Distributions

Pre-lecture 4 video

Relevant reading for this lecture

Topics: Continuous pdfs, Expectation, Variance, Entropy/Information, Gaussian pdf (including multivariate), Covariance matrices, MLE with Gaussian pdf, Bayesian estimation with Gaussian pdf, MLE and MAP with Gaussian.

1.2, 1.2.2, 1.2.4, 1.6, 1.6.1, 2.1.1, 2.2, 2.3, 2.3.4, 2.3.6

One of the most important operations involving probabilities is that of finding weighted averages of functions. The average value of some function $f(x)$ under a probability distribution $p(x)$ is called the *expectation* of $f(x)$ and will be denoted by $\mathbb{E}[f]$. For a discrete distribution, it is given by

$$\mathbb{E}[f] = \sum_x p(x) f(x) \tag{1.33}$$

so that the average is weighted by the relative probabilities of the different values of x. In the case of continuous variables, expectations are expressed in terms of an integration with respect to the corresponding probability density

$$\mathbb{E}[f] = \int p(x) f(x) \, \mathrm{d}x. \tag{1.34}$$

Variance

$$Var(X) = E(X - E(X))^2 = \sum P(X)(X - E(X))^2$$

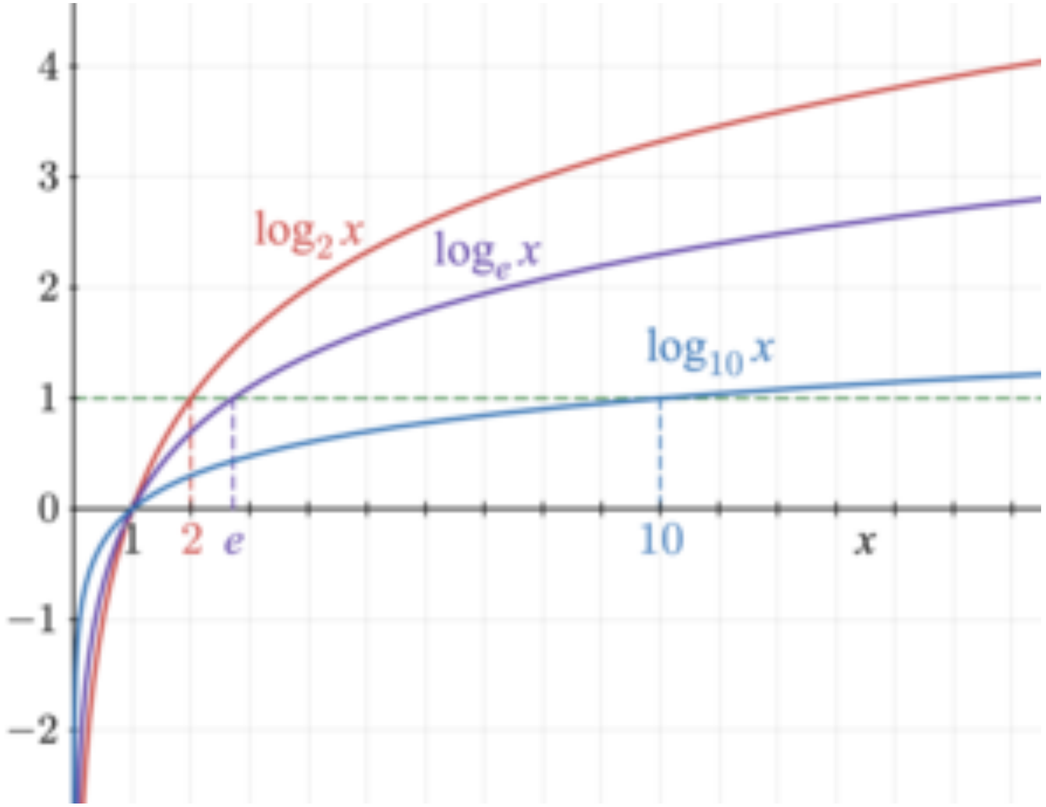
$$Var(X) = \int (x - E(x))^2 p(x) dx$$

Let's show that

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

$$\begin{aligned} Var(X) &= E(X - E(X))^2 \\ &= E((X - E(X))(X - E(X))) \\ &= E(X^2) - E(X)E(X) - E(X)E(X) + E(X)E(X) \\ &= E(X^2) - (E(X))^2 \end{aligned}$$

