12: Continuous Joint Distributions

David Varodayan

February 3, 2020

Adapted from slides by Lisa Yan

Hello dice rolls, my old friends

A 6-sided die is rolled 7 times.

What is the probability of getting:

- 1 one
- 1 two
- 0 threes
- 2 fours
- 0 fives
- 3 sixes

# of times a six appears

$$P(X_1 = 1, X_2 = 1, X_3 = 0, X_4 = 2, X_5 = 0, X_6 = 3)$$

$$= \binom{7}{1,1,0,2,0,3} \left(\frac{1}{6}\right)^1 \left(\frac{1}{6}\right)^1 \left(\frac{1}{6}\right)^0 \left(\frac{1}{6}\right)^2 \left(\frac{1}{6}\right)^0 \left(\frac{1}{6}\right)^3 = 420 \left(\frac{1}{6}\right)^7$$

choose where the sixes appear

probability of rolling a six

this many times

Today's plan

- Text analysis
- Continuous joint distribution
- Joint cumulative distribution functions (CDFs)

Probabilistic text analysis

Ignoring the order of words…

What is the probability of any given word that you write in English?

- $P(\text{word} = \text{“the”}) > P(\text{word} = \text{“multinomial”})$
- $P(\text{word} = \text{“Stanford”}) > P(\text{word} = \text{“Cal”})$

Probabilities of *counts* of words = Multinomial distribution

A document is a large multinomial.

(according to the Global Language Monitor,
there are 988,968 words in the English language
used on the internet.)

Probabilistic text analysis

Probabilities of *counts* of words = Multinomial distribution

Example document:

#words: $n = 48$

"When my late husband was alive he deposited some amount of Money with china Bank in which the amount will be declared to you once you respond to this message indicating your interest in helping to receive the fund and use it for Gods work as my wish."

$$P\left(\begin{array}{l} \text{bank} = 1 \\ \text{fund} = 1 \\ \text{money} = 1 \\ \text{wish} = 1 \\ \dots \\ \text{to} = 3 \end{array} \middle| \text{spam} \right) = \frac{n!}{1!\,1!\,1!\,1!\cdots 3!} p_{\text{bank}}^{1} p_{\text{fund}}^{1} \cdots p_{\text{to}}^{3}$$

Note: $P\left(\text{bank} \middle| \begin{array}{c}\text{spam} \\ \text{writer}\end{array}\right) \gg P\left(\text{bank} \middle| \begin{array}{c}\text{writer=} \\ \text{you}\end{array}\right)$

Probabilistic text analysis

Probabilities of *counts* of words = Multinomial distribution

What about probability of those same words in someone else's writing?

- $P\left(\text{word = "probability"} \middle| \begin{matrix}\text{writer =} \\ \text{you}\end{matrix}\right) > P\left(\text{word = "probability"} \middle| \begin{matrix}\text{writer =} \\ \text{non−CS109 student}\end{matrix}\right)$

To determine authorship:

1. Estimate $P(\text{word}|\text{writer})$ from known writings
2. Use Bayes' Theorem to determine $P(\text{writer}|\text{document})$ for a new writing!

Who wrote the Federalist Papers?

Old and New Analysis

Authorship of The Federalist Papers

- 85 essays advocating ratification of the US constitution
- Written under the pseudonym “Publius” (really, Alexander Hamilton, James Madison, John Jay)

Who wrote which essays?

- Analyze probability of words in each essay and compare against word distributions from known writings of three authors

Let’s write a program!

http://web.stanford.edu/class/cs109/demos/federalistpapers.html

Announcements

<u>Midterm exam</u>

When: Monday, February 10, 7:00pm-9:00pm
Where: Cubberley Auditorium

Not permitted: book/computer/calculator
Permitted: <u>**Three**</u> 8.5”x11” double-sided sheets of notes

Covers: Up to (and including) week 4 + Lecture Notes 11
Practice: http://web.stanford.edu/class/cs109/exams/midterm.html
Review session: Saturday, 3-5pm, STLC 111
not recorded; materials will be posted though

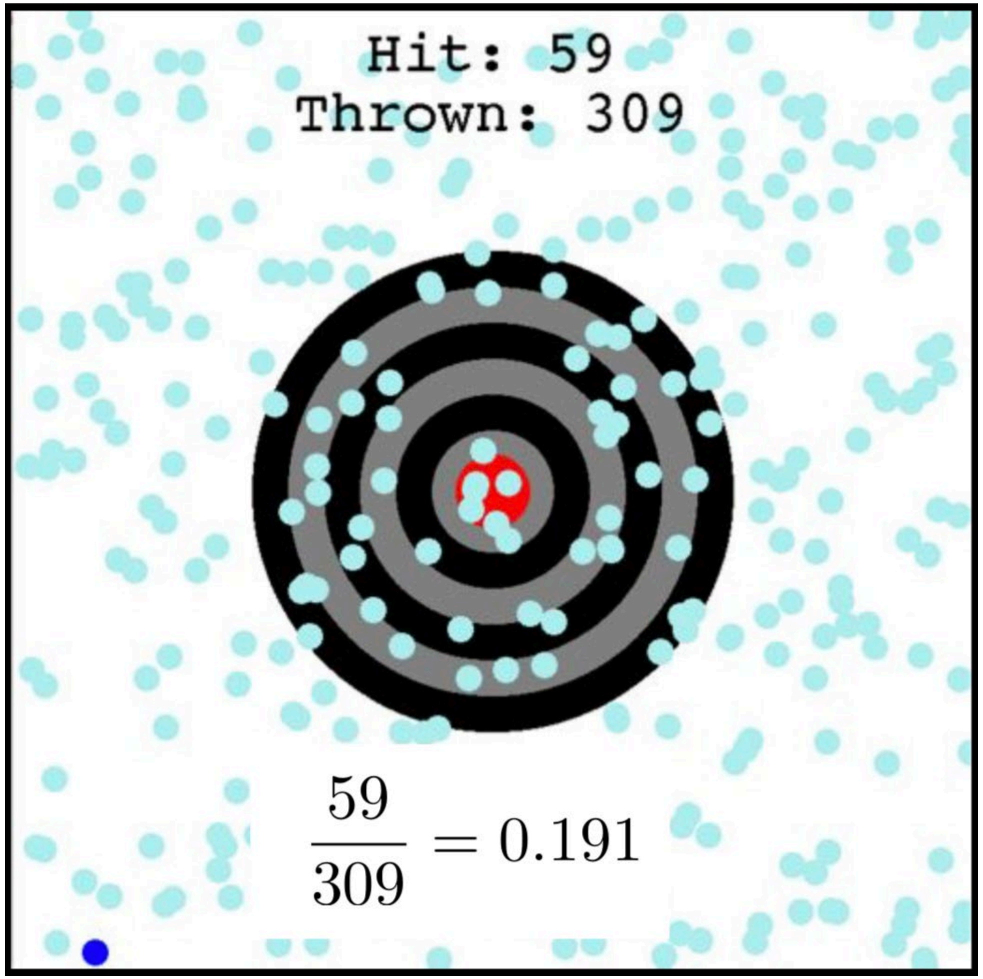
Today's plan

Text analysis

Continuous joint distribution

Joint cumulative distribution functions (CDFs)

Remember target?

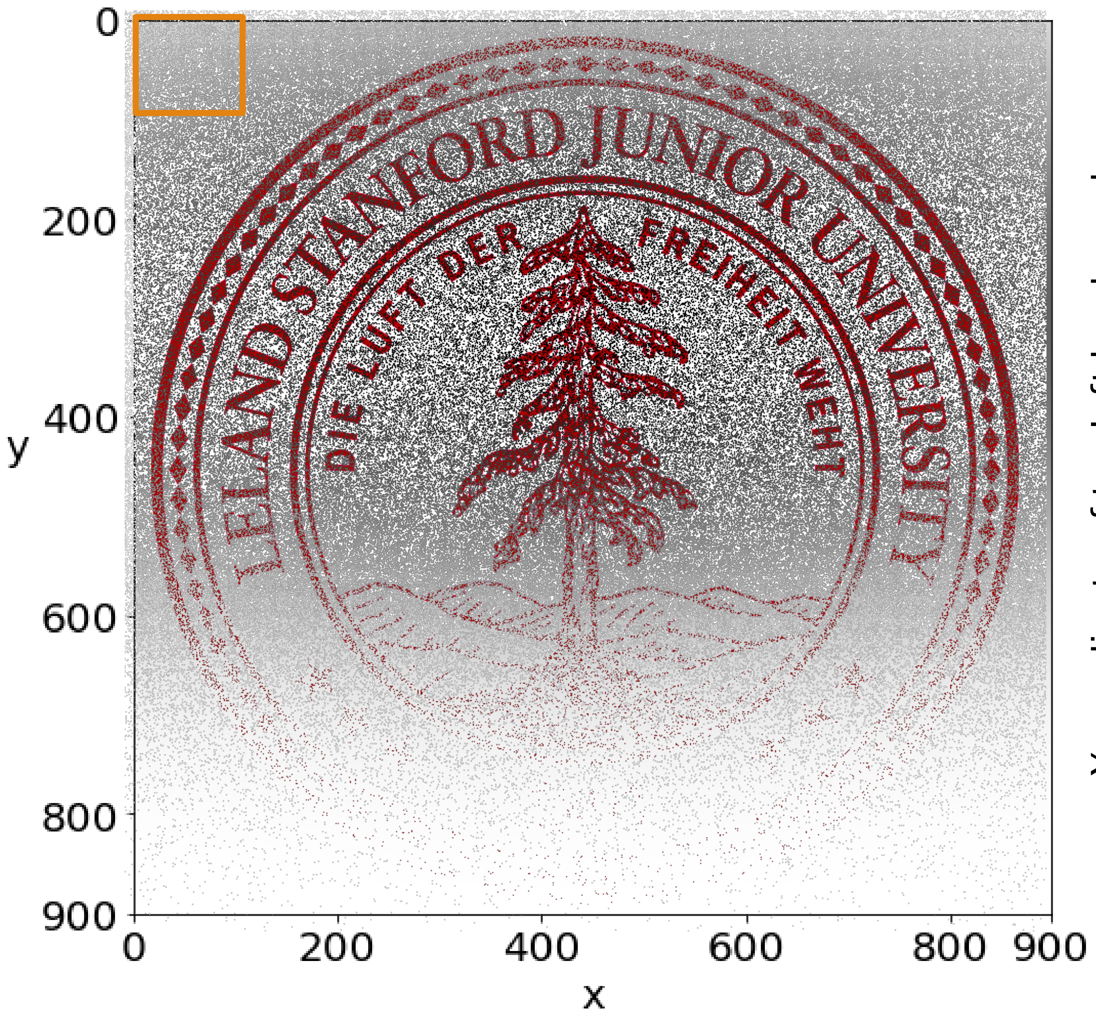


CS109 logo with darts

The CS109 logo was created by throwing 500,000 darts according to a joint distribution.

