Probability

CMPUT 467: Machine Learning II

Chapter 2

PMFs and PDFs

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

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

Discrete case:

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

Continuous case:

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

Can also write it this way

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

Discrete case:

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

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

Continuous case:

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

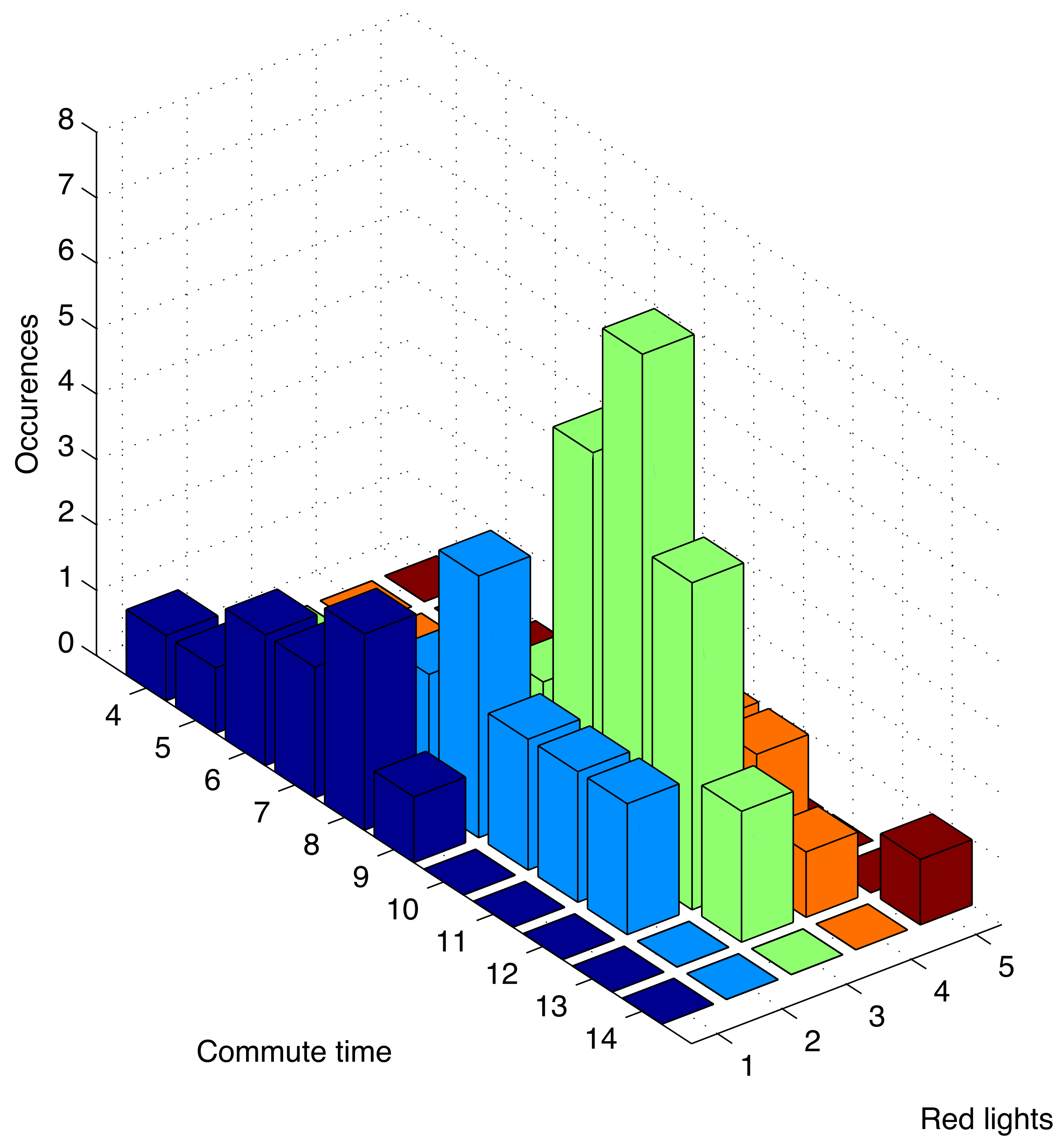
$$\int_{\mathscr{X}_1} \int_{\mathscr{X}_2} \cdots \int_{\mathscr{X}_d} p(x_1, x_2, \ldots, x_d)\, dx_1 dx_2 \ldots dx_d = 1$$

