Probability

CMPUT 367: Intermediate Machine Learning

Chapter 2

PMFs and PDFs of Many Variables

We can consider a d-dimensional random variable $\overrightarrow{X} = (X_1, \ldots, X_d)$ with vector-valued outcomes $\overrightarrow{x} = (x_1, \ldots, x_d)$, with each x_i chosen from some $\mathcal{X}_i$. Then,

Discrete case:

$p : \mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_d \to [0,1]$ is a **(joint) probability mass function** if

$$\sum_{x_1 \in \mathcal{X}_1} \sum_{x_2 \in \mathcal{X}_2} \cdots \sum_{x_d \in \mathcal{X}_d} p(x_1, x_2, \ldots, x_d) = 1$$

Continuous case:

$p : \mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_d \to [0,\infty)$ is a **(joint) probability density function** if

$$\int_{\mathcal{X}_1} \int_{\mathcal{X}_2} \cdots \int_{\mathcal{X}_d} p(x_1, x_2, \ldots, x_d)\, dx_1 dx_2 \ldots dx_d = 1$$

Rules of Probability Already Covered the Multidimensional Case

Outcome space is $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_d$

Outcomes are multidimensional variables $\mathbf{x} = [x_1, x_2, \ldots, x_d]$

Discrete case:

$p : \mathcal{X} \rightarrow [0,1]$ is a **(joint) probability mass function** if $\sum_{\mathbf{x} \in \mathcal{X}} p(\mathbf{x}) = 1$

Continuous case:

$p : \mathcal{X} \rightarrow [0,\infty)$ is a **(joint) probability density function** if $\int_{\mathcal{X}} p(\mathbf{x})\, d\mathbf{x} = 1$

But useful to recognize that we have multiple variables

Marginal Distributions

A **marginal distribution** is defined for a subset of $\overrightarrow{X}$ by summing or integrating out the remaining variables. (We will often say that we are "marginalizing over" or "marginalizing out" the remaining variables).

Discrete case: $p(x_i) = \sum_{x_1 \in \mathcal{X}_1} \cdots \sum_{x_{i-1} \in \mathcal{X}_{i-1}} \sum_{x_{i+1} \in \mathcal{X}_{i+1}} \cdots \sum_{x_d \in \mathcal{X}_d} p(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_d)$

Continuous:

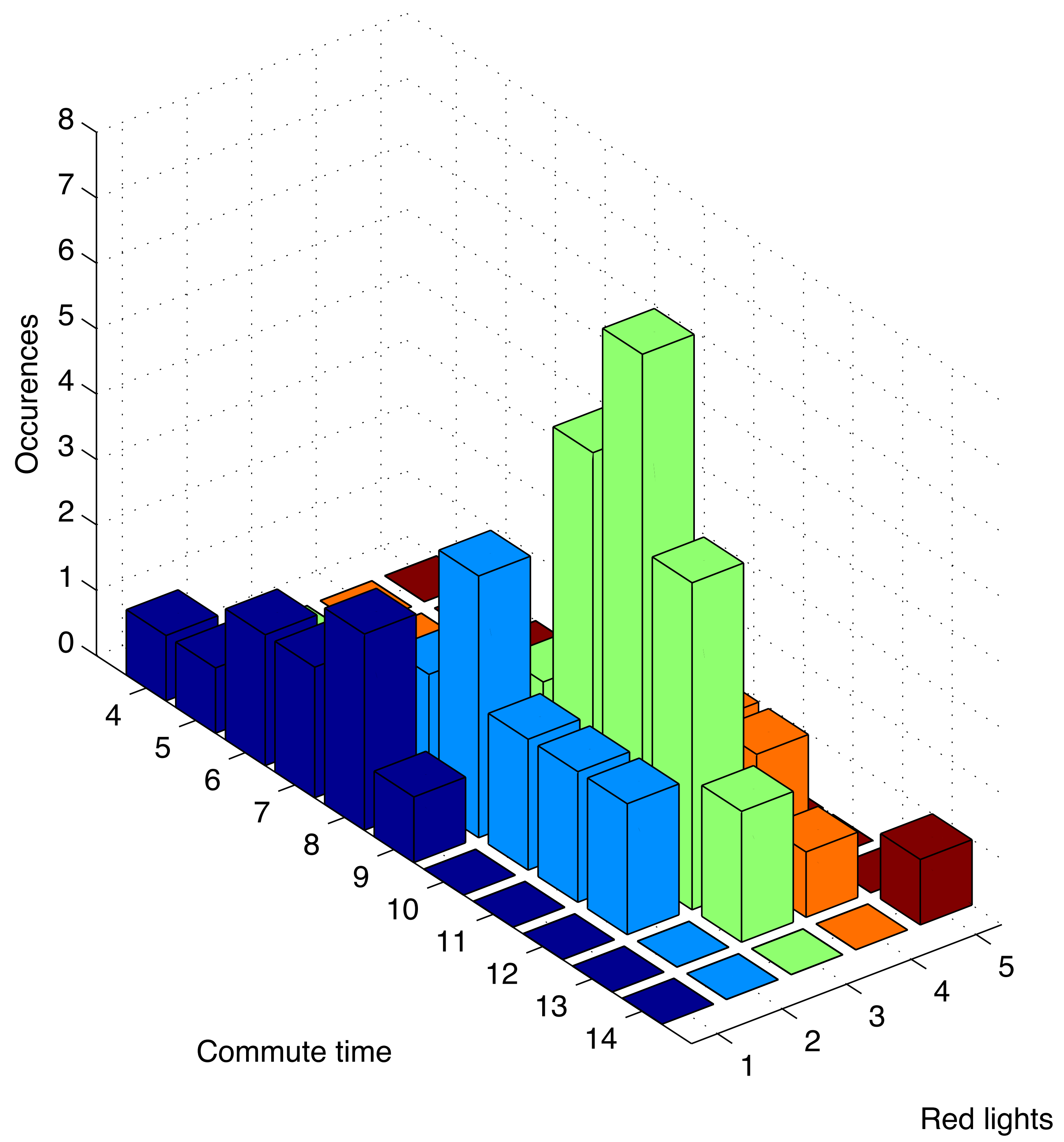
$$p(x_i) = \int_{\mathcal{X}_1} \cdots \int_{\mathcal{X}_{i-1}} \int_{\mathcal{X}_{i+1}} \cdots \int_{\mathcal{X}_d} p(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_d)\, dx_1 \ldots dx_{i-1} dx_{i+1} \ldots dx_d$$

Multidimensional PMF often is simply a multi-dimensional array

Now record both commute time and number red lights

$\Omega = \{4, \ldots, 14\} \times \{1, 2, 3, 4, 5\}$

PMF is normalized 2-d table (histogram) of occurrences



Multivariate PMF: Multinomial Distribution

- Sample space: $\mathscr{X} = \{0,1,\ldots,n\}^d$
- $$p(x_1, x_2, \ldots, x_d) = \begin{cases} \binom{n}{x_1, x_2, \ldots, x_d} \alpha_1^{x_1}\alpha_2^{x_2}\ldots\alpha_d^{x_d} & \text{if } x_1 + x_2 + \cdots + x_d = n \\ 0 & \text{otherwise} \end{cases}$$
- where $\alpha_i \geq 0, \quad \sum_{i=1}^{d} \alpha_i = 1$
- α_i gives probability
- Coefficient says how we can distribute n balls into d boxes such that the first box contains k1 balls, the second box k2 balls, etc.

Example: Multiple Rolls

- n tosses of a 6-sided dice
- d = 6, with $x_i =$ number of times we saw a i
    - $(x_1, x_2, \ldots, x_6) = (3,2,2,1,4,1)$ means we saw 3 ones, 2 twos, 2 threes, 1 four, 4 fives and 1 six. This means n = 13
- All the $\alpha_i = 1/6$
- $p(x_1, x_2, \ldots, x_6) =$ probability of seeing x_1 ones, x_2 twos, etc. (regardless of the order)

More usefully for us: Multi-class classification

- Want to categorize an item into one of d classes
- Only one "roll": n = 1, $x_i = 1$ if the item is categorized as class i
- Sample space: $\mathcal{X} = \{0,1\}^d$ (e.g., outcome is $(0,1,0,0)$ for d = 4)
- $p(x_1, x_2, \ldots, x_d) = \begin{cases} \alpha_1^{x_1}\alpha_2^{x_2}\ldots\alpha_d^{x_d} & \text{if } x_1 + x_2 + \cdots + x_d = 1 \\ 0 & \text{otherwise} \end{cases}$
- When d = 2, then this is the Bernoulli
- For d > 2, this is called a Categorical distribution

Sampling from a categorical distribution

- The same as sampling proportionally to a table of probabilities
- d items, with associated probabilities $\alpha_1, \ldots, \alpha_{d-1}$ where the probability for the last item is simply $\alpha_d = 1 - \sum_{j=}^{d-1} \alpha_j$

1. Sample u uniformly from $[0, 1]$ ($u \in [0, 1]$)
2. Set $s = 0, k = 1$
3. While $s < u$
   (a) $s \leftarrow s + w_k$
   (b) if $s \geq u$, return k
   (c) $k \leftarrow k + 1$

Sampling from a table of probabilities

- For probability values $w_1, \ldots, w_d$

1. Sample u uniformly from $[0, 1]$ ($u \in [0, 1]$)
2. Set $s = 0, k = 1$
3. While $s < u$
   (a) $s \leftarrow s + w_k$
   (b) if $s \geq u$, return k
   (c) $k \leftarrow k + 1$

More usefully for us: Multi-class classification

- Want to categorize an item into one of d classes
- Only one "roll": n = 1, $x_i = 1$ if the item is categorized as class i
- Sample space: $\mathcal{X} = \{0,1\}^d$ (e.g., outcome is $(0,1,0,0)$ for d = 4)
- $p(x_1, x_2, \ldots, x_d) = \begin{cases} \alpha_1^{x_1}\alpha_2^{x_2}\ldots\alpha_d^{x_d} & \text{if } x_1 + x_2 + \cdots + x_d = 1 \\ 0 & \text{otherwise} \end{cases}$
- When d = 2, then this is the Bernoulli
- **Question**: If you have a dataset with classes $\mathcal{Y} = \{\text{apple}, \text{banana}, \text{orange}\}$, how would you convert it to use this distribution?