$$Cov(\vec{x}) = E((\vec{x} - E(\vec{x}))(\vec{x} - E(\vec{x}))^T)$$



$$\log_2(x) = \frac{\ln(x)}{\ln(2)}$$

$$\log(ab) = \log(a) + \log(b)$$

$$\log(a/b) = \log(a) - \log(b)$$

$$\log(a^b) = b\log(a)$$

$$\exp(\ln(a)) = a$$

Note: base of logarithm is matter of convention, in information theory base 2 is typically used: information measured in bits in contrast to nats for base e (natural logarithm).

How to Quantify Information?

consider discrete random variable X taking values from the set of "outcomes" $\{x_k, k=1,\ldots,N\}$ with probability P_k: $0 \le P_k \le 1$; $\Sigma P_k = 1$

Question: what would be a good measure for the *information* gained by observing outcome $X = x_k$?

Idea: Improbable outcomes should somehow provide more information

Let's try: $$I(x_k) = \log(1/P_k) = -\log(P_k)$$ "Shannon information content"

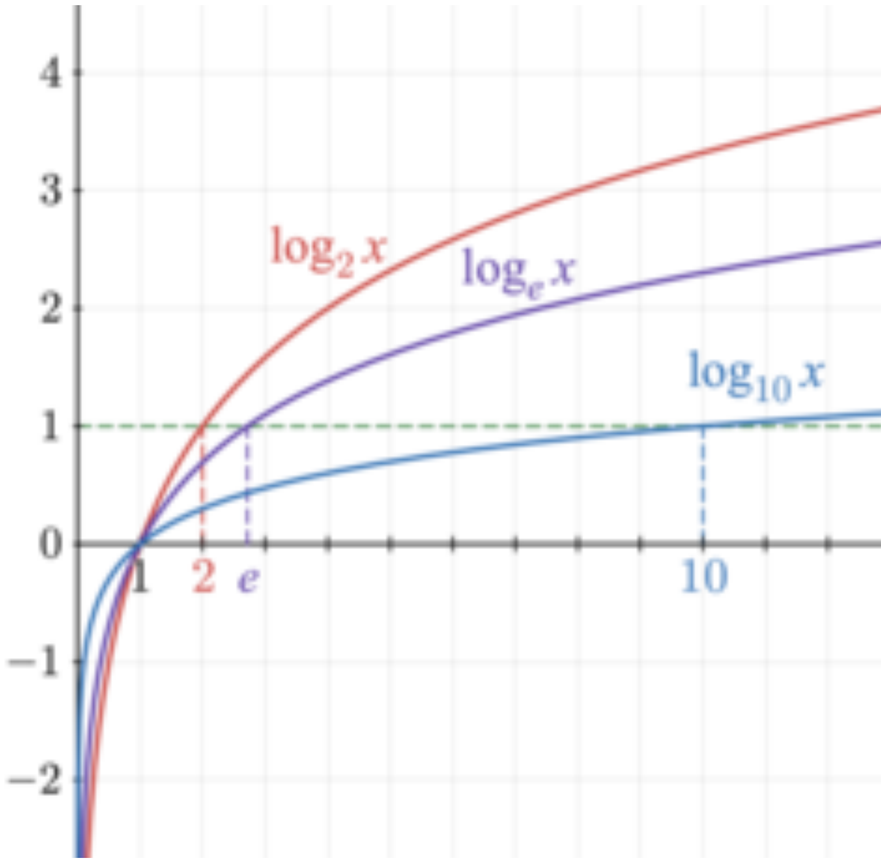
Properties:

$I(x_k) \geq 0$ since $0 \leq P_k \leq 1$

$I(x_k) = 0$ for $P_k = 1$ (certain event gives no information)

$I(x_k) > I(x_i)$ for $P_k < P_i$

logarithm is monotonic, i.e:
if a<b then log(a)<log(b)



Information gained for sequence of independent events

Consider observing the outcomes x_a, x_b, x_c in succession;
the probability for observing this sequence is $P(x_a,x_b,x_c) = P_a P_b P_c$

Let's look at the information gained: $I(x_k) = \log(1/P_k) = -\log(P_k)$

$$I(x_a,x_b,x_c) = -\log(P(x_a,x_b,x_c)) = -\log(P_a P_b P_c)$$

$$= -\log(P_a) - \log(P_b) - \log(P_c)$$

$$= I(x_a) + I(x_b) + I(x_c)$$

Information gained is just the sum of individual information gains

Entropy

Question: What is the average information gained when observing a random variable over and over again?

Answer: ***Entropy!***

$$H(X) \equiv E\left[I(X)\right] = -\sum_k P_k \log(P_k)$$

Notes:

- entropy always bigger than or equal to zero for discrete r.v.s
- entropy is a measure of the uncertainty in a random variable
- can be seen as generalization of variance
- entropy related to minimum average code length for variable
- related concept in physics and physical chemistry: there entropy is a measure of the "disorder" of a system.

Examples

1. Binary random variable: outcomes are {0,1} where outcome '1' occurs with probability P and outcome '0' occurs with probability $Q=1-P$.

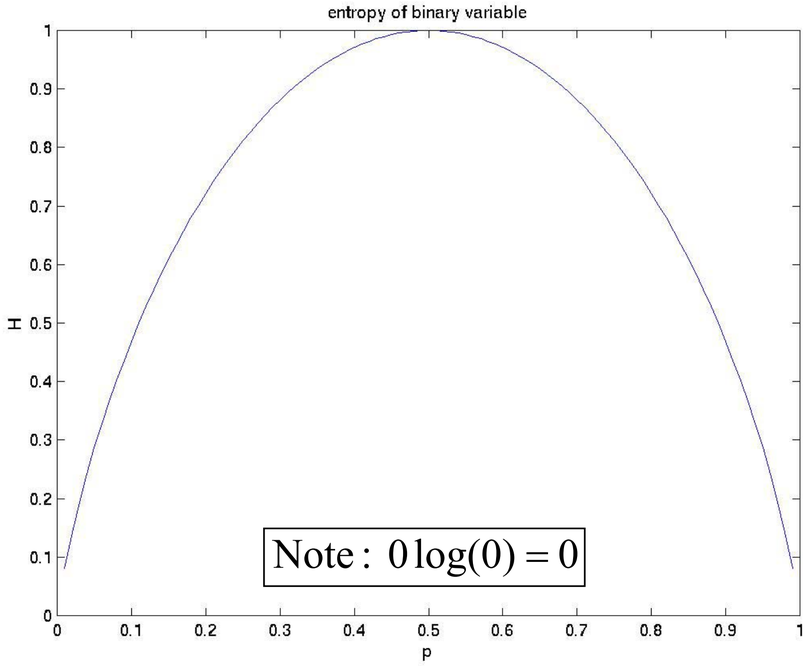
Question: What is the entropy of this random variable?

$$H(X) \equiv -\sum_k P_k \log(P_k)$$

Answer: (just apply definition)

$$H(X) = -P\log(P) - Q\log(Q)$$

Note: entropy zero if one outcome certain, entropy maximized if both outcomes equally likely (1 bit)



$$H(X) \equiv E\left[I(X)\right] = -\sum_k P_k \log(P_k)$$

---

2. Horse Race: eight horses are starting and their respective odds of winning are: 1/2, 1/4, 1/8, 1/16, 1/64, 1/64, 1/64, 1/64

What is the entropy?