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



Utility for classification

- Want to categorize an item into one of d classes
- Sample space: $\mathcal{X} = \{0,1\}^d$ (e.g., outcome is $(0,1,0,0)$ for class 2 for d = 4)
- PMF is a table of probabilities, but we can write is compactly as
- $p(x_1, x_2, \ldots, x_d) = \begin{cases} \alpha_1^{x_1}\alpha_2^{x_2}\ldots\alpha_d^{x_d} & \text{if } x_1 + x_2 + \cdots + x_d = 1 \\ 0 & \text{otherwise} \end{cases}$
- When d = 2, then this is the Bernoulli $p(x) = \alpha^x(1-\alpha)^{(1-x)}$ for $\alpha_1 = \alpha, \alpha_2 = 1 - \alpha$
- For d > 2, this is called a Categorical distribution

Utility for classification

- Sample space: $\mathscr{X} = \{0,1\}^d$ (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 $p(x) = \alpha^x(1-\alpha)^{(1-x)}$ for $\alpha_1 = \alpha, \alpha_2 = 1 - \alpha$
- For d > 2, this is called a Categorical distribution
- **Exercise**: how do we write the Categorical using only $\alpha_1, \alpha_2, \ldots, \alpha_{d-1}$?

Utility for classification (simpler)

- Sample space: $\mathscr{X} = \{0,1\}^d$ (e.g., outcome is $(0,1,0,0)$ for d = 4)
- $p(x_1, x_2, \ldots, x_d) = \alpha_1^{x_1}\alpha_2^{x_2}\ldots\alpha_d^{x_d}$ assuming $x_1 + x_2 + \cdots + x_d = 1$
- When d = 2, then this is the Bernoulli $p(x) = \alpha^x(1-\alpha)^{(1-x)}$ for $\alpha_1 = \alpha, \alpha_2 = 1 - \alpha$
- For d > 2, this is called a Categorical distribution
- **Exercise**: how do we write the Categorical using only $\alpha_1, \alpha_2, \ldots, \alpha_{d-1}$?

Exercise Answer

- Sample space: $\mathcal{X} = \{0,1\}^d$ (e.g., outcome is $(0,1,0,0)$ for d = 4)
- $p(x_1, x_2, \ldots, x_d) = \alpha_1^{x_1}\alpha_2^{x_2}\ldots\alpha_d^{x_d}$ assuming $x_1 + x_2 + \cdots + x_d = 1$
- When d = 2, then this is the Bernoulli $p(x) = \alpha^x(1-\alpha)^{(1-x)}$ for $\alpha_1 = \alpha, \alpha_2 = 1 - \alpha$
- For d > 2, this is called a Categorical distribution
- **Exercise**: how do we write the Categorical using only $\alpha_1, \alpha_2, \ldots, \alpha_{d-1}$?
- $p(x_1, x_2, \ldots, x_d) = \alpha_1^{x_1}\alpha_2^{x_2}\ldots\alpha_{d-1}^{x_{d-1}}\left(1 - \sum_{j=1}^{d-1}\alpha_j\right)^{x_d}$ because $\alpha_d = 1 - \sum_{j=1}^{d-1}\alpha_j$

Utility for classification

- Want to categorize an item into one of d classes
- Sample space: $\mathcal{X} = \{0,1\}^d$ (e.g., outcome is $(0,1,0,0)$ for d = 4)
- PMF is a table of probabilities, but we can write is compactly as
- $p(x_1, x_2, \ldots, x_d) = \alpha_1^{x_1}\alpha_2^{x_2}\ldots\alpha_d^{x_d}$ assuming $x_1 + x_2 + \cdots + x_d = 1$
- **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?