More usefully for us: Multi-class classification

- Sample space: $\mathscr{Z} = \{0,1\}^d$ (e.g., outcome is $(0,1,0,0)$ for d = 4)
- $p(z_1, z_2, \ldots, z_d) = \begin{cases} \alpha_1^{z_1}\alpha_2^{z_2}\ldots\alpha_d^{z_d} & \text{if } z_1 + z_2 + \cdots + z_d = 1 \\ 0 & \text{otherwise} \end{cases}$
- **Question**: If you have a dataset with classes $\mathscr{Y} = \{\text{apple}, \text{banana}, \text{orange}\}$, how would you convert it to use this distribution?
- Can rewrite RV Y to vector-valued RV Z that is a multinomial with d = 3
- $p(y = \text{apple} \mid \mathbf{x}) = p(z = (1,0,0) \mid \mathbf{x})) = \alpha_1(\mathbf{x})$
- $p(y = \text{banana} \mid \mathbf{x}) = p(z = (0,1,0) \mid \mathbf{x})) = \alpha_2(\mathbf{x})$
- $p(y = \text{banana} \mid \mathbf{x}) = p(z = (0,0,1) \mid \mathbf{x})) = \alpha_3(\mathbf{x}) = 1 - \alpha_1(\mathbf{x}) - \alpha_2(\mathbf{x})$

* Later we see how to parameterize α_1, α_2 in multinomial logistic regression

Multivariate Gaussian

- $p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^d |\mathbf{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$
- with $\mathbf{\Sigma} \in \mathbb{R}^{d \times d}$ and $\boldsymbol{\mu} \in \mathbb{R}^d$
- The covariance matrix $\mathbf{\Sigma}$ consists of the covariance between each variable
- $\Sigma_{ij} = \text{Cov}(X_i, X_j)$

Important note! This Sigma matrix is not the same as singular values!
We re-use this symbol to mean two different things

The Covariance Matrix

$$\boldsymbol{X} = [X_1, \ldots, X_d]$$

$$\begin{aligned}\Sigma_{ij} &= \mathrm{Cov}[X_i, X_j] \\ &= \mathbb{E}\left[\left(X_i - \mathbb{E}\left[X_i\right]\right)\left(X_j - \mathbb{E}\left[X_j\right]\right)\right]\end{aligned}$$

$$\begin{aligned}\boldsymbol{\Sigma} &= \mathrm{Cov}[\boldsymbol{X}, \boldsymbol{X}] \quad \in \mathbb{R}^{d \times d} \\ &= \mathbb{E}[(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])(\boldsymbol{X} - \mathbb{E}(\boldsymbol{X})^\top] \\ &= \mathbb{E}[\boldsymbol{X}\boldsymbol{X}^\top] - \mathbb{E}[\boldsymbol{X}]\mathbb{E}[\boldsymbol{X}]^\top.\end{aligned}$$

The Covariance Matrix

$$\boldsymbol{X} = [X_1, \ldots, X_d]$$

$$\begin{aligned}\boldsymbol{\Sigma} &= \mathrm{Cov}[\boldsymbol{X}, \boldsymbol{X}] \in \mathbb{R}^{d \times d} \\ &= \mathbb{E}[(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])(\boldsymbol{X} - \mathbb{E}(\boldsymbol{X})^\top] \\ &= \mathbb{E}[\boldsymbol{X}\boldsymbol{X}^\top] - \mathbb{E}[\boldsymbol{X}]\mathbb{E}[\boldsymbol{X}]^\top.\end{aligned}$$

$$\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$$

Dot product

$$\mathbf{x}^\top \mathbf{y} = \sum_{i=1}^{d} x_i y_i$$

Outer product

$$\mathbf{x}\mathbf{y}^\top = \begin{bmatrix} x_1 y_1 & x_1 y_2 & \ldots & x_1 y_d \\ x_2 y_1 & x_2 y_2 & \ldots & x_2 y_d \\ \vdots & \vdots & & \vdots \\ x_d y_1 & x_d y_2 & \ldots & x_d y_d \end{bmatrix}$$

Covariance for two dimensions

$$\boldsymbol{X} = [X_1, \ldots, X_d]$$

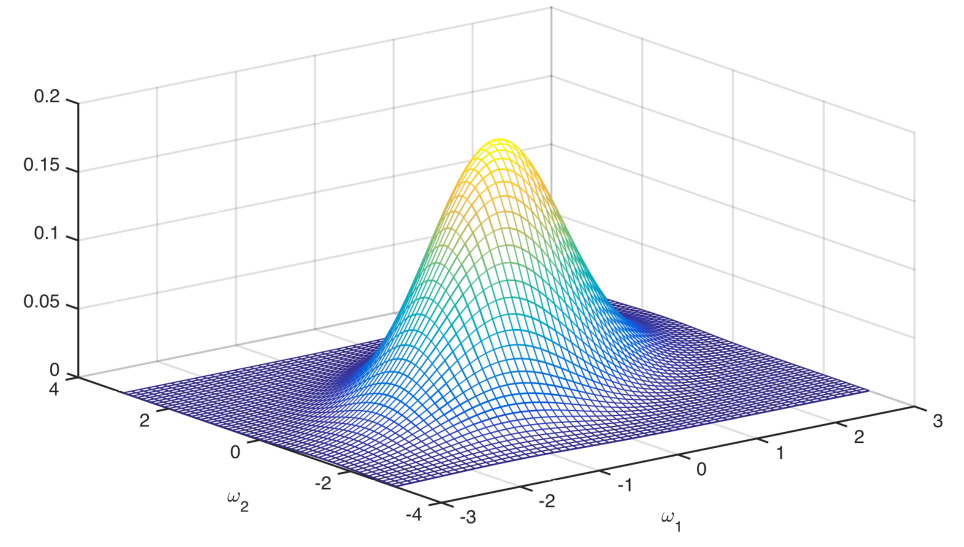
$$\begin{aligned}\boldsymbol{\Sigma} &= \mathrm{Cov}[\boldsymbol{X}, \boldsymbol{X}] \in \mathbb{R}^{d\times d} \\ &= \mathbb{E}[(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])(\boldsymbol{X} - \mathbb{E}(\boldsymbol{X})^\top] \\ &= \mathbb{E}[\boldsymbol{X}\boldsymbol{X}^\top] - \mathbb{E}[\boldsymbol{X}]\mathbb{E}[\boldsymbol{X}]^\top.\end{aligned}$$

$$\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$$

Example:

$$\mathbb{E}\begin{bmatrix} X_1^2 & X_1 X_2 \\ X_2 X_1 & X_2^2 \end{bmatrix} - \begin{bmatrix} \mathbb{E}[X_1]^2 & \mathbb{E}[X_1]\mathbb{E}[X_2] \\ \mathbb{E}[X_2]\mathbb{E}[X_1] & \mathbb{E}[X_2]^2 \end{bmatrix}$$

Multivariate Gaussian Example



$$p(\boldsymbol{\omega}) = \frac{1}{\sqrt{(2\pi)^k|\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\boldsymbol{\omega}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\omega}-\boldsymbol{\mu})\right)$$