Quick check: What is the probability that a dart hits at (456.2344132343, 532.1865739012)?

CS109 logo with darts

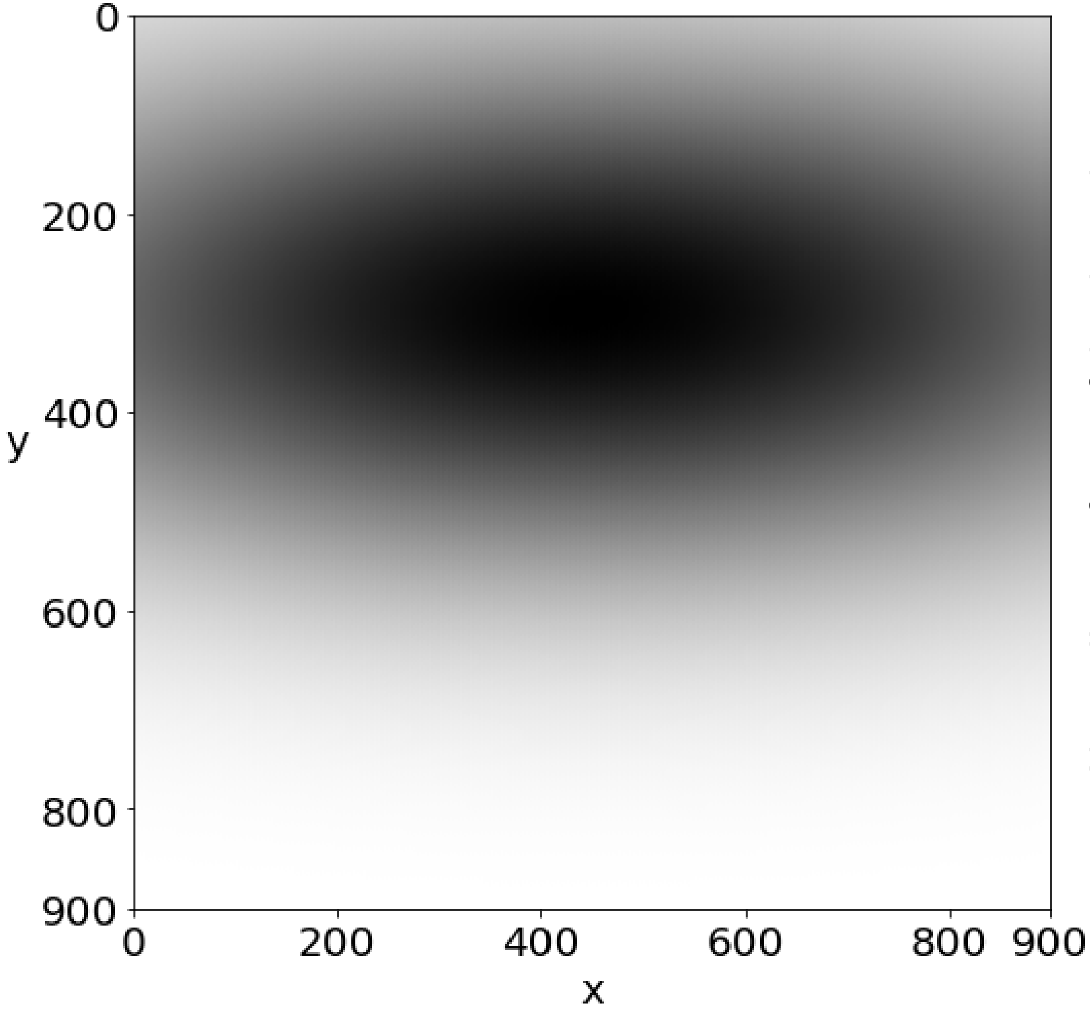


X-coordinate of top-left box bound (columns); Y-coordinate of top-left box bound (rows)

Y \ X	0	100	200	300	400	500	600	700	800
0	3906	4641	5249	5653	5795	5655	5254	4648	3914
100	7140	8483	9596	10333	10593	10338	9605	8496	7154
200	9660	11478	12983	13980	14332	13987	12996	11495	9679
300	9675	11495	13002	14001	14353	14008	13015	11512	9694
400	7172	8522	9639	10380	10641	10385	9648	8534	7186
500	3935	4676	5289	5695	5838	5698	5294	4683	3943
600	1598	1898	2147	2312	2371	2314	2150	1901	1601
700	480	570	645	694	712	695	646	571	481
800	106	126	143	154	158	154	143	127	106

Possible dart counts (in 100x100 boxes)

CS109 logo with darts



X-coordinate of top-left box bound (columns); Y-coordinate of top-left box bound (rows)

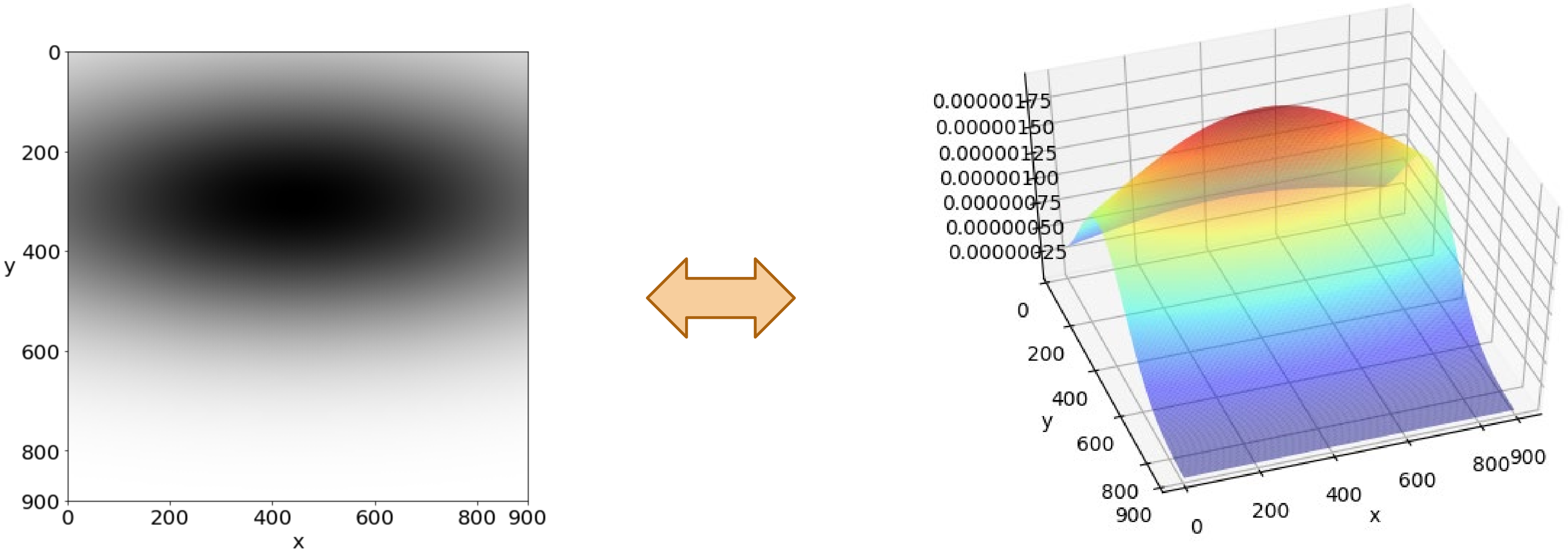
	0	100	200	300	400	500	600	700	800
0	3906	4641	5249	5653	5795	5655	5254	4648	3914
100	7140	8483	9596	10333	10593	10338	9605	8496	7154
200	9660	11478	12983	13980	14332	13987	12996	11495	9679
300	9675	11495	13002	14001	14353	14008	13015	11512	9694
400	7172	8522	9639	10380	10641	10385	9648	8534	7186
500	3935	4676	5289	5695	5838	5698	5294	4683	3943
600	1598	1898	2147	2312	2371	2314	2150	1901	1601
700	480	570	645	694	712	695	646	571	481
800	106	126	143	154	158	154	143	127	106

Possible dart counts (in 100x100 boxes)