Exercise Answer

- Sample space: $\mathcal{X} = \{0,1\}^d$ (e.g., outcome is $(0,1,0,0)$ for d = 4)
- $p(x_1, x_2, \ldots, x_d) = \alpha_1^{x_1} \alpha_2^{x_2} \ldots \alpha_d^{x_d}$ assuming $x_1 + x_2 + \cdots + x_d = 1$
- **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?
- Can rewrite RV Y to vector-valued RV $\mathbf{X}$ with d = 3, where
- $p(y = \text{apple}) = p(\mathbf{x} = (1,0,0)) = \alpha_1$
- $p(y = \text{banana}) = p(\mathbf{x} = (0,1,0)) = \alpha_2$
- $p(y = \text{orange}) = p(\mathbf{x} = (0,0,1)) = \alpha_3 = 1 - \alpha_1 - \alpha_2$

We did not have to call it X, can use any term for categorical variable

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

Conditional PMF

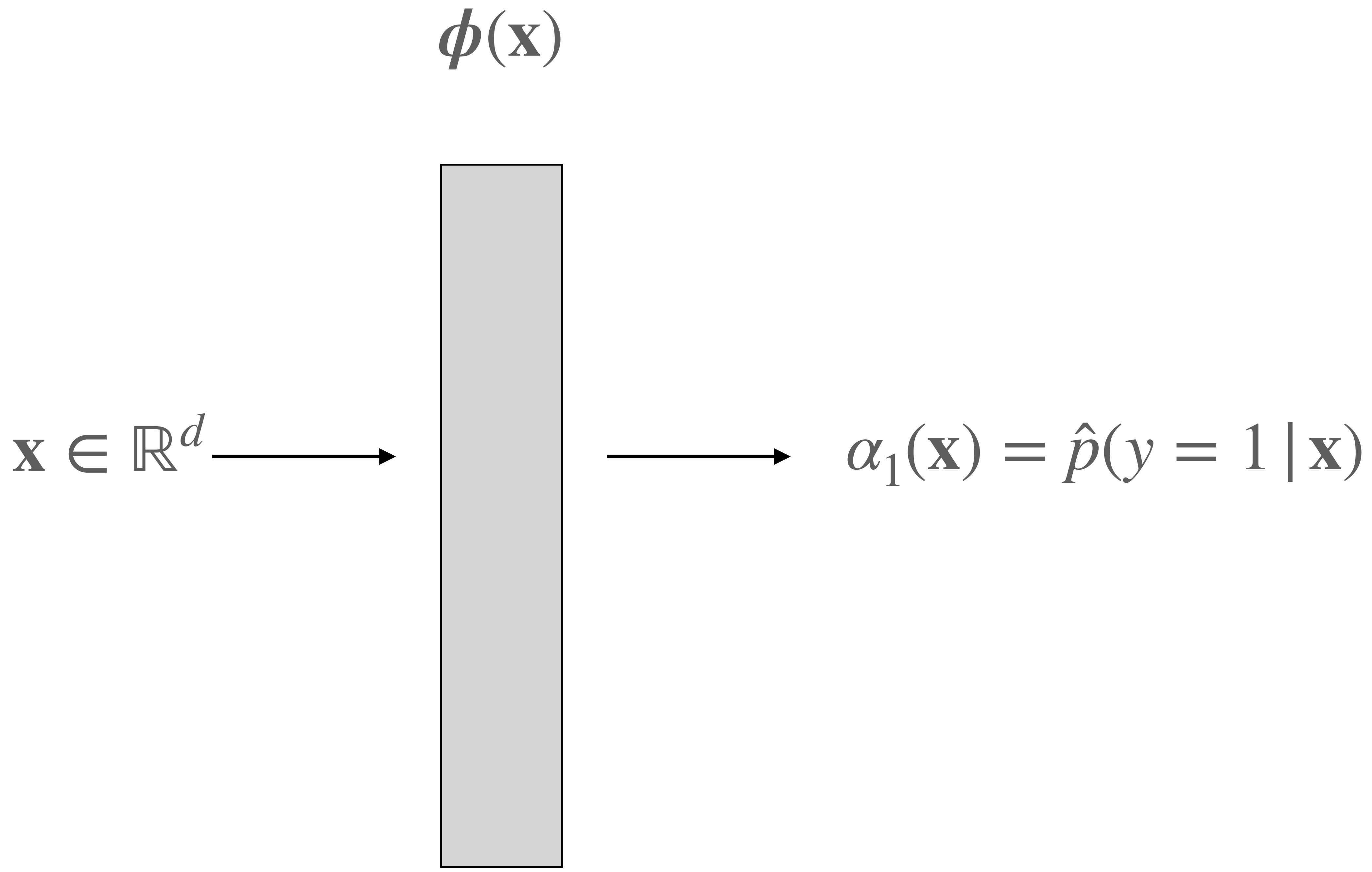
- In classification, we actually learned a conditional PMF on inputs $\mathbf{x} \in \mathbb{R}^d$
- How do we write the conditional distribution for $\mathscr{Y} = \{\text{apple}, \text{banana}, \text{orange}\}$?