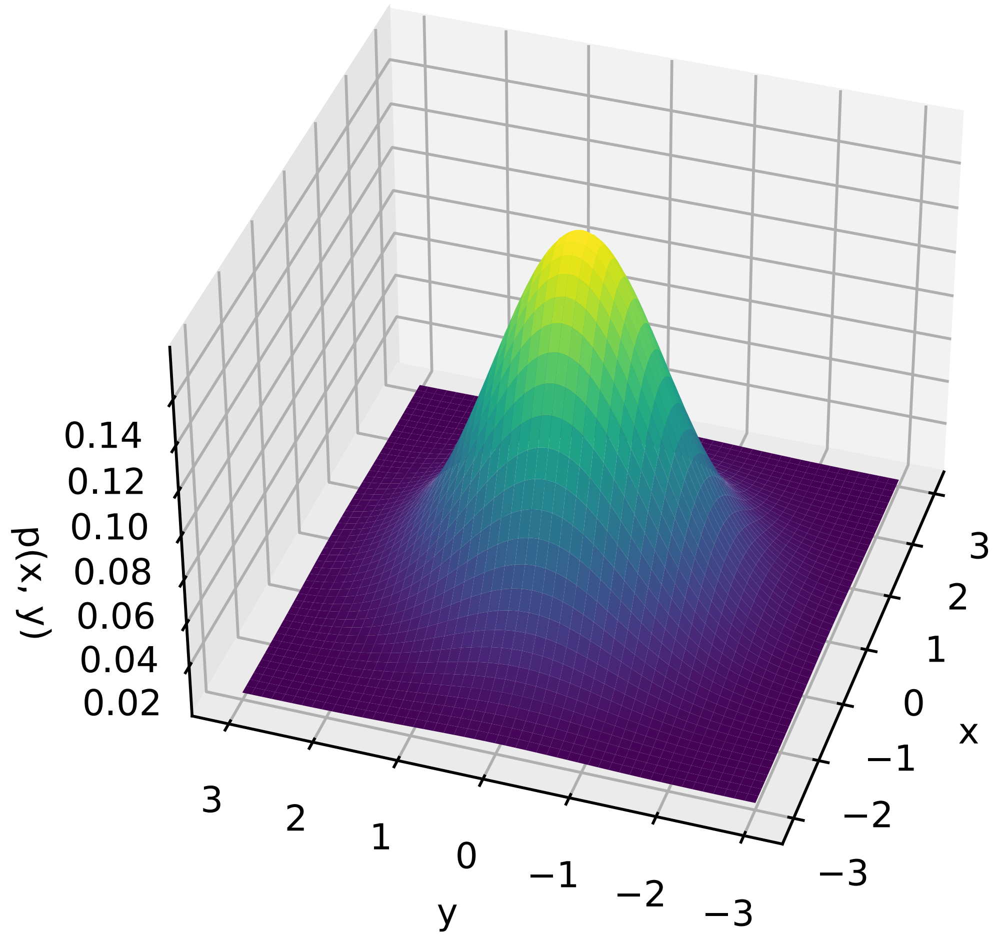
$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \quad \boldsymbol{\Sigma} = \begin{bmatrix} 10 & 0 \\ 0 & 2 \end{bmatrix} \quad \boldsymbol{\Sigma}^{-1} = \begin{bmatrix} \frac{1}{10} & 0 \\ 0 & \frac{1}{2} \end{bmatrix}$$

$$\boldsymbol{\omega} - \boldsymbol{\mu} = \begin{bmatrix} \omega_1 - \mu_1 \\ \omega_2 - \mu_2 \end{bmatrix}$$

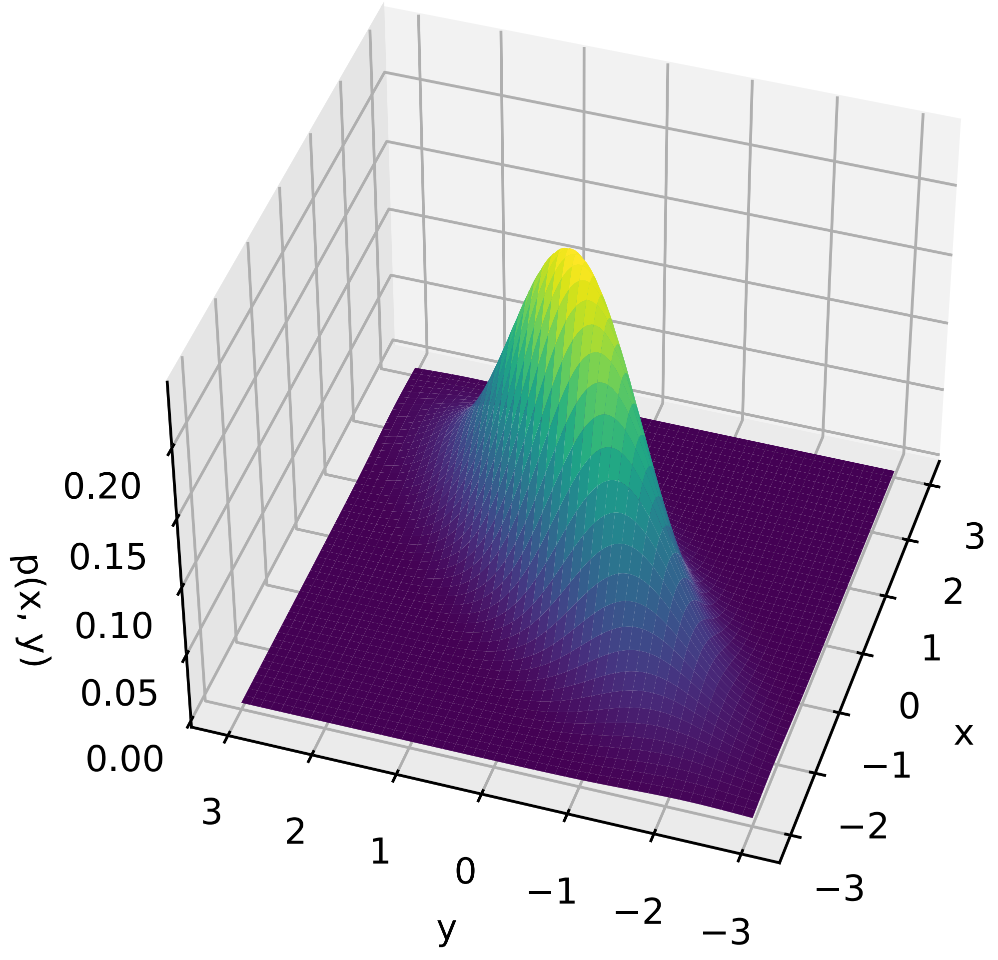
$$\begin{bmatrix} \omega_1 - \mu_1 \\ \omega_2 - \mu_2 \end{bmatrix} \begin{bmatrix} \frac{1}{10} & 0 \\ 0 & \frac{1}{2} \end{bmatrix} = \begin{bmatrix} \frac{1}{10}(\omega_1 - \mu_1) \\ \frac{1}{2}(\omega_2 - \mu_2) \end{bmatrix}$$

$$\begin{bmatrix} \frac{1}{10}(\omega_1 - \mu_1) \\ \frac{1}{2}(\omega_2 - \mu_2) \end{bmatrix}^\top \begin{bmatrix} \omega_1 - \mu_1 \\ \omega_2 - \mu_2 \end{bmatrix} = \frac{1}{10}(\omega_1 - \mu_1)^2 + \frac{1}{2}(\omega_2 - \mu_2)^2$$

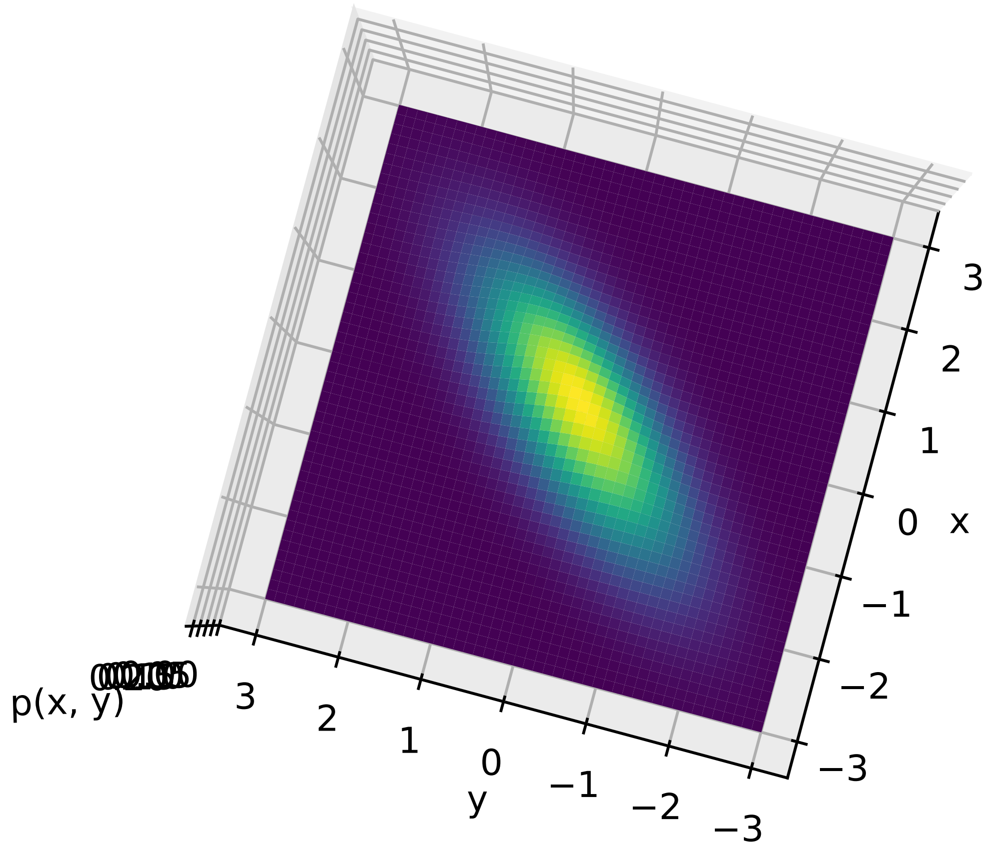
Visually



$$\mathbf{\Sigma} = \begin{bmatrix} 1.0 & 0 \\ 0 & 1.0 \end{bmatrix}$$



$$\mathbf{\Sigma} = \begin{bmatrix} 1.0 & 0.75 \\ 0.75 & 1.0 \end{bmatrix}$$



$$\mathbf{\Sigma} = \begin{bmatrix} 1.0 & 0.75 \\ 0.75 & 1.0 \end{bmatrix}$$

$$\mathbf{\Sigma}^{-1} = \begin{pmatrix} 2.3 & -1.7 \\ -1.7 & 2.3 \end{pmatrix}$$