H = -(1/2 log(1/2) + 1/4 log(1/4) + 1/8 log(1/8) + 1/16 log(1/16) + 4 * 1/64 log(1/64)
= 2 bits

What if each horse had chance of 1/8 of winning?

H = -8 * 1/8 log(1/8)
= 3 bits (maximum uncertainty)

3. Uniform: for N outcomes entropy maximized if all equally likely:

$$H(X) = -\sum_{k=1}^{N} P_k \log(P_k) = -N \frac{1}{N} \log\left(\frac{1}{N}\right) = \log(N)$$

Differential Entropy

Idea: generalize to continuous random variables described by pdf:

$$H(X) \equiv -\int_{-\infty}^{\infty} p(x)\log(p(x))dx$$

Notes:

• differential entropy can be negative, in contrast to entropy of discrete random variable

• but still: the smaller differential entropy, the "less random" is X

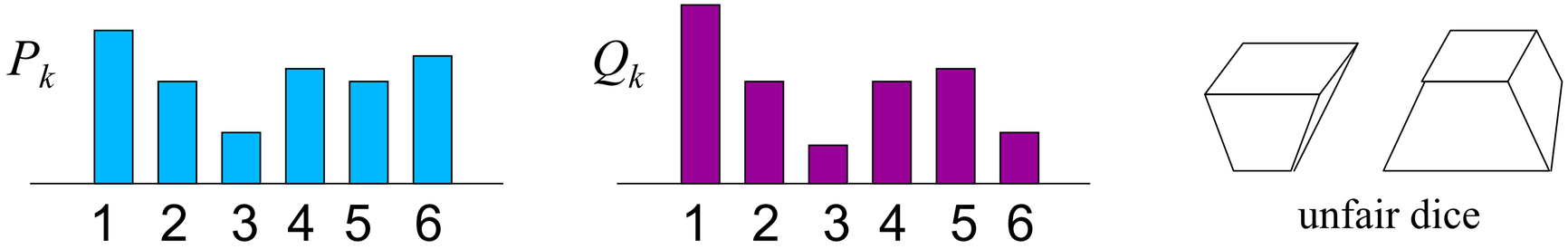
Example: uniform distribution

$$p(x) = \begin{cases} 1/a, \text{ for } 0 \le x \le a \\ 0, \text{otherwise} \end{cases}$$

$$H(X) = -\int_{0}^{a} 1/a \log(1/a)dx = \log(a)$$

Kullback Leibler Divergence

Idea: Consider you want to compare two probability distributions P and Q that are defined over the same set of outcomes.



A "natural" way of defining a "distance" between two distributions is the so-called *Kullback-Leibler divergence (KL-distance), or relative entropy*:

$$D(P \,\|\, Q) \equiv E_P\left[\log \frac{P(X)}{Q(X)}\right] = \sum_k P(x_k) \log \frac{P(x_k)}{Q(x_k)}$$

$$D(P \,||\, Q) \equiv E_P\left[\log \frac{P(X)}{Q(X)}\right] = \sum_k P(x_k) \log \frac{P(x_k)}{Q(x_k)}$$

Properties of KL-divergence:

D(P||Q) ≥ 0 and D(P||Q)=0 if and only if P=Q, i.e., if two distributions are the same, their KL-divergence is zero otherwise it's bigger.
D(P||Q) in general is not equal to D(Q||P) (*i.e.* D(.||.) is not a *metric*)

The KL-divergence is a quantitative measure of how "alike" two probability distributions are.

Generalization to continuous distributions:

$$D(p(x) \,||\, q(x)) \equiv E_p\left[\log \frac{p(x)}{q(x)}\right] = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)}\, dx$$

The same properties as above hold.