Continuous joint probability density functions

If two random variables X and Y are jointly continuous, then there exists a joint probability density function $f_{X,Y}$ defined over $-\infty < x, y < \infty$ such that:

$$P(a_1 < X \le a_2,\ b_1 < Y \le b_2) = \int_{a_1}^{a_2} \int_{b_1}^{b_2} f_{X,Y}(x,y)\,dy\,dx$$

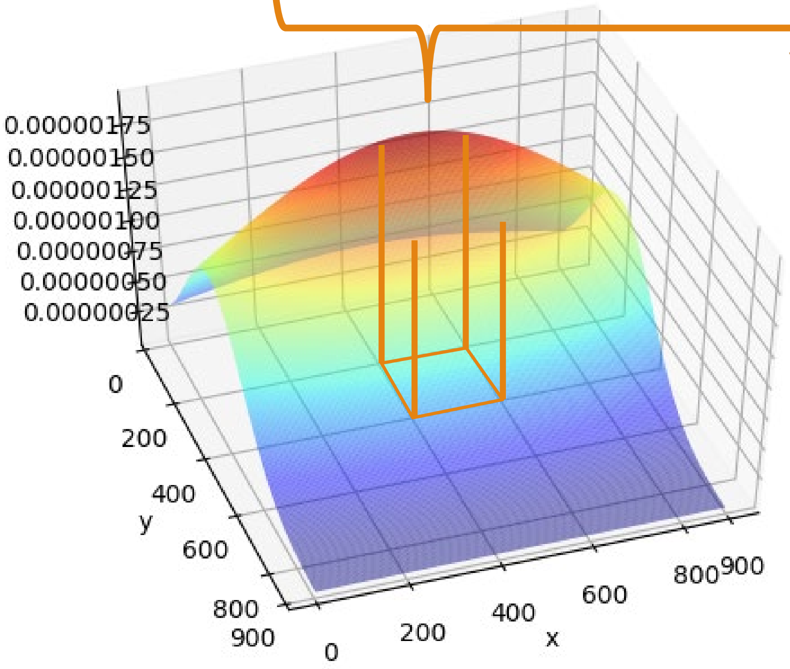


Continuous joint probability density functions

If two random variables X and Y are jointly continuous, then there exists a joint probability density function $f_{X,Y}$ defined over $-\infty < x, y < \infty$ such that:

$$P(a_1 < X \leq a_2,\ b_1 < Y \leq b_2) = \int_{a_1}^{a_2} \int_{b_1}^{b_2} f_{X,Y}(x,y)\,dy\,dx$$

volume



Note:

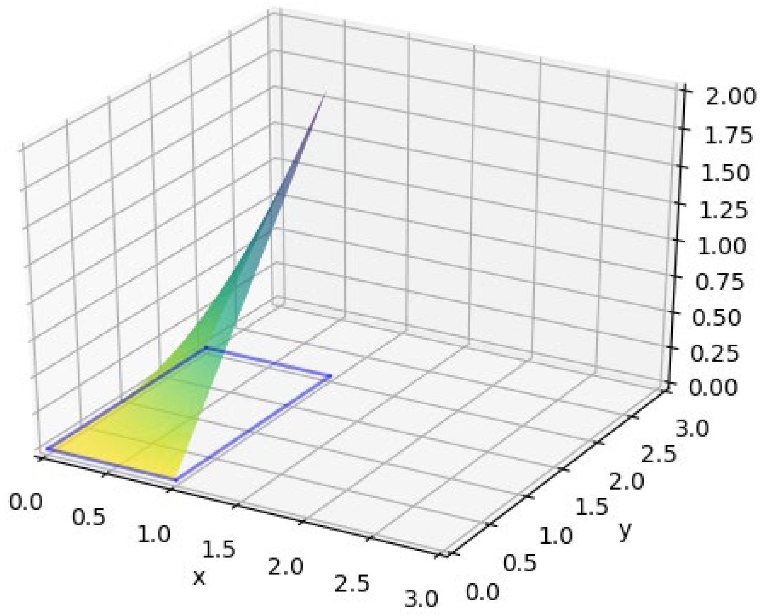
$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy\,dx = 1$$

Double integrals without tears

Let X and Y be two continuous random variables.

- Support: $0 \le X \le 1$, $0 \le Y \le 2$.

Is $g(x, y) = xy$ a valid joint PDF over X and Y?



Want to prove (circle all that apply):

A. $$\int_{y=0}^{2} \int_{x=0}^{1} xy \, dx \, dy = 1$$

B. $$\int_{y=0}^{1} \int_{x=0}^{2} xy \, dx \, dy = 1$$

C. $$\int_{x=0}^{1} \int_{y=0}^{2} xy \, dy \, dx = 1$$

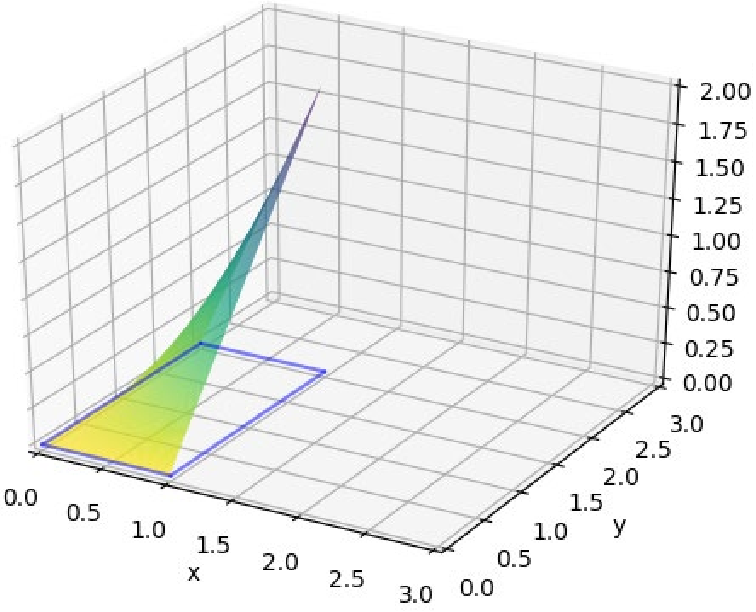
D. $$\int_{x=0}^{1} x \, dx = 1$$

E. $$\int_{y=0}^{2} xy \, dy = 1$$

Double integrals without tears

Let X and Y be two continuous random variables.

- Support: $0 \le X \le 1$, $0 \le Y \le 2$.



Is $g(x, y) = xy$ a valid joint PDF over X and Y?

0. Set up integral:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x, y)dx\,dy = \int_{y=0}^{2}\int_{x=0}^{1} xy\,dx\,dy$$

1. Evaluate inside integral by treating y as a constant:

$$\int_{y=0}^{2}\left(\int_{x=0}^{1} xy\,dx\right)dy = \int_{y=0}^{2} y\left(\int_{x=0}^{1} x\,dx\right)dy = \int_{y=0}^{2} y\left[\frac{x^2}{2}\right]_0^1 dy = \int_{y=0}^{2} y\frac{1}{2}dy$$

2. Evaluate remaining (single) integral:

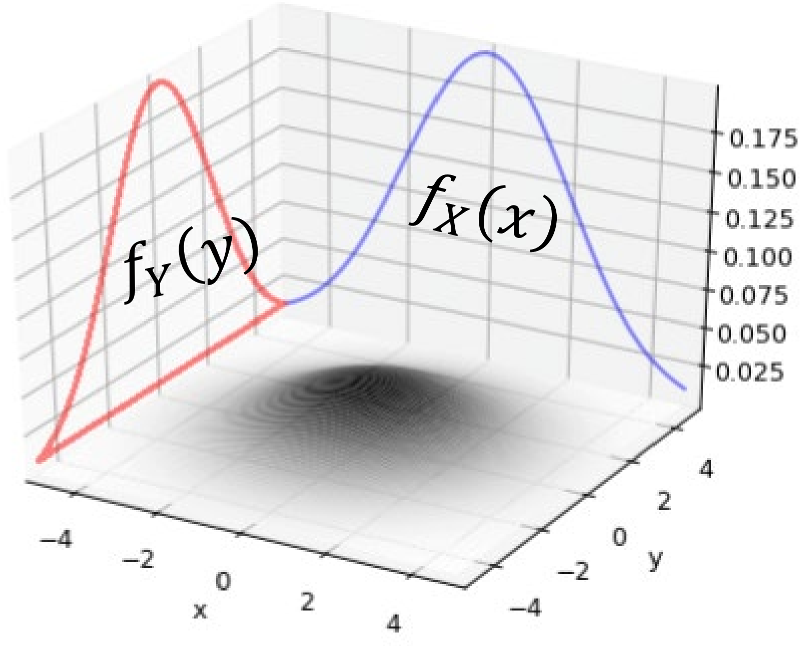
$$\int_{y=0}^{2} y\frac{1}{2}dy = \left[\frac{y^2}{4}\right]_{y=0}^{2} = 1 - 0 = 1$$

$g(x, y)$is a valid joint PDF

Marginal distributions

Suppose X and Y are continuous random variables with joint PDF:

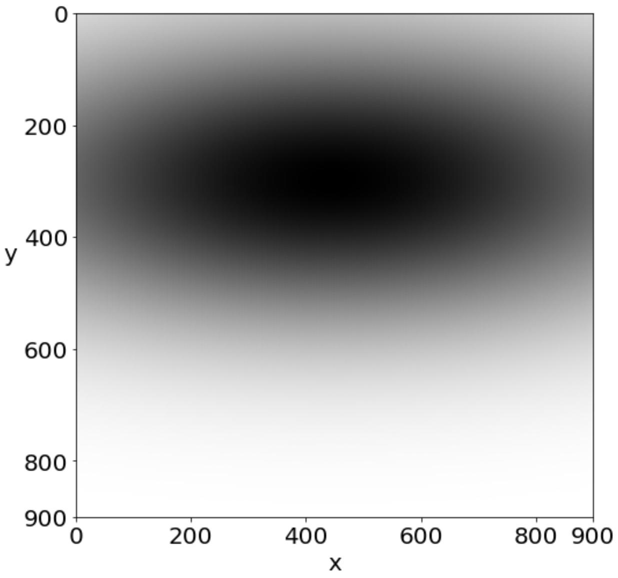
$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy\,dx = 1$$

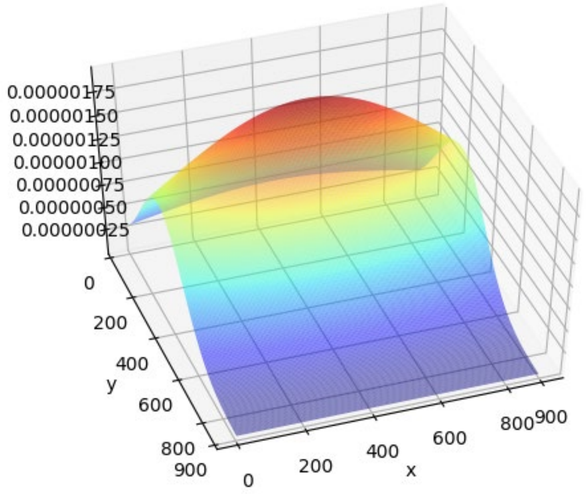


The marginal density functions (marginal PDFs) are therefore:

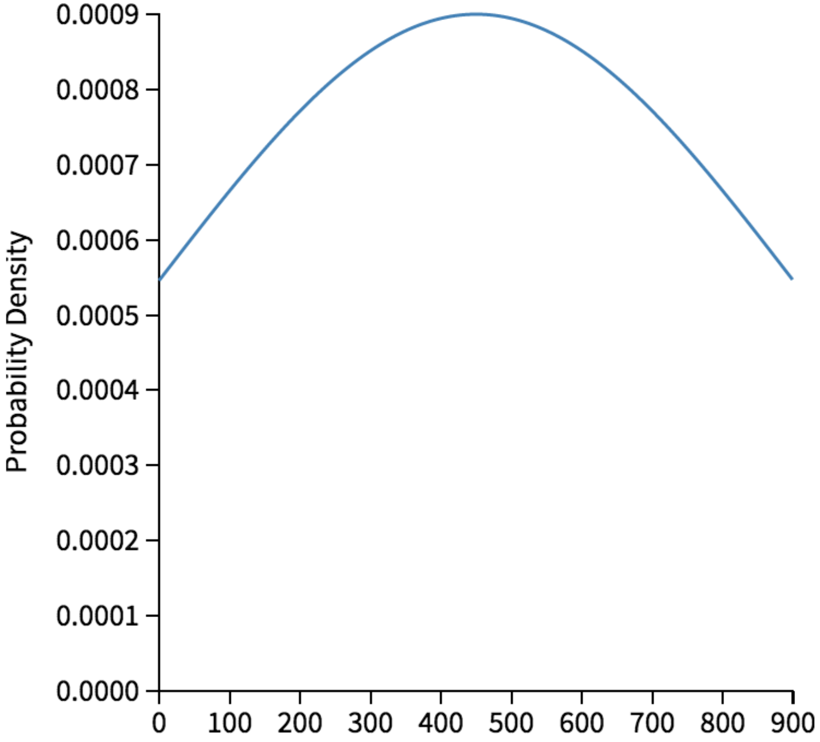
$$f_X(a) = \int_{-\infty}^{\infty} f_{X,Y}(a,y)\,dy \qquad f_Y(b) = \int_{-\infty}^{\infty} f_{X,Y}(x,b)\,dx$$

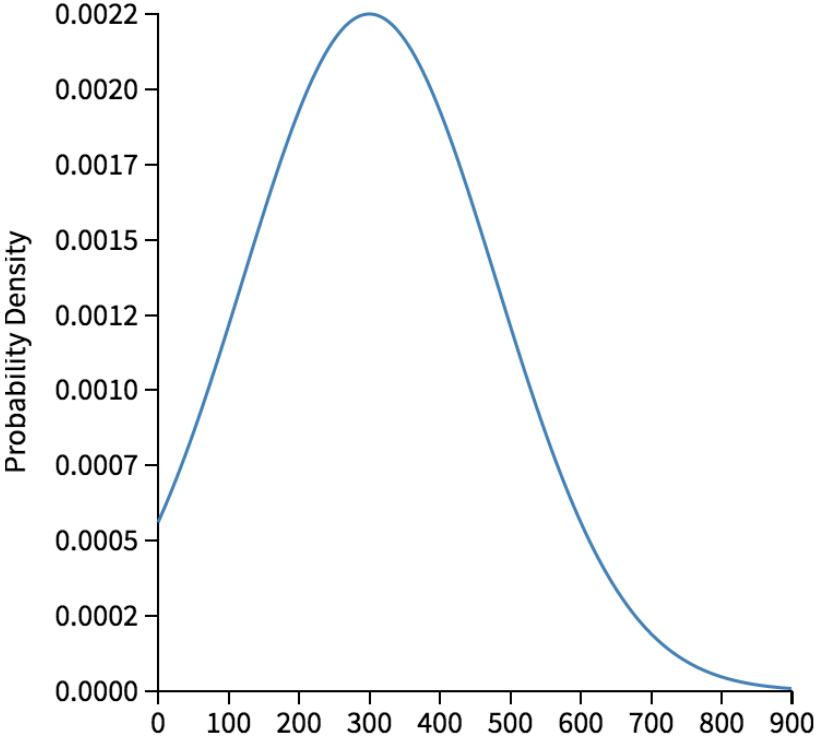
Back to darts!





Match X and Y to their respective marginal PDFs:





Today's plan

Text analysis

Continuous joint distribution

Joint cumulative distribution functions (CDFs)

Joint cumulative distribution function

For two random variables X and Y, there can be a **joint cumulative distribution function** $F_{X,Y}$:

$$F_{X,Y}(a, b) = P(X \leq a, Y \leq b)$$

For discrete X and Y:

$$F_{X,Y}(a, b) = \sum_{x \leq a} \sum_{y \leq b} p_{X,Y}(x, y)$$

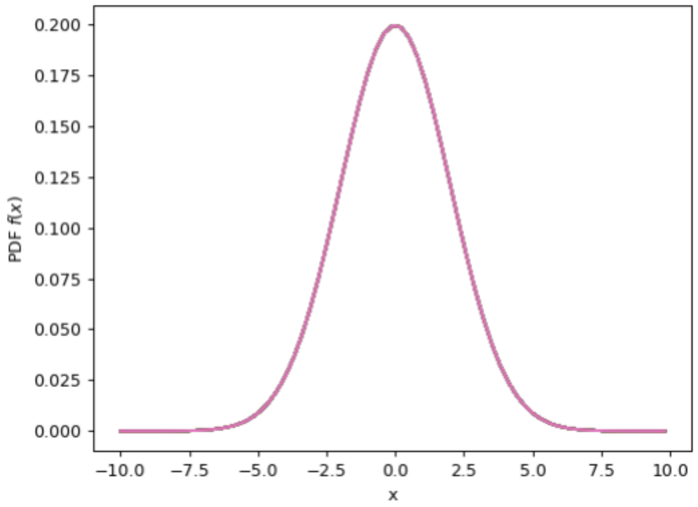
For continuous X and Y:

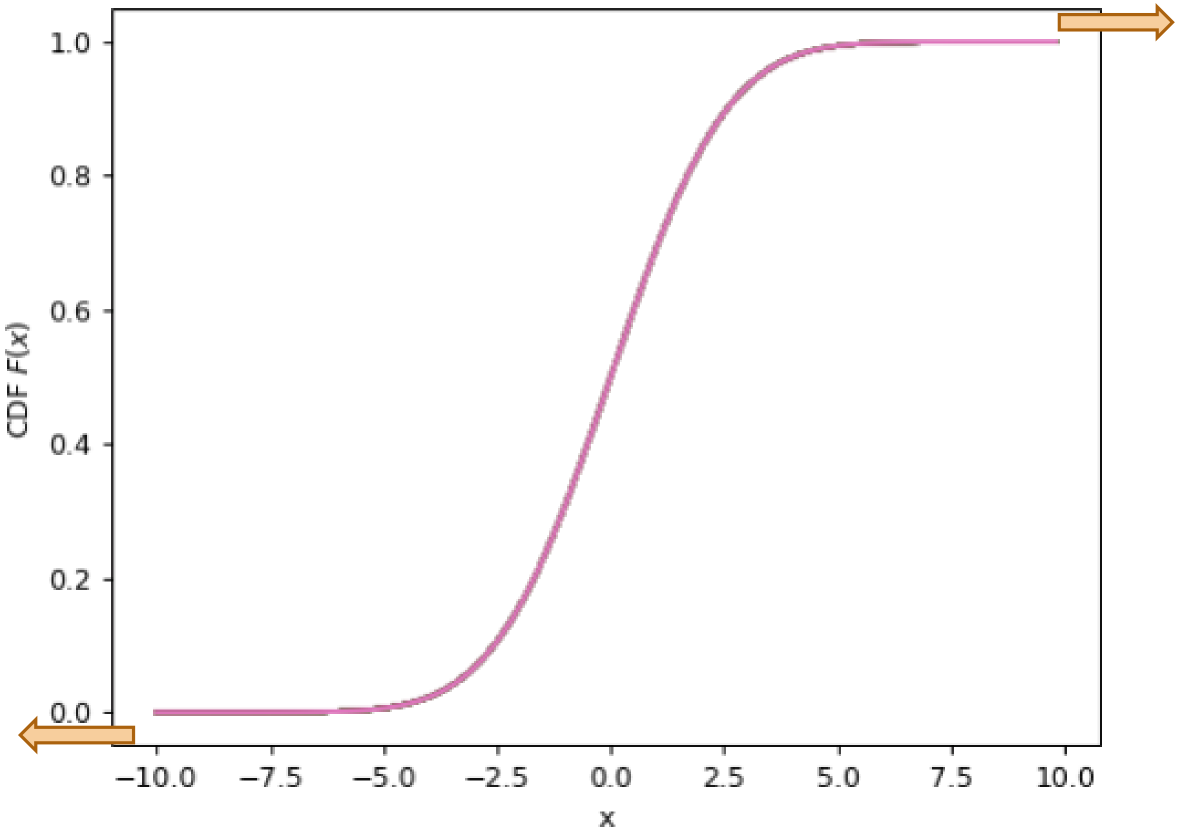
$$F_{X,Y}(a, b) = \int_{-\infty}^{a} \int_{-\infty}^{b} f_{X,Y}(x, y) dy\, dx$$