The weighted norm with correlations

$$\begin{bmatrix} e_1 \\ e_2 \end{bmatrix} \doteq \begin{bmatrix} x_1 - u_1 \\ x_2 - u_2 \end{bmatrix}$$

- The weighted norm gives a distance to the mean, for the covariance

$$\begin{bmatrix} e_1 \\ e_2 \end{bmatrix}^\top \begin{bmatrix} 2.3 & -1.7 \\ -1.7 & 2.3 \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \end{bmatrix} = \begin{bmatrix} 2.3e_1 - 1.7e_2 \\ -1.7e_1 + 2.3e_2 \end{bmatrix}^\top \begin{bmatrix} e_1 \\ e_2 \end{bmatrix}$$

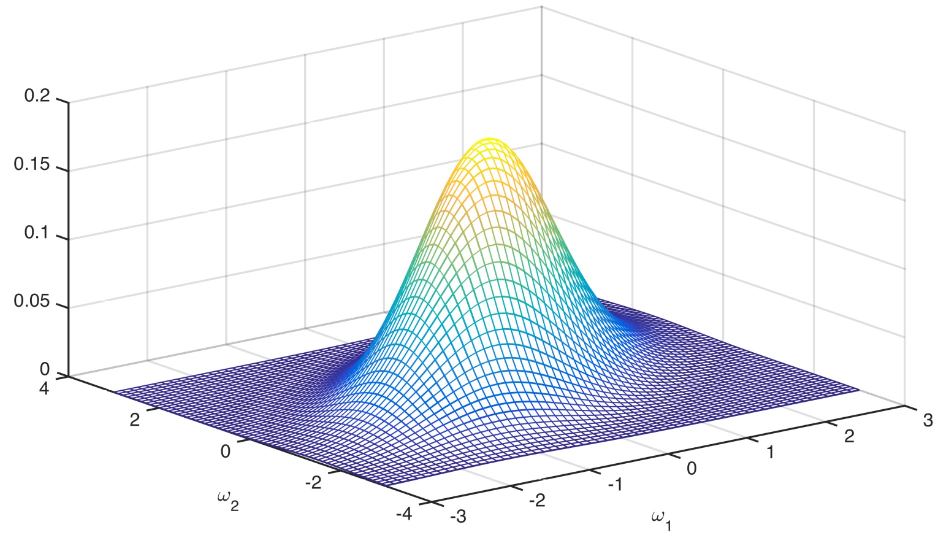
$$= 2.3e_1^2 + 2.3e_2^2 - 2.4e_1e_2$$

- If e_1 is the opposite sign from e_2, then the distance is larger (-2.4 * negative number = positive number added to distance)

- If e_1 is the same sign as e_2, then the distance is larger (-2.4 * positive = negative)

The determinant component

$$p(\boldsymbol{\omega}) = \frac{1}{\sqrt{(2\pi)^k |\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\boldsymbol{\omega} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\omega} - \boldsymbol{\mu})\right)$$



$$\boldsymbol{\Sigma} = \begin{bmatrix} 10 & 0 \\ 0 & 2 \end{bmatrix}$$

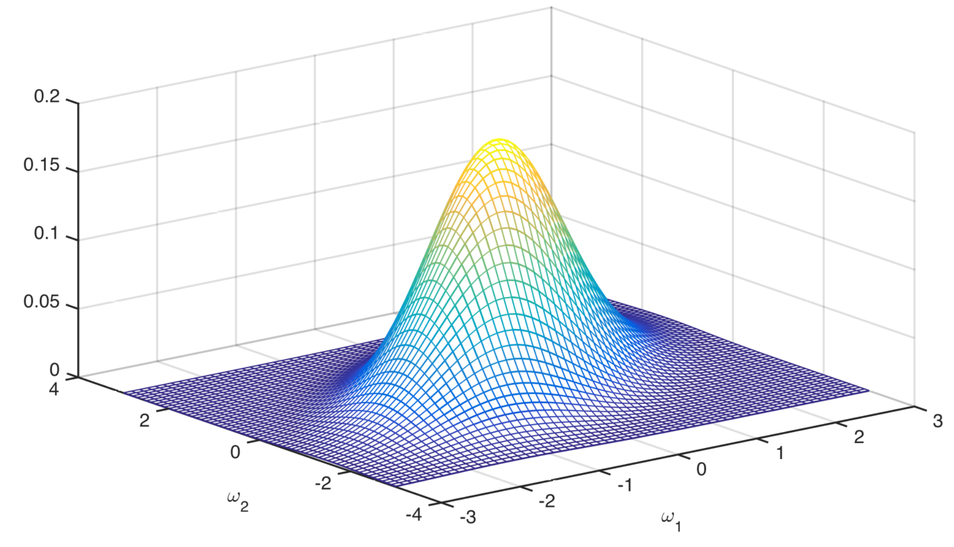
$|\boldsymbol{\Sigma}| = \det(\boldsymbol{\Sigma}) =$ product of singular values

(reflects the magnitude of the covariance)

What is the determinant of this Sigma?

The determinant component

$$p(\boldsymbol{\omega}) = \frac{1}{\sqrt{(2\pi)^k |\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\boldsymbol{\omega} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\omega} - \boldsymbol{\mu})\right)$$



$$\boldsymbol{\Sigma} = \left[\begin{array}{cc} 10 & 0 \\ 0 & 2 \end{array}\right]$$

$|\boldsymbol{\Sigma}| = \det(\boldsymbol{\Sigma}) =$ product of singular values

(reflects the magnitude of the covariance)

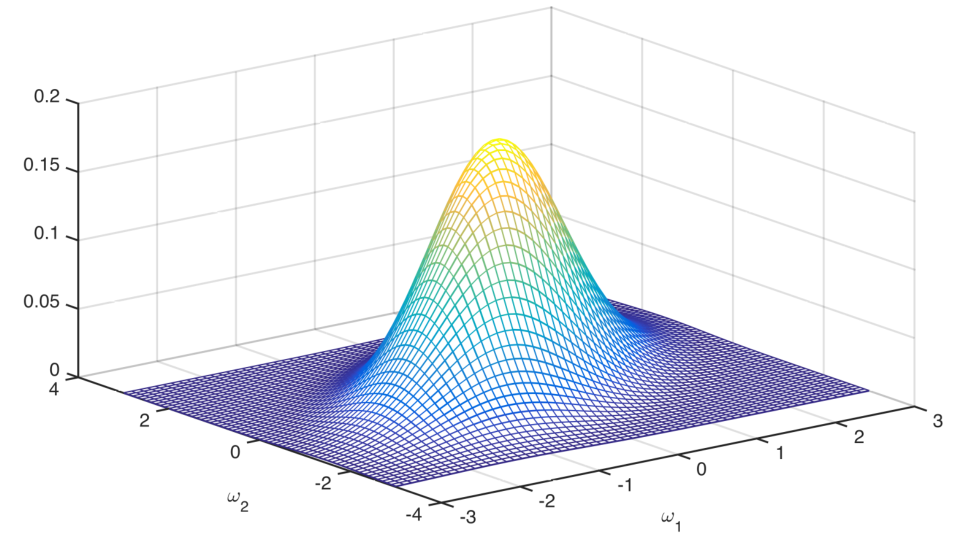
What is the determinant of this other Sigma?

$$\boldsymbol{\Sigma} = \left[\begin{array}{cc} 1.0 & 0.75 \\ 0.75 & 1.0 \end{array}\right]$$

It has singular values: $\sigma_1 = 1.75, \sigma_2 = 0.25$

The determinant component

$$p(\boldsymbol{\omega}) = \frac{1}{\sqrt{(2\pi)^k |\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\boldsymbol{\omega} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\omega} - \boldsymbol{\mu})\right)$$



$$\boldsymbol{\Sigma} = \begin{bmatrix} 10 & 0 \\ 0 & 2 \end{bmatrix}$$

$|\boldsymbol{\Sigma}| = \det(\boldsymbol{\Sigma}) =$ product of singular values

(reflects the magnitude of the covariance)

What is the determinant of this other Sigma?

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1.0 & 0.75 \\ 0.75 & 1.0 \end{bmatrix}$$

It has singular values: $\sigma_1 = 1.75, \sigma_2 = 0.25$

Answer: $\sigma_1 \times \sigma_2 \approx 0.44$

Mixture of Distributions

Mixture model:

A set of m probability distributions, $\{p_i(x)\}_{i=1}^m$

$$p(x) = \sum_{i=1}^{m} w_i p_i(x)$$

where $\boldsymbol{w} = (w_1, w_2, \ldots, w_m)$ and non-negative and

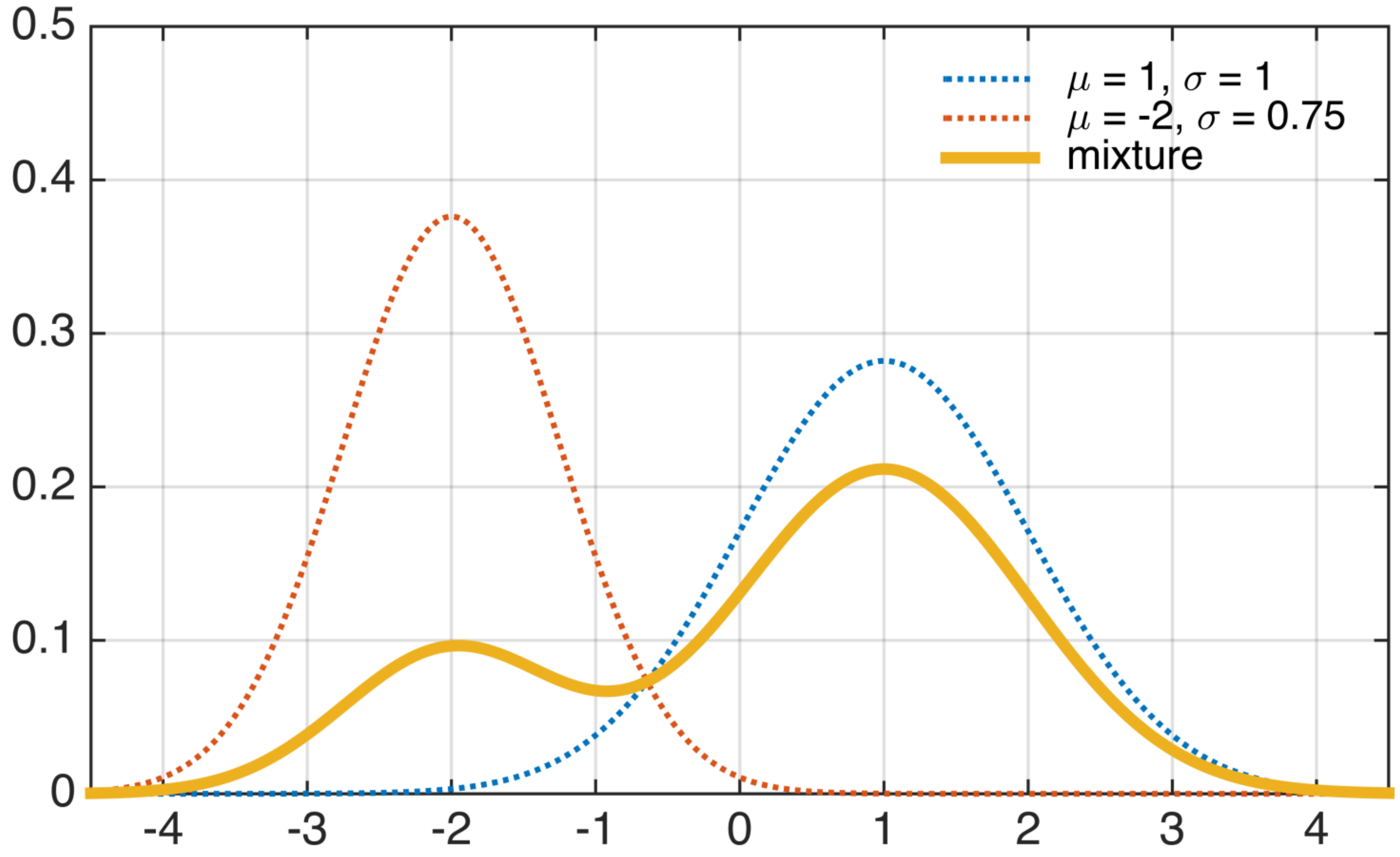
$\sum_{i=1}^m w_i = 1$

Mixture of Gaussians

$$p(x) = \sum_{i=1}^{m} w_i p_i(x)$$

Mixture of $m = 2$ Gaussian distributions:

$$w_1 = 0.75, \; w_2 = 0.25$$



Exercise

- Show that $p(x) = \sum_{i=1}^{m} w_i p_i(x)$ is a valid pmf if the p_i are valid pmfs
  - when $\sum_{i=1}^{m} w_i = 1$ and $w_i \geq 0$
- Show this also for the case where p is a pdf and the p_i are pdfs

Exercise Solution for PMFs

- $p(x) = \sum_{i=1}^{m} w_i p_i(x)$
- $p(x) \geq 0$ because $w_i p_i(x) \geq 0$, sum of nonnegative #s is nonnegative

Exercise Solution for PMFs

$$\begin{aligned}
\sum_{x \in \mathcal{X}} p(x) &= \sum_{x \in \mathcal{X}} \sum_{i=1}^{m} w_i p_i(x) \\
&= \sum_{i=1}^{m} \sum_{x \in \mathcal{X}} w_i p_i(x) \\
&= \sum_{i=1}^{m} w_i \underbrace{\sum_{x \in \mathcal{X}} p_i(x)}_{=1} \\
&= \sum_{i=1}^{m} w_i = 1
\end{aligned}$$

Exercise Solution for PDFs

$$\sum_{x \in \mathcal{X}} p(x) = \sum_{x \in \mathcal{X}} \sum_{i=1}^{m} w_i p_i(x)$$

$$= \sum_{i=1}^{m} \sum_{x \in \mathcal{X}} w_i p_i(x)$$

$$= \sum_{i=1}^{m} w_i \underbrace{\sum_{x \in \mathcal{X}} p_i(x)}_{=1}$$

$$= \sum_{i=1}^{m} w_i = 1$$

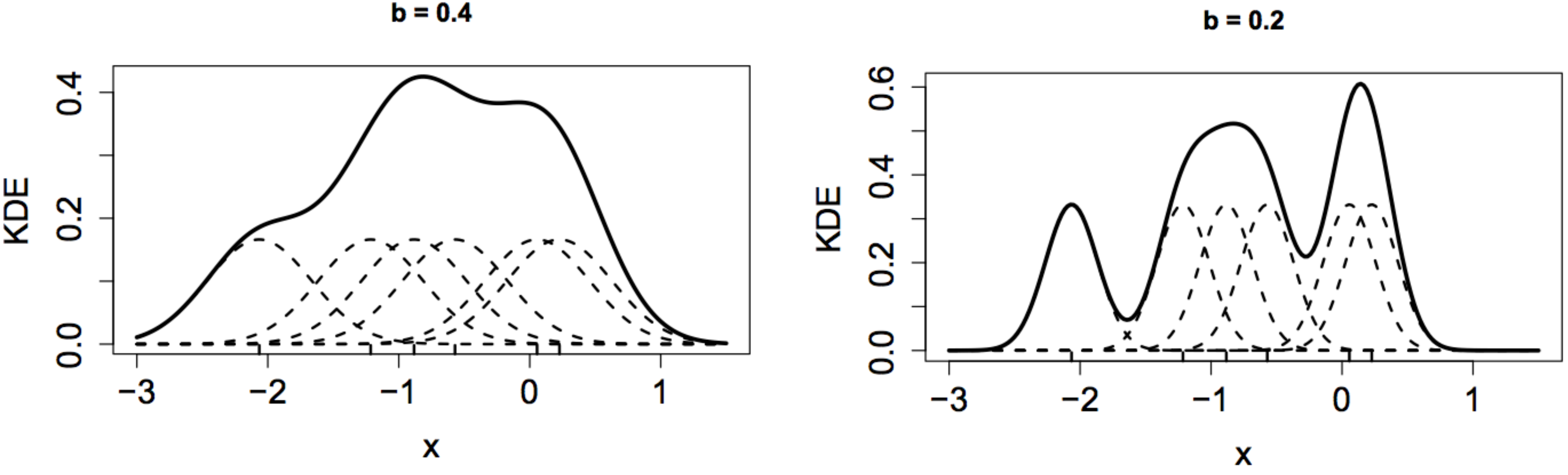
$$\int_{\mathcal{X}} p(x) dx = \int_{\mathcal{X}} \sum_{i=1}^{m} w_i p_i(x) dx$$

$$= \sum_{i=1}^{m} \int_{\mathcal{X}} w_i p_i(x) dx$$

$$= \sum_{i=1}^{m} w_i \underbrace{\int_{\mathcal{X}} p_i(x) dx}_{=1}$$

$$= \sum_{i=1}^{m} w_i = 1$$

Mixture Can Produce Complex Distributions



* Image from https://people.ucsc.edu/~ealdrich/Teaching/Econ114/LectureNotes/kde.html

Exercise Question

- Multidimensional PMFs essentially allow any distribution (table of probabilities)
- Densities for Continuous RVs are more restricted (even with mixtures)
- Why not just discretize our variables and use PMFs?
- Example: imagine the RV is in the range [-10, 10]
- You discretize into chunks of size 0.1. How many parameters do you have to learn?
- What if you use a Gaussian mixture with 5 components?

Contrast to Sum of Gaussians

- Let $Y = w_1 X_1 + w_2 X_2$ for $w_1, w_2 \geq 0, w_1 + w_2 = 1$
- Let X be an RV with a pdf that is Gaussian mixture model with two components, and the same weights $w_1, w_2 \geq 0, w_1 + w_2 = 1$
- $X \neq Y$
- Y is a Gaussian RV, so they can’t be the same (bimodal vs unimodal)
- Mixture model uses **convex combo of pdfs**, not of RVs

Independence and Decorrelation

- Recall if X and Y are independent, then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$
- Independent RVs have zero correlation

  Recall: $\text{Cov}[X, Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$

- Uncorrelated RVs (i.e., $\text{Cov}(X, Y) = 0$) **might be dependent** (i.e., $p(x, y) \neq p(x)p(y)$).
  - Correlation (**Pearson's correlation coefficient**) shows linear relationships; but can miss nonlinear relationships
  - **Example:** $X \sim \text{Uniform}\{-2, -1, 0, 1, 2\}$, $Y = X^2$
    - $\mathbb{E}[XY] = .2(-2 \times 4) + .2(2 \times 4) + .2(-1 \times 1) + .2(1 \times 1) + .2(0 \times 0)$
    - $\mathbb{E}[X] = 0$
    - So $\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0 - 0\mathbb{E}[Y] = 0$

Alternative: Mutual Information (using the KL Divergence)

Mutual information $I(X;Y) = D_{KL}(p_{xy}||p_x p_y)$

Only zero when X and Y independent

Entropy

- $$H(X) = \begin{cases} -\sum_{x \in \mathcal{X}} p(x) \log p(x) & X \text{ discrete} \\ -\int_{\mathcal{X}} p(x) \log p(x) dx & X \text{ continuous} \end{cases}$$

- Entropy measures level of dispersion (like variance), but looks at the total spread in probabilities, rather than deviation from the mean

- For a zero-mean $\mathbf{X}$, $H(\mathbf{X}) \leq \dfrac{d}{2}(\ln 2\pi + 1 + \ln \det \mathbf{\Sigma})$
  - equal if X is a multivariate Gaussian

- Another example: entropy of exponential distribution is $-ln\lambda + 1$, whereas the variance is $1/\lambda^2$ (mean is $1/\lambda$)

Exponential Distribution

An **exponential distribution** is a distribution over the positive reals. It has one parameter $\lambda > 0$.