Distributions

COGS 118B Winter 2024

Lecture4

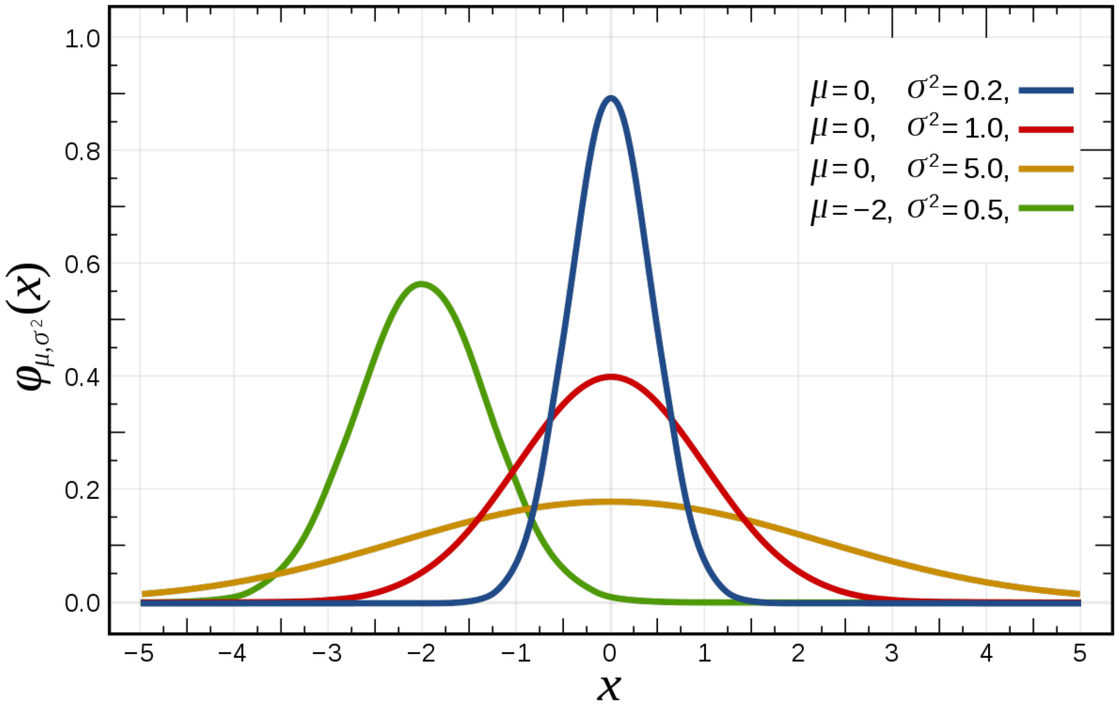
Jason G. Fleischer, PhD
Department of Cognitive Science
University of California San Diego

https://jgfleischer.com
Book a slot in my office hours

Normal distribution

- “Normal” because its very common
- Central limit theorem: sum of a large number of independent RVs is normally distributed
- Analytically tractable
- Exponential family
- Maximum entropy of all distributions with a given mean and variance