Conditional PMF Example

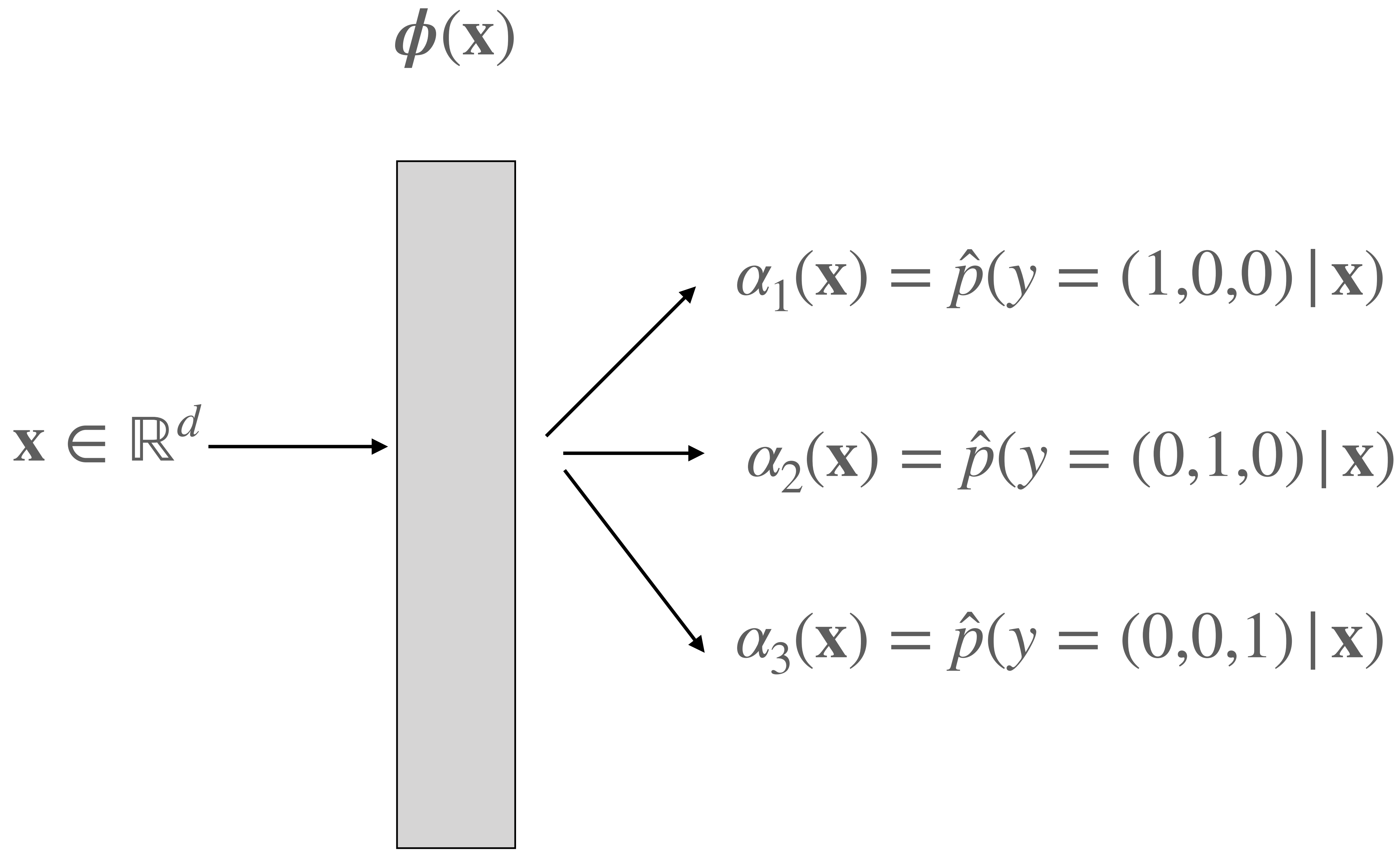
- Classes $\mathscr{Y} = \{\text{apple, banana, orange}\}$, inputs $\mathbf{x} \in \mathbb{R}^d$
- As before, we rewrite RV Y to vector-valued RV $\mathbf{Z}$ that is a multinomial with d = 3
- But now probabilities are functions of inputs $\mathbf{x} \in \mathbb{R}^d$
- $p(y = \text{apple} \,|\, \mathbf{x}) = p(z = (1,0,0) \,|\, \mathbf{x})) = \alpha_1(\mathbf{x})$
- $p(y = \text{banana} \,|\, \mathbf{x}) = p(z = (0,1,0) \,|\, \mathbf{x})) = \alpha_2(\mathbf{x})$
- $p(y = \text{banana} \,|\, \mathbf{x}) = p(z = (0,0,1) \,|\, \mathbf{x})) = \alpha_3(\mathbf{x})$