$\Omega = \mathbb{R}^+$

entropy = $-ln\lambda + 1$

variance = $1/\lambda^2$ (mean is $1/\lambda$)

$$p(\omega) = \lambda \exp(-\lambda\omega)$$

lambda = 0.5

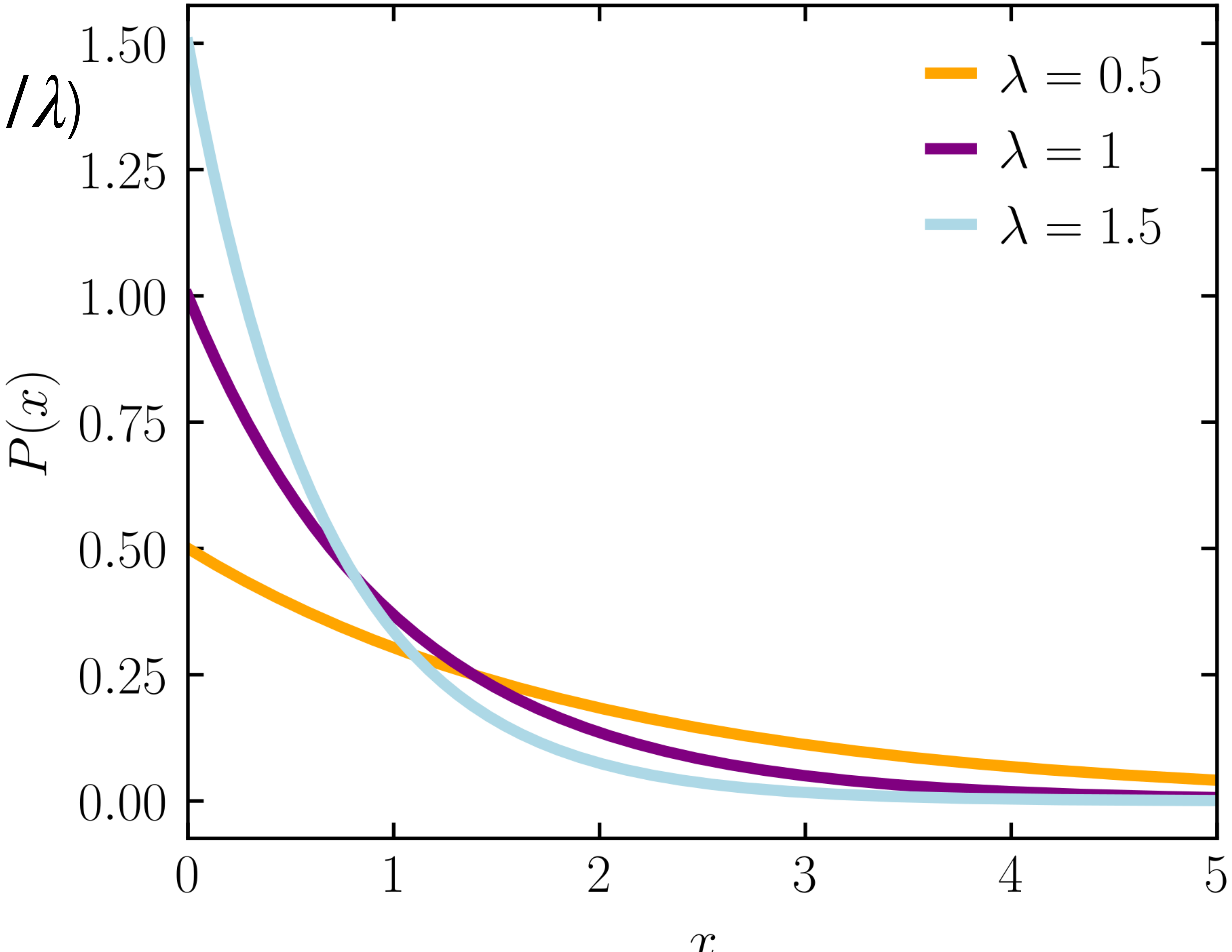
entropy = $-ln0.5 + 1 \approx 1.7$

variance = $1/0.5^2 = 4$

lambda = 1.5

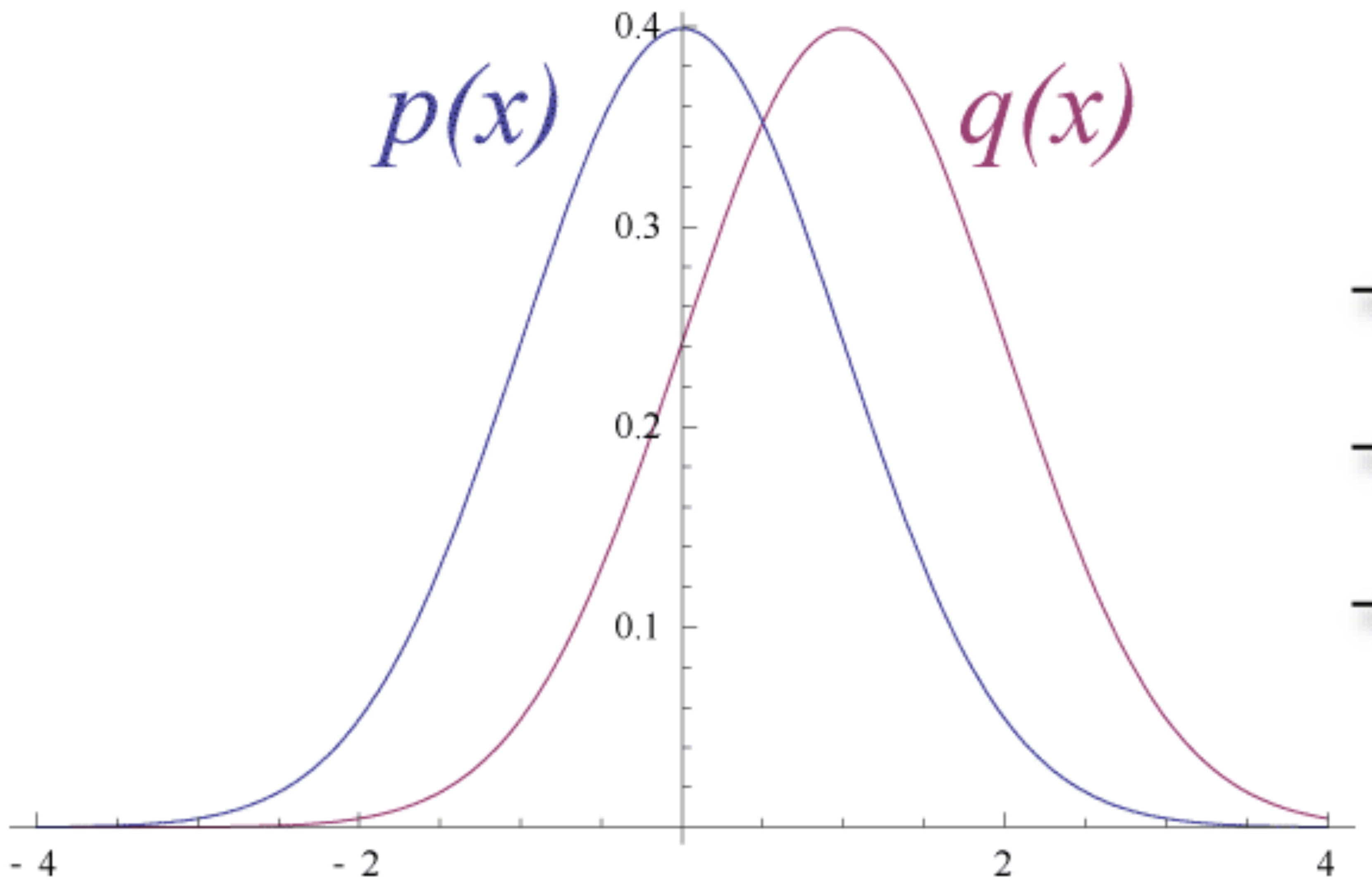