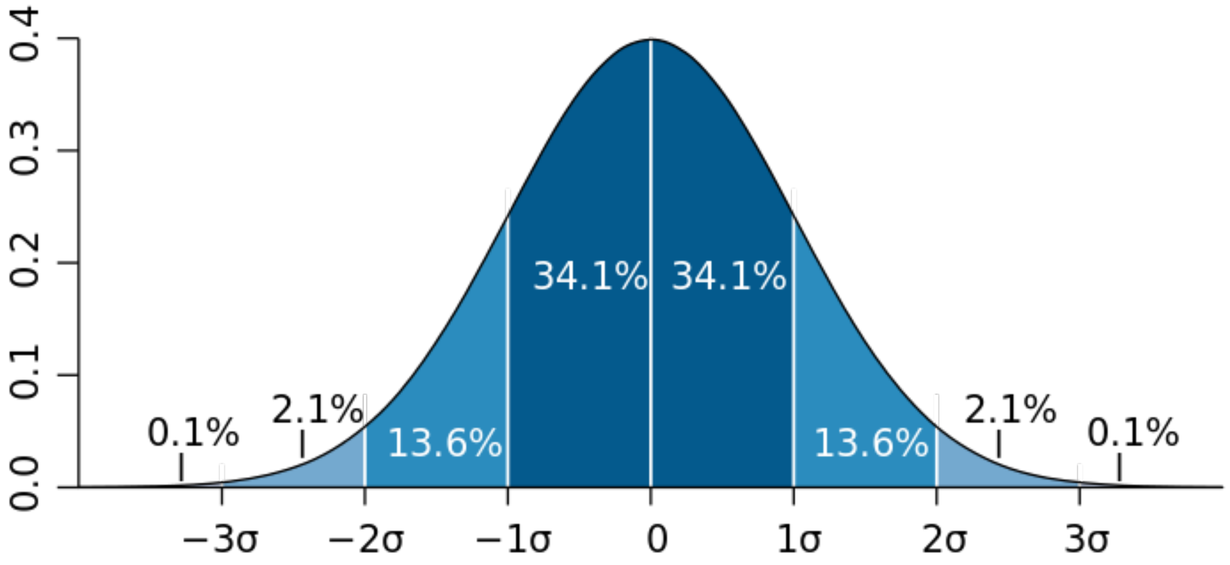
$$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



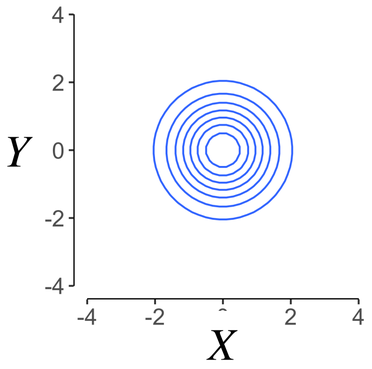
Normal distribution

- Exponential decay with distance squared from μ
- Distance is normalized in units of σ
- Constant in front normalizes PDF to 1

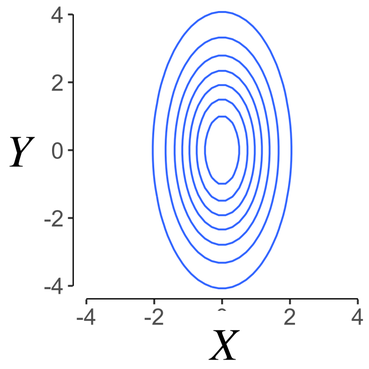
$$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