Single variable CDF, graphically

Review

$$\lim_{x \to +\infty} F_X(x) = 1$$



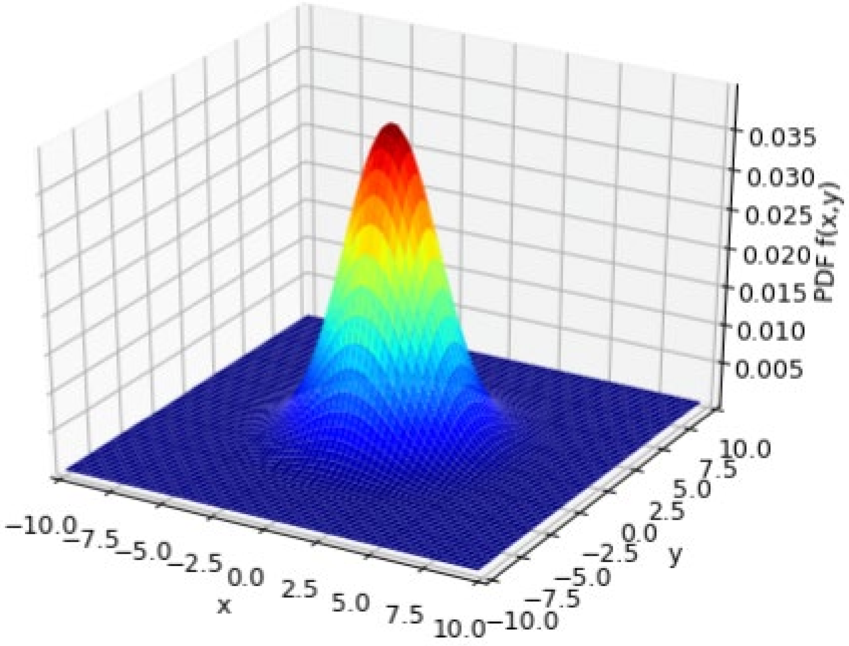


$$\lim_{x \to -\infty} F_X(x) = 0$$

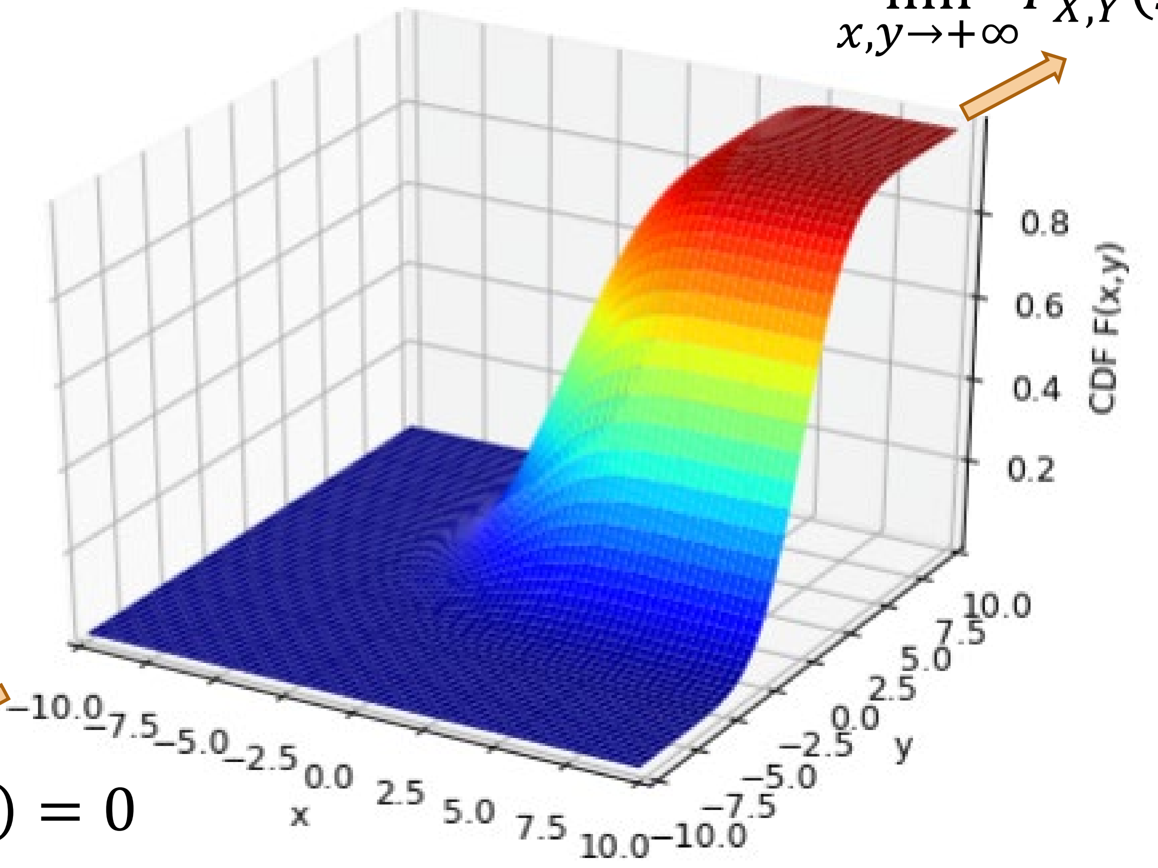
$$f_X(x)$$

$$F_X(x) = P(X \le x)$$

Joint CDF, graphically



$$\lim_{x,y \to +\infty} F_{X,Y}(x,y) = 1$$



$$\lim_{x,y \to -\infty} F_{X,Y}(x,y) = 0$$

$$f_{X,Y}(x,y)$$

$$F_{X,Y}(x,y) = P(X \le x, Y \le y)$$

Probabilities from joint CDFs

Recall for a single RV X with CDF F_X:

CDF: $P(X \leq x) = F_X(x)$

$$P(a < X \leq b) = F_X(b) - F(a)$$

For two RVs X and Y with joint CDF $F_{X,Y}$:

Joint CDF: $P(X \leq x, Y \leq y) = F_{X,Y}(x, y)$

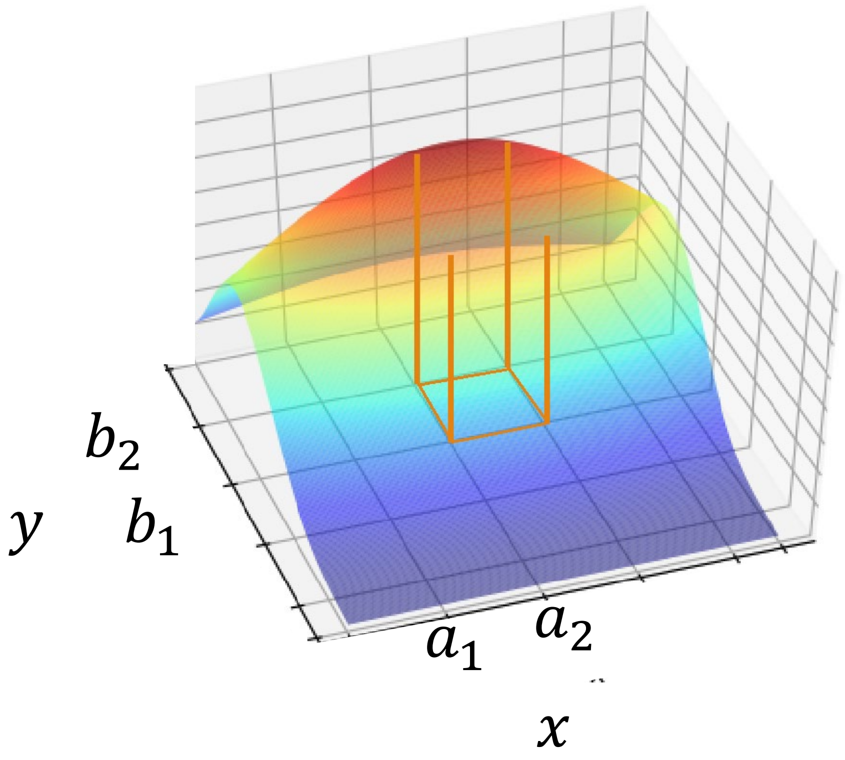
$$P(a_1 < X \leq a_2, b_1 < Y \leq b_2) =$$
$$F_{X,Y}(a_{2,} b_2) - F_{X,Y}(a_{1,} b_2) - F_{X,Y}(a_{2,} b_1) + F_{X,Y}(a_{1,} b_1)$$