Contrasting binary versus multiclass

Binary Classification



Multiclass Classification



* Later we see how to parameterize these functions in multinomial logistic regression

Multivariate Gaussian

- $p(\mathbf{x}) = \dfrac{1}{\sqrt{(2\pi)^d |\mathbf{\Sigma}|}} \exp\left(-\dfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$
- 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

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

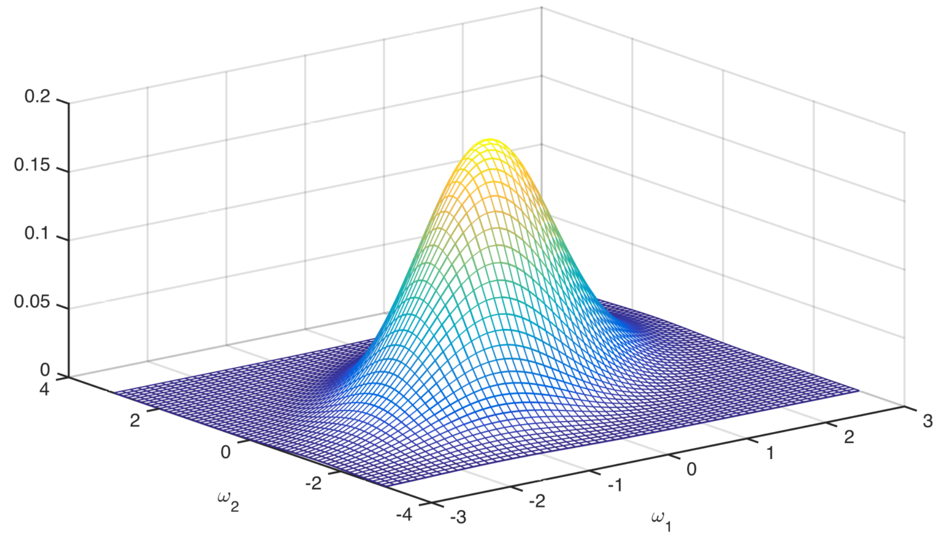
$$\begin{aligned}\boldsymbol{\Sigma} &= \mathrm{Cov}[\boldsymbol{X}, \boldsymbol{X}] \in \mathbb{R}^{d \times d} \\ &= \mathbb{E}[(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])(\boldsymbol{X} - \mathbb{E}(\boldsymbol{X})^\top] \\ &= \mathbb{E}[\boldsymbol{X}\boldsymbol{X}^\top] - \mathbb{E}[\boldsymbol{X}]\mathbb{E}[\boldsymbol{X}]^\top.\end{aligned}$$