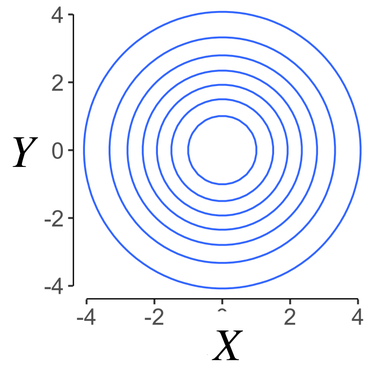
$\rho = 0, \sigma_X = \sigma_Y = 1$



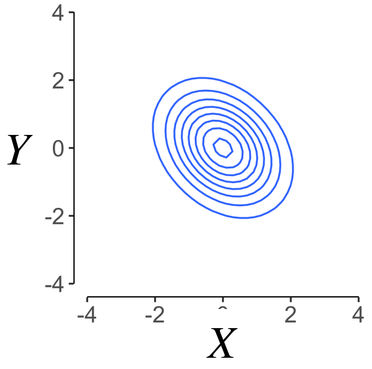
$\rho = 0, \sigma_X = 1, \sigma_Y = 2$



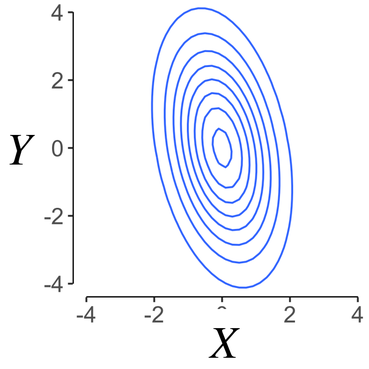
$\rho = 0, \sigma_X = \sigma_Y = 2$



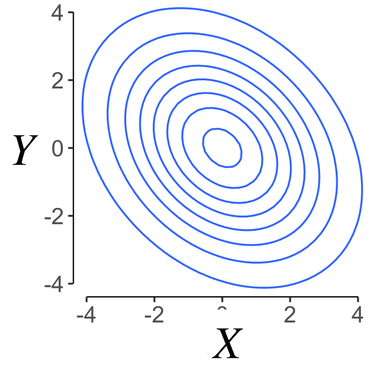
$\rho = -0.3, \sigma_X = \sigma_Y = 1$



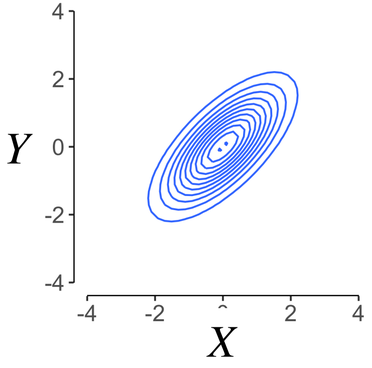
$\rho = -0.3, \sigma_X = 1, \sigma_Y = 2$



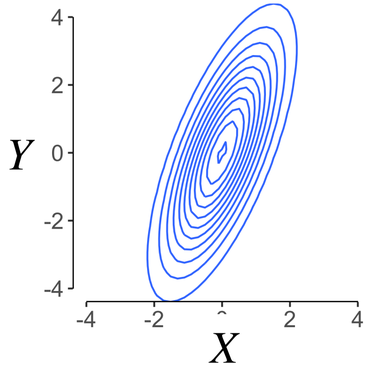
$\rho = -0.3, \sigma_X = \sigma_Y = 2$



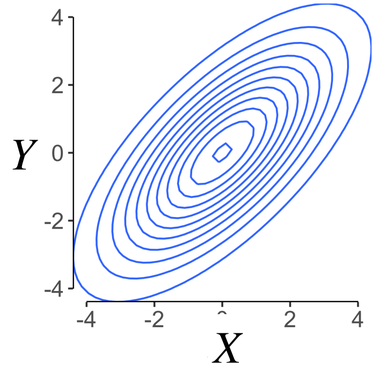
$\rho = 0.7, \sigma_X = \sigma_Y = 1$



$\rho = 0.7, \sigma_X = 1, \sigma_Y = 2$



$\rho = 0.7, \sigma_X = \sigma_Y = 2$



$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}$$

Axis-aligned: $\rho = 0$

Spherical: all σs are the same

Bayes with Normals

Suppose we have a measurement $x \sim N(\theta, \sigma^2)$ where the variance σ^2 is known. That is, the mean θ is our unknown parameter of interest and we are given that the likelihood comes from a normal distribution with variance σ^2. If we choose a normal prior pdf

$$f(\theta) \sim \mathrm{N}(\mu_{\text{prior}}, \sigma^2_{\text{prior}})$$

then the posterior pdf is also normal: $f(\theta|x) \sim \mathrm{N}(\mu_{\text{post}}, \sigma^2_{\text{post}})$ where

$$\frac{\mu_{\text{post}}}{\sigma^2_{\text{post}}} = \frac{\mu_{\text{prior}}}{\sigma^2_{\text{prior}}} + \frac{x}{\sigma^2}, \qquad \frac{1}{\sigma^2_{\text{post}}} = \frac{1}{\sigma^2_{\text{prior}}} + \frac{1}{\sigma^2} \tag{1}$$