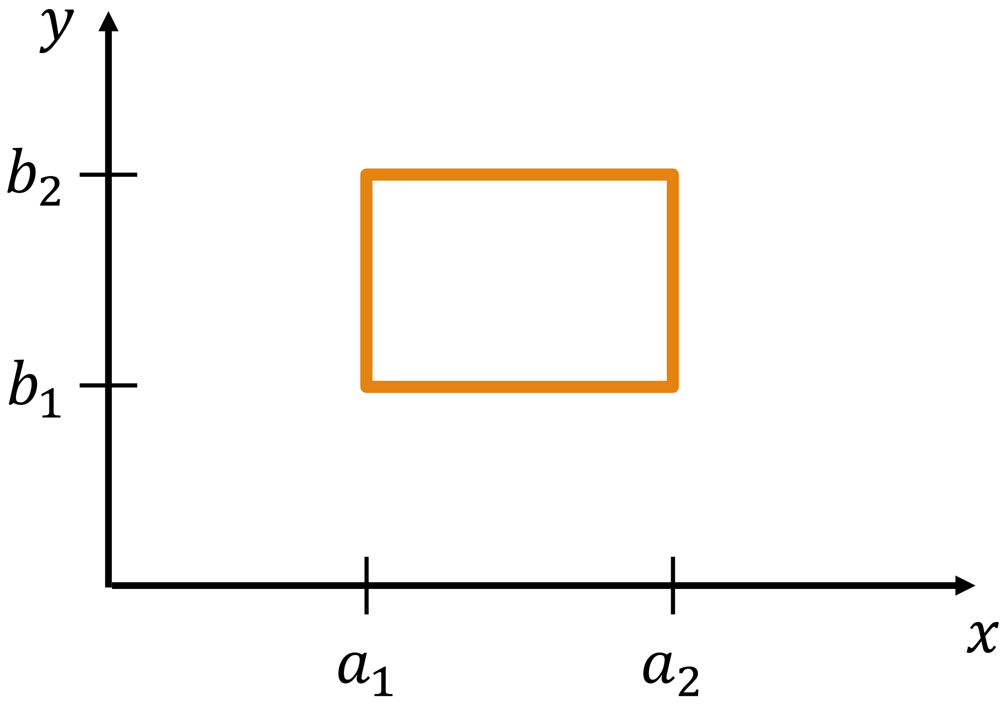
Note strict inequalities; these properties hold for both discrete and continuous RVs.

Probabilities from joint CDFs

Joint CDF: $P(X \leq x, Y \leq y) = F_{X,Y}(x, y)$

$$P(a_1 < X \leq a_2, b_1 < Y \leq b_2) =$$
$$F_{X,Y}(a_2, b_2) - F_{X,Y}(a_1, b_2) - F_{X,Y}(a_2, b_1) + F_{X,Y}(a_1, b_1)$$



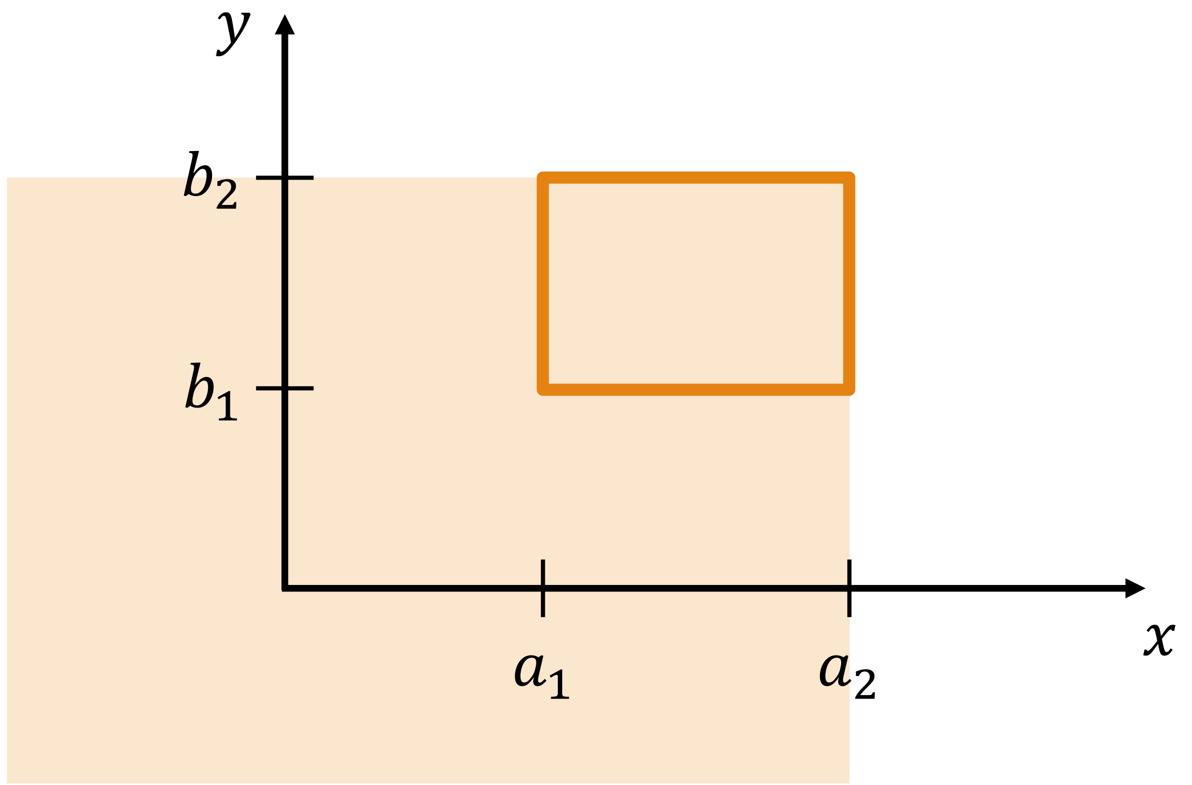


Probabilities from joint CDFs

Joint CDF: $P(X \leq x, Y \leq y) = F_{X,Y}(x, y)$

$$P(a_1 < X \leq a_2, b_1 < Y \leq b_2) =$$

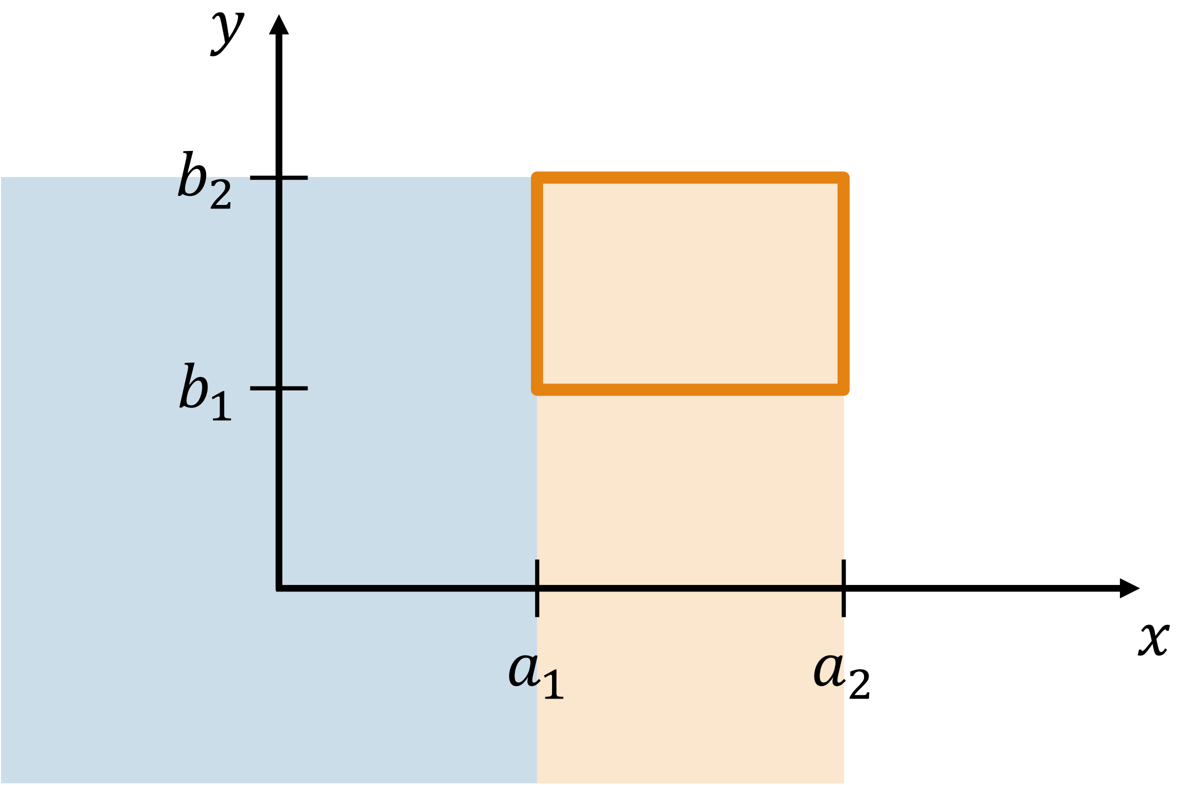
$$F_{X,Y}(a_2, b_2) - F_{X,Y}(a_1, b_2) - F_{X,Y}(a_2, b_1) + F_{X,Y}(a_1, b_1)$$