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

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



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

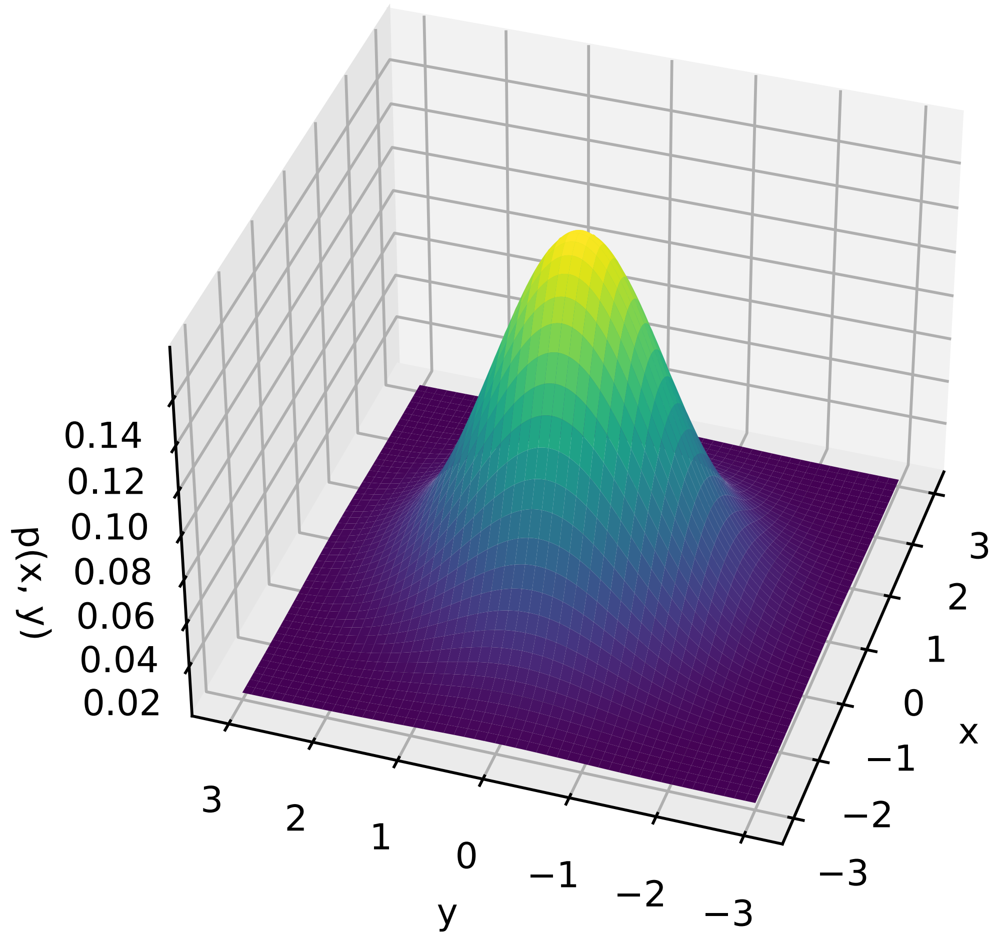
$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \quad \boldsymbol{\Sigma} = \begin{bmatrix} 10 & 0 \\ 0 & 2 \end{bmatrix} \quad \boldsymbol{\Sigma}^{-1} = \begin{bmatrix} \frac{1}{10} & 0 \\ 0 & \frac{1}{2} \end{bmatrix}$$