The following form of these formulas is easier to read and shows that μ_{post} is a weighted average between μ_{prior} and the data x.

$$a = \frac{1}{\sigma^2_{\text{prior}}} \qquad b = \frac{1}{\sigma^2}, \qquad \mu_{\text{post}} = \frac{a\mu_{\text{prior}} + bx}{a+b}, \qquad \sigma^2_{\text{post}} = \frac{1}{a+b}. \tag{2}$$

hypothesis	data	prior	likelihood	posterior
θ	x	$f(\theta) \sim \mathrm{N}(\mu_{\text{prior}}, \sigma^2_{\text{prior}})$ $= c_1 \exp\left(\frac{-(\theta-\mu_{\text{prior}})^2}{2\sigma^2_{\text{prior}}}\right)$	$\phi(x\|\theta) \sim \mathrm{N}(\theta, \sigma^2)$ $= c_2 \exp\left(\frac{-(x-\theta)^2}{2\sigma^2}\right)$	$f(\theta\|x) \sim \mathrm{N}(\mu_{\text{post}}, \sigma^2_{\text{post}})$ $= c_3 \exp\left(\frac{-(\theta-\mu_{\text{post}})^2}{2\sigma^2_{\text{post}}}\right)$

Bayes with Normals

Example 2. Suppose we have prior $\theta \sim \mathrm{N}(4, 8)$, and likelihood function likelihood $x \sim \mathrm{N}(\theta, 5)$. Suppose also that we have one measurement $x_1 = 3$. Show the posterior distribution is normal.

answer: We will show this by grinding through the algebra which involves completing the square.

$$\text{prior: } f(\theta) = c_1\, \mathrm{e}^{-(\theta-4)^2/16}; \qquad \text{likelihood: } \phi(x_1|\theta) = c_2\, \mathrm{e}^{-(x_1-\theta)^2/10} = c_2\, \mathrm{e}^{-(3-\theta)^2/10}$$

We multiply the prior and likelihood to get the posterior:

$$\begin{aligned} f(\theta|x_1) &= c_3\, \mathrm{e}^{-(\theta-4)^2/16}\, \mathrm{e}^{-(3-\theta)^2/10} \\ &= c_3 \exp\left(-\frac{(\theta-4)^2}{16} - \frac{(3-\theta)^2}{10}\right) \end{aligned}$$

We complete the square in the exponent

$$\begin{aligned} -\frac{(\theta-4)^2}{16} - \frac{(3-\theta)^2}{10} &= -\frac{5(\theta-4)^2 + 8(3-\theta)^2}{80} \\ &= -\frac{13\theta^2 - 88\theta + 152}{80} \\ &= -\frac{\theta^2 - \frac{88}{13}\theta + \frac{152}{13}}{80/13} \\ &= -\frac{(\theta - 44/13)^2 + 152/13 - (44/13)^2}{80/13}. \end{aligned}$$

Therefore the posterior is

$$f(\theta|x_1) = c_3\, \mathrm{e}^{-\frac{(\theta-44/13)^2 + 152/13 - (44/13)^2}{80/13}} = c_4\, \mathrm{e}^{-\frac{(\theta-44/13)^2}{80/13}}.$$

This has the form of the pdf for $\mathrm{N}(44/13, 40/13)$. QED

For practice we check this against the formulas (2).

$$\mu_{\text{prior}} = 4, \quad \sigma^2_{\text{prior}} = 8, \quad \sigma^2 = 5 \;\Rightarrow\; a = \frac{1}{8}, \quad b = \frac{1}{5}.$$

Therefore

$$\begin{aligned} \mu_{\text{post}} &= \frac{a\mu_{\text{prior}} + bx}{a+b} = \frac{44}{13} = 3.38 \\ \sigma^2_{\text{post}} &= \frac{1}{a+b} = \frac{40}{13} = 3.08. \end{aligned}$$