Probabilities from joint CDFs

Joint CDF: $P(X \leq x, Y \leq y) = F_{X,Y}(x, y)$

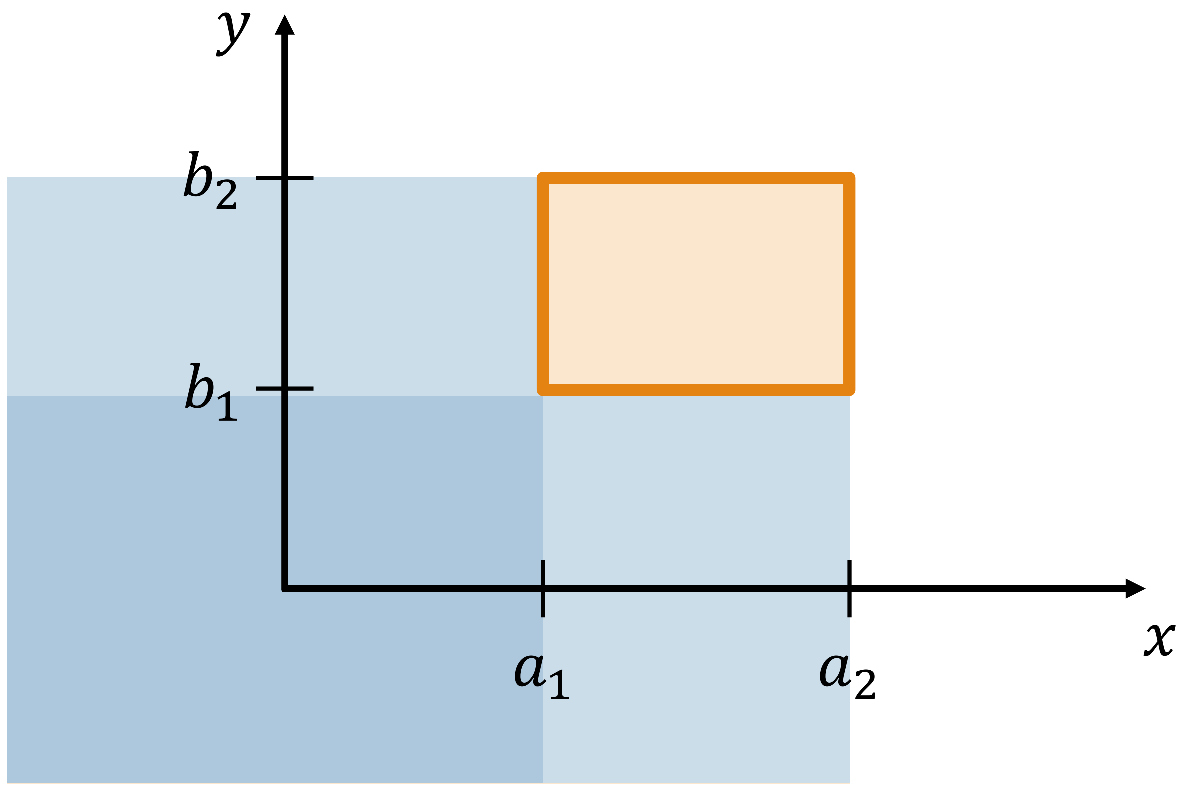
$$P(a_1 < X \leq a_2, b_1 < Y \leq b_2) =$$
$$F_{X,Y}(a_2, b_2) - F_{X,Y}(a_1, b_2) - F_{X,Y}(a_2, b_1) + F_{X,Y}(a_1, b_1)$$



Probabilities from joint CDFs

Joint CDF: $P(X \leq x, Y \leq y) = F_{X,Y}(x, y)$

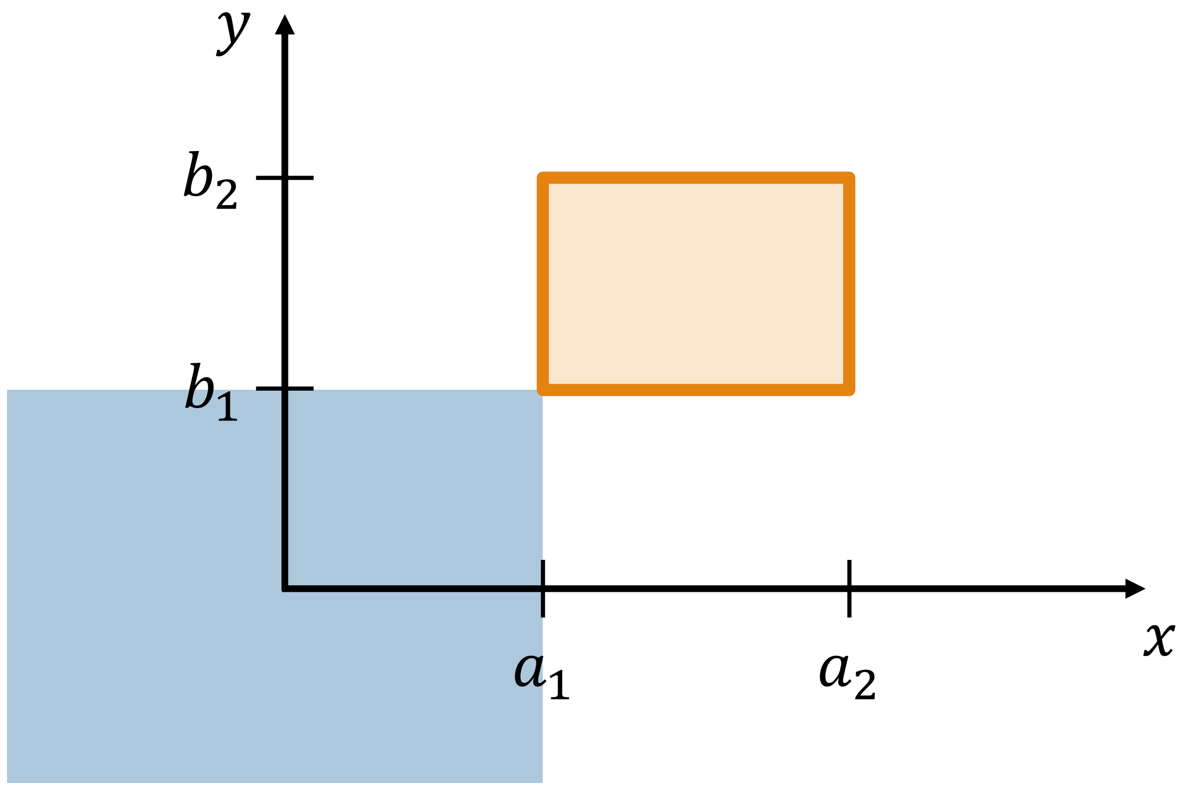
$$P(a_1 < X \leq a_2, b_1 < Y \leq b_2) =$$
$$F_{X,Y}(a_2, b_2) - F_{X,Y}(a_1, b_2) - F_{X,Y}(a_2, b_1) + F_{X,Y}(a_1, b_1)$$



Probabilities from joint CDFs

Joint CDF: $P(X \leq x, Y \leq y) = F_{X,Y}(x, y)$

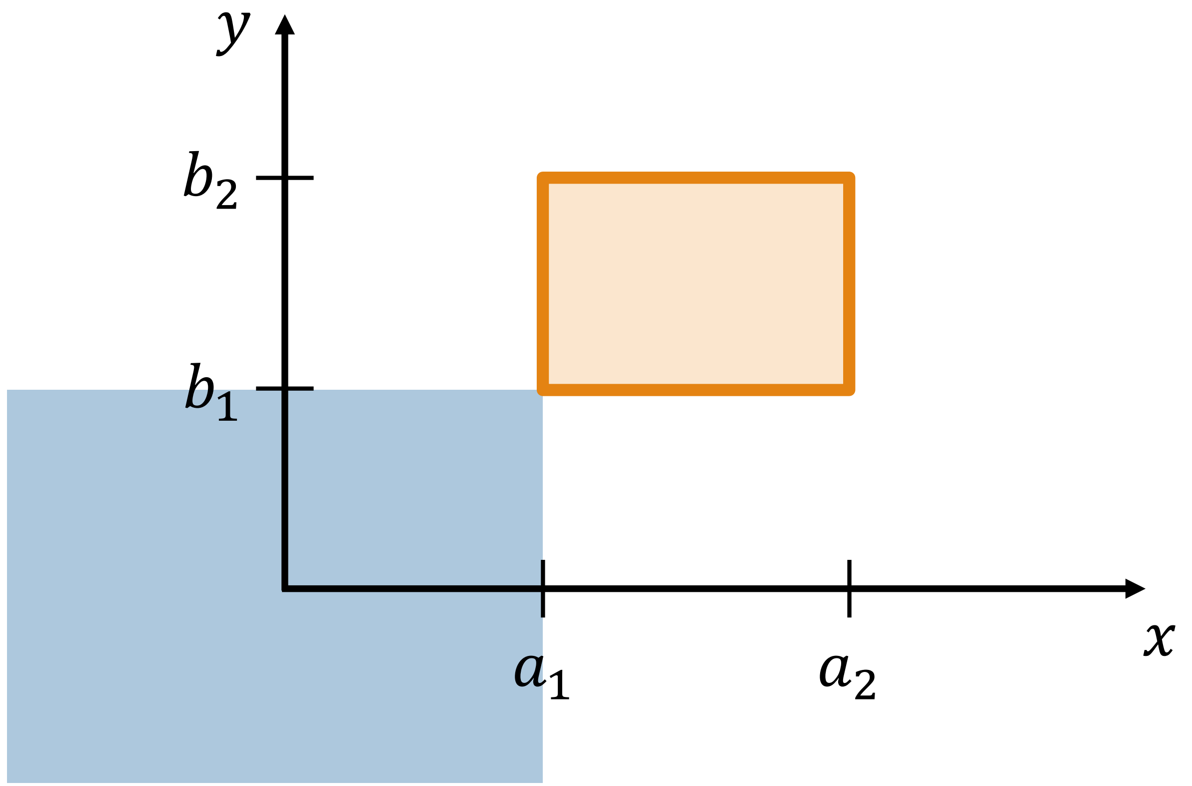
$$P(a_1 < X \leq a_2, b_1 < Y \leq b_2) =$$
$$F_{X,Y}(a_2, b_2) - F_{X,Y}(a_1, b_2) - F_{X,Y}(a_2, b_1) + F_{X,Y}(a_1, b_1)$$



