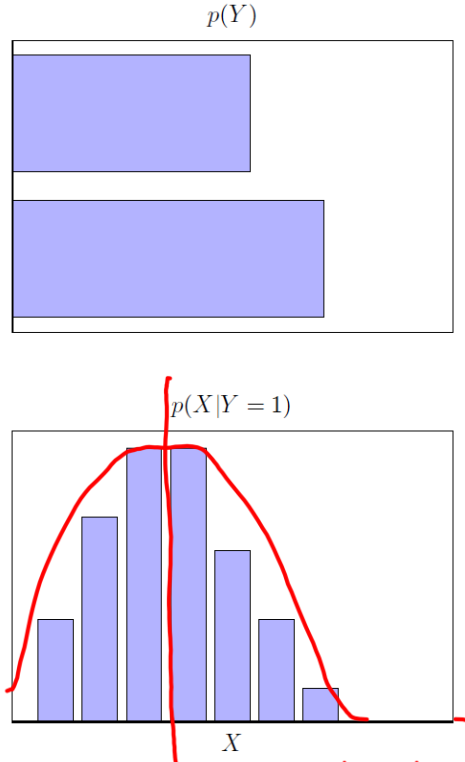
- **CLT in practice - Normal approximations**
  - We can treat Z_n as a standard normal random variable
  - So $S_n = \sqrt{n}\sigma Z_n + n\mu$ can also be treated as normal random variable
  - Hence $S_n \approx N(n\mu, n\sigma^2)$ and sample mean $\hat{X}_n \approx N(\mu, \frac{\sigma^2}{n})$
- **Can value of n be moderate?**
  - Usually yes
  - Symmetry and closeness to Normal distribution.
  - Unimodal (Single peak)

Appendix on…

- A) Transformed densities (pdf -- change-of-variables using Jacobian)
- B) Why Gaussian is a celebrated distbn.? (Central Limit Theorem (CLT), and related LLN)
- **C) Conditional Expectation (more examples)**