entropy = $-ln1.5 + 1 \approx 0.6$

variance = $1/1.5^2 \approx 0.44$

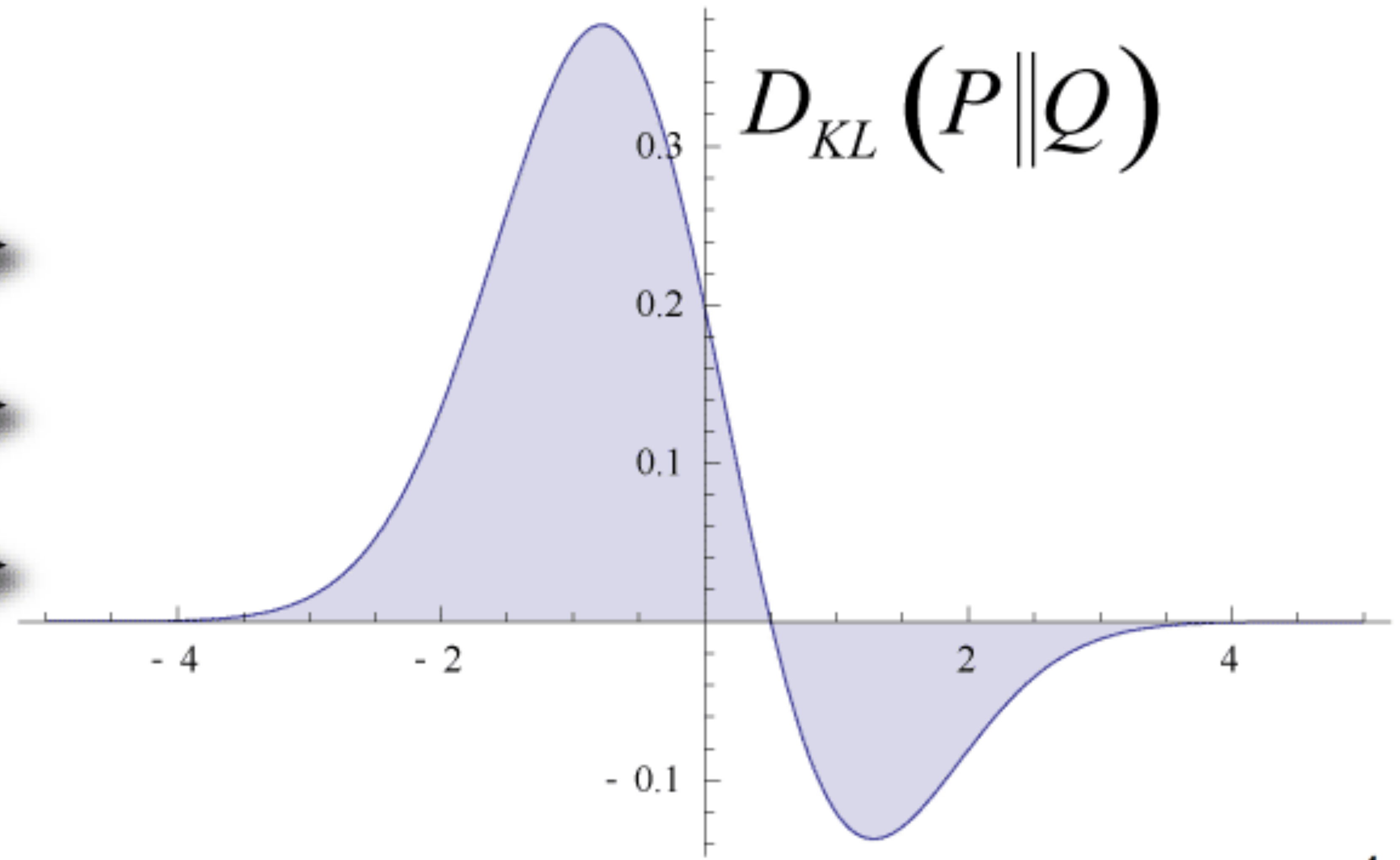


KL Divergence

* Images from Wikipedia



Original Gaussian PDF's



KL Area to be Integrated

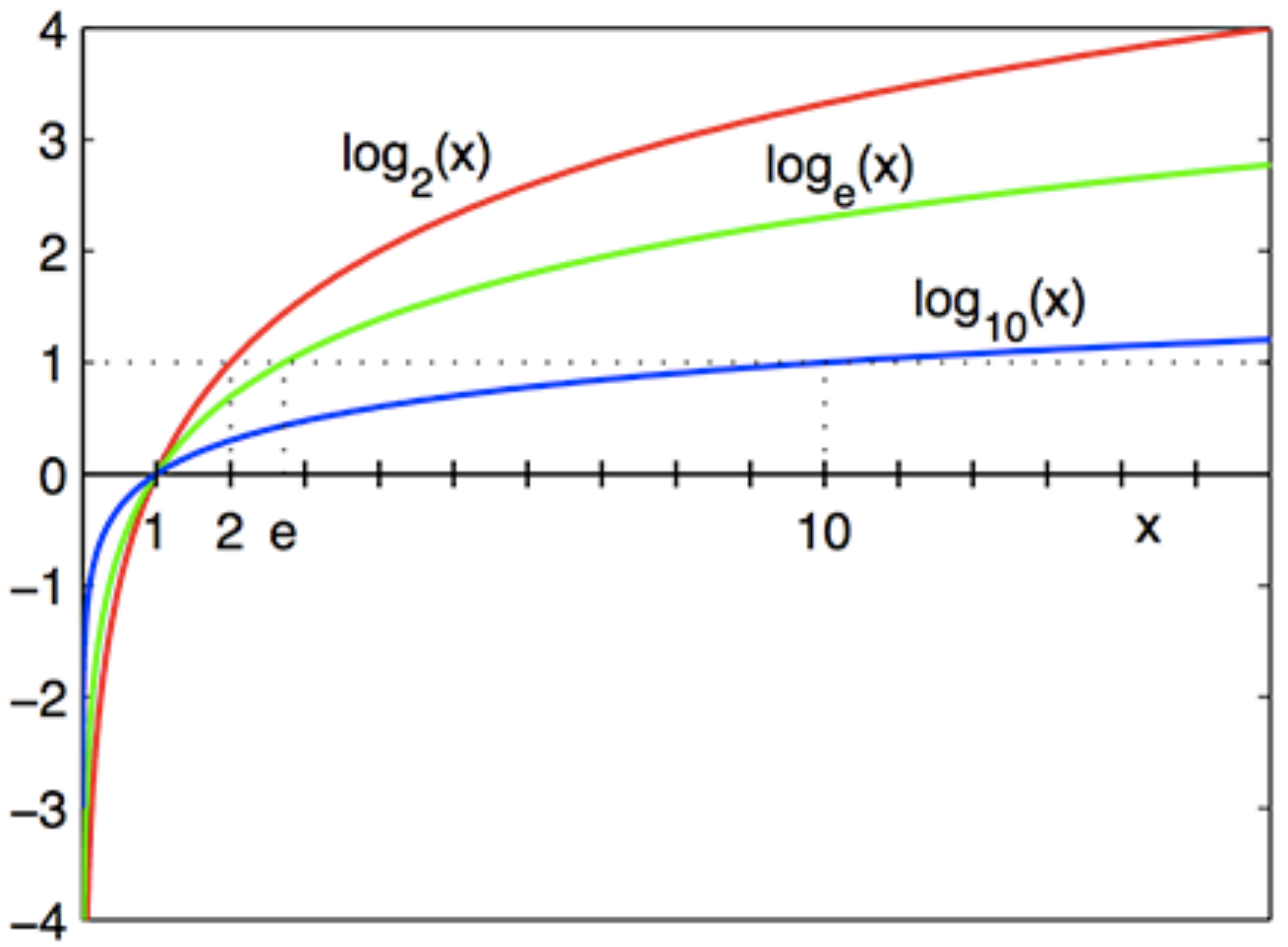
$$\mathrm{KL}(p||q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}$$