Probabilities from joint CDFs

Joint CDF: $P(X \leq x, Y \leq y) = F_{X,Y}(x, y)$

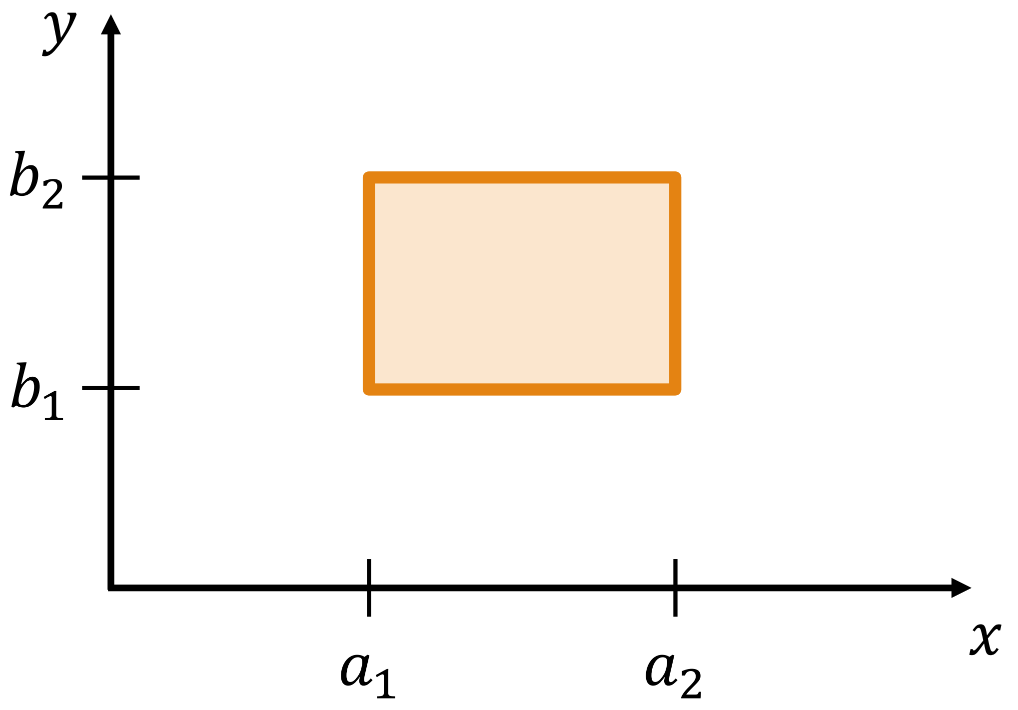
$$P(a_1 < X \leq a_2, b_1 < Y \leq b_2) =$$
$$F_{X,Y}(a_2, b_2) - F_{X,Y}(a_1, b_2) - F_{X,Y}(a_2, b_1) + F_{X,Y}(a_1, b_1)$$



Probabilities from joint CDFs

Joint CDF: $P(X \leq x, Y \leq y) = F_{X,Y}(x, y)$

$$P(a_1 < X \leq a_2, b_1 < Y \leq b_2) =$$
$$F_{X,Y}(a_2, b_2) - F_{X,Y}(a_1, b_2) - F_{X,Y}(a_2, b_1) + F_{X,Y}(a_1, b_1)$$



Probability with Instagram!

$$P(a_1 < X \leq a_2, b_1 < Y \leq b_2) = F_{X,Y}(a_2, b_2) - F_{X,Y}(a_1, b_2) - F_{X,Y}(a_2, b_1) + F_{X,Y}(a_1, b_1)$$

In image processing, a Gaussian blur is the result of blurring an image by a Gaussian function. It is a widely used effect in graphics software, typically to reduce image noise.

Gaussian blur

$$P(a_1 < X \le a_2, b_1 < Y \le b_2) =$$
$$F_{X,Y}(a_2, b_2) - F_{X,Y}(a_1, b_2) - F_{X,Y}(a_2, b_1) + F_{X,Y}(a_1, b_1)$$

In a Gaussian blur, for every pixel:

- Weight each pixel by the probability that X and Y are both within the pixel bounds
- The weighting function is a Gaussian joint PDF with a standard deviation parameter σ.

Gaussian blurring with $\sigma = 3$

Joint PDF:

$$f_{X,Y}(x, y) = \frac{1}{2\pi \cdot 3^2} e^{-(x^2+y^2)/2\cdot 3^2}$$

Joint CDF:

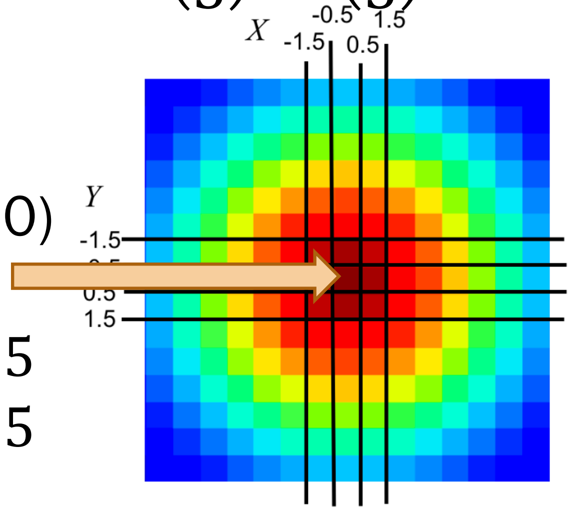
$$F_{X,Y}(x, y) = \Phi\left(\frac{x}{3}\right) \Phi\left(\frac{y}{3}\right)$$

Weight matrix:

Center pixel: (0, 0)

Pixel bounds:

$$-0.5 < x \le 0.5$$
$$-0.5 < y \le 0.5$$



Gaussian blur

$$P(a_1 < X \le a_2, b_1 < Y \le b_2) = F_{X,Y}(a_2, b_2) - F_{X,Y}(a_1, b_2) - F_{X,Y}(a_2, b_1) + F_{X,Y}(a_1, b_1)$$

In a Gaussian blur:

- Weight each pixel by the probability that X and Y are both within the pixel bounds

What is the weight of the center pixel?

$$P(-0.5 < X \le 0.5, -0.5 < Y \le 0.5)$$
$$= F_{X,Y}(0.5, 0.5) - F_{X,Y}(-0.5, 0.5) - F_{X,Y}(0.5, -0.5) + F_{X,Y}(-0.5, -0.5)$$
$$= \Phi\left(\frac{0.5}{3}\right)\Phi\left(\frac{0.5}{3}\right) - 2 \cdot \Phi\left(\frac{-0.5}{3}\right)\Phi\left(\frac{0.5}{3}\right) - \Phi\left(\frac{-0.5}{3}\right)\Phi\left(\frac{-0.5}{3}\right)$$
$$= 0.5662^2 - 2 \cdot 0.5662 \cdot 0.4338 + 0.4338^2$$
$$= 0.206$$

Gaussian blurring with $\sigma = 3$

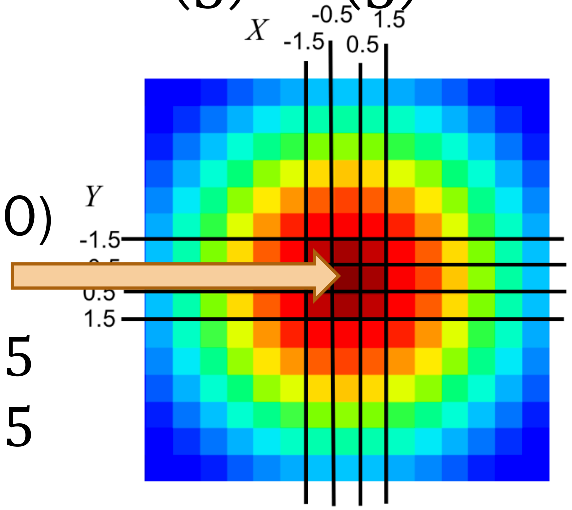
Joint PDF:

$$f_{X,Y}(x, y) = \frac{1}{2\pi \cdot 3^2} e^{-(x^2+y^2)/2 \cdot 3^2}$$

Joint CDF:

$$F_{X,Y}(x, y) = \Phi\left(\frac{x}{3}\right)\Phi\left(\frac{y}{3}\right)$$

Weight matrix:



Center pixel: (0, 0)

Pixel bounds:

$-0.5 < x \le 0.5$

$-0.5 < y \le 0.5$