An example





$E(X|Y=y) = F(y)$

$p(X|Y=2)$

$E(X|Y=1) \approx 3.5$

$E(X|Y=2) \approx 7$

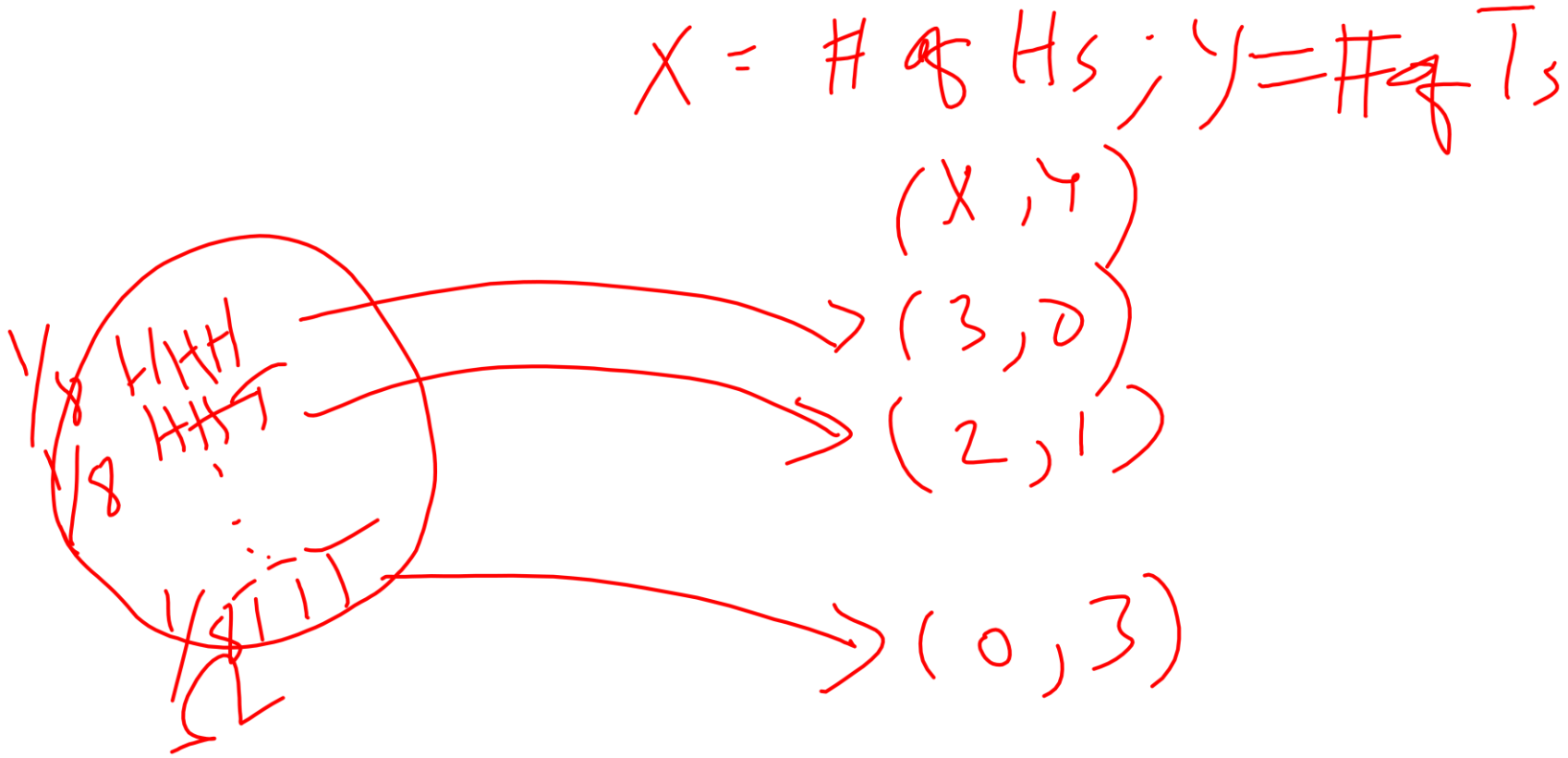
[CMB]

Another example: Expectation of Geometric(θ) r.v.

Exercise: Use conditional expectation property to derive the mean of a rv X that follows the Geometric(θ) distbn.

(Hint: Let $Y = 1$ if first toss is H, and 0 ow. Then, use Law of Total Expectation, i.e., compute $E[E[X|Y]]$ to obtain $E[X]$.)

Another example (3 coin tosses): E(X | Y=y)



$E(X \mid Y=y) = E_X(X \mid Y=y)$

$= \sum_{x=0,1,2,3} x \cdot P(X=x \mid Y=y)$

$= (3-y) \cdot 1$

$E(X|y) = 3-y$

$E(X \mid Y=y) = E_X(X \mid Y=y)$

$= E_X(3-y)$

$= 3-y$

$E(X \mid y) = 3-y$

Backup slides

Covariance

$$
\begin{aligned}
\operatorname{cov}[x, y] &= \mathbb{E}_{x,y}\left[\{x - \mathbb{E}[x]\}\{y - \mathbb{E}[y]\}\right] \\
&= \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y]
\end{aligned}
$$

$$
\begin{aligned}
\operatorname{cov}[\mathbf{x}, \mathbf{y}] &= \mathbb{E}_{\mathbf{x},\mathbf{y}}\left[\{\mathbf{x} - \mathbb{E}[\mathbf{x}]\}\{\mathbf{y}^{\mathrm{T}} - \mathbb{E}[\mathbf{y}^{\mathrm{T}}]\}\right] \\
&= \mathbb{E}_{\mathbf{x},\mathbf{y}}[\mathbf{x}\mathbf{y}^{\mathrm{T}}] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{y}^{\mathrm{T}}]
\end{aligned}
$$

$$\mathbb{E}[\,(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])\big(\boldsymbol{Y} - \mathbb{E}(\boldsymbol{Y})\big)^{T}]$$

[CMB]