or

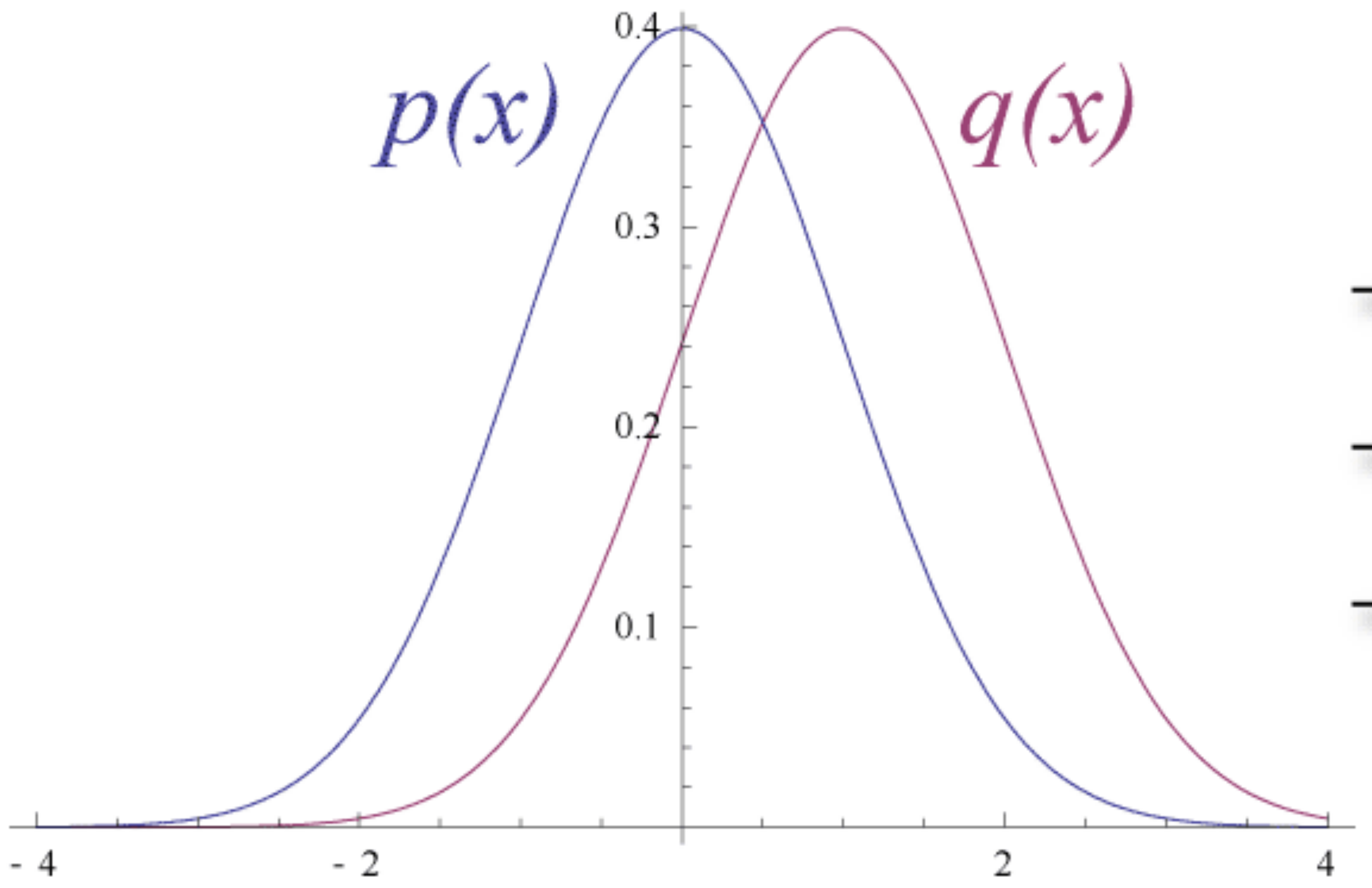
$$\mathrm{KL}(p||q) = \int_{\mathcal{X}} p(x) \log \frac{p(x)}{q(x)} dx$$

Called a divergence, does not satisfy requirements to be a metric/distance

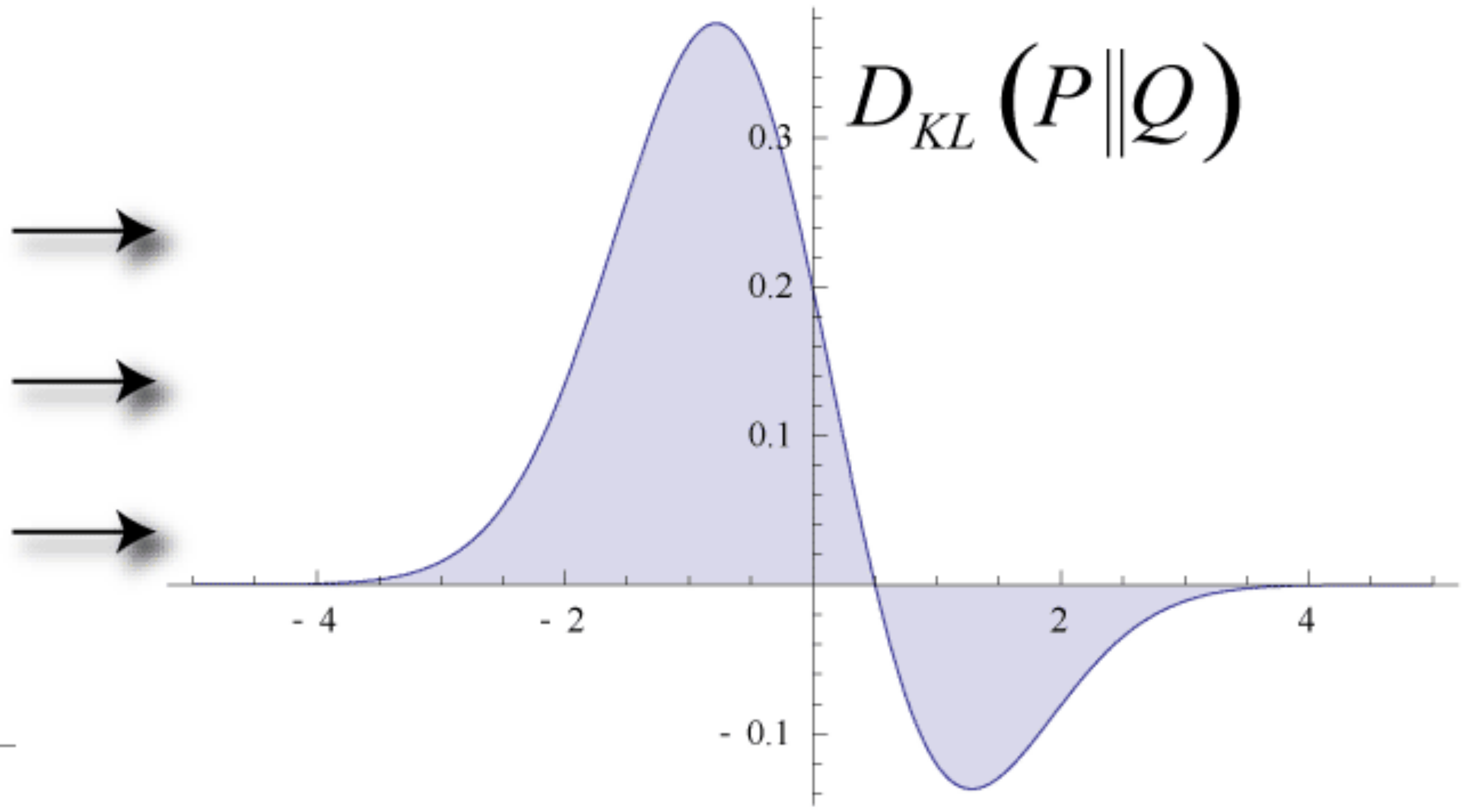
- Not symmetric
- But does satisfy $D_{\text{KL}}(p\,||\,q) \geq 0$ and
- $D_{\text{KL}}(p\,||\,q) = 0$ if and only if (iff) $p = q$



Alternative: Mutual Information (using the KL Divergence)



Original Gaussian PDF's



KL Area to be Integrated

$$\mathrm{KL}(p||q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}$$

or

$$\mathrm{KL}(p||q) = \int_{\mathcal{X}} p(x) \log \frac{p(x)}{q(x)} dx$$

Mutual information $I(X;Y) = D_{KL}(p_{xy}||p_x p_y)$

Revisiting Our Example

- **Example:** $X \sim$ Uniform$\{-2, -1, 0, 1, 2\}$, $Y = X^2$
  - $\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0 - 0\mathbb{E}[Y] = 0$
- $\mathcal{X} = \{-2, -1, 0, 1, 2\}$ and $\mathcal{Y} = \{0, 1, 4\}$
- $p(x, y) = 0$ if $y \neq x^2$, and else is 1/5 (is this a valid pmf? how do you know?)
- $p_x(x) = 1/5$ and $p_y(0) = 1/5, p_y(1) = 2/5, p_y(4) = 2/5$
- KL$(p || p_x p_y) = \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} p(x, y) \log \frac{p(x, y)}{p_x(x) p_y(y)}$

Revisiting Our Example

- $p(x, y) = 0$ if $y \neq x^2$, and else is 1/5 (is this a valid pmf? how do you know?)
- $p_x(x) = 1/5$ and $p_y(0) = 1/5, p_y(1) = 2/5, p_y(4) = 2/5$
-

$$\text{KL}(p \,||\, p_x p_y) = \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} p(x, y) \log \frac{p(x, y)}{p_x(x) p_y(y)}$$

$$= \sum_{x \in \mathcal{X}, y = x^2} \frac{1}{5} \log \frac{1/5}{1/5 p_y(y)}$$

$$= \frac{1}{5} \sum_{x \in \mathcal{X}, y = x^2} \log \frac{1}{p_y(y)}$$

$$= \frac{1}{5}[\log \frac{1}{1/5} + 4 \log \frac{1}{2/5}] = \frac{1}{5}[\log 5 + 4 \log \frac{5}{2}] \approx 1.05 \neq 0$$

Fun Fact

- Imagine you want to learn a distribution. There is some true underlying distribution p_0, but you do not know even what type it is
  - Might be Gaussian, might be a mixture model, might be something we don't have a name for
- Minimizing the KL to the true distribution corresponds to minimizing the negative log likelihood in expectation over all data
- $\arg\min_\theta D_{\text{KL}}(p_0 || p_\theta) = \arg\min_\theta - \mathbb{E}[\ln p_\theta(X)]$
- Further motivates using MLE, since with more data we get closer and closer to minimizing $-\mathbb{E}[\ln p_\theta(X)] \approx \frac{1}{n}\sum_{i=1}^{n} - \ln p_\theta(x_i)$

Fun Fact

- Imagine you want to learn a distribution. There is some true underlying distribution p_0, but you do not know even what type it is
  - Might be Gaussian, might be a mixture model, might be something we don't have a name for
- $\arg\min_\theta D_{\text{KL}}(p_0 || p_\theta) = \arg\min_\theta - \mathbb{E}[\ln p_\theta(X)]$
- **Question**: Imagine we learn a Gaussian, and the true distribution is Gaussian. Is there a p_θ that can get zero $D_{\text{KL}}(p_0 || p_\theta)$?
- What if we learn a Gaussian, but p_θ is a mixture model?