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

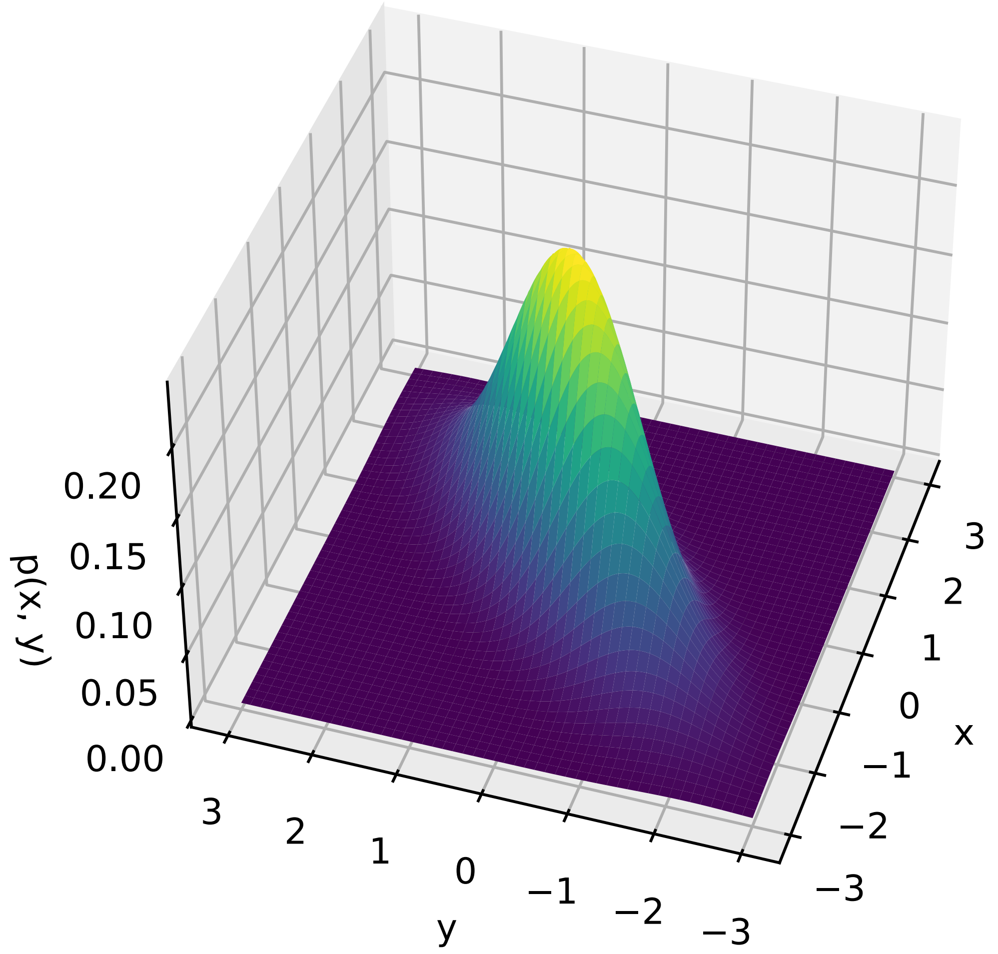
$$\begin{bmatrix} \omega_1 - \mu_1 \\ \omega_2 - \mu_2 \end{bmatrix} \begin{bmatrix} \frac{1}{10} & 0 \\ 0 & \frac{1}{2} \end{bmatrix} = \begin{bmatrix} \frac{1}{10}(\omega_1 - \mu_1) \\ \frac{1}{2}(\omega_2 - \mu_2) \end{bmatrix}$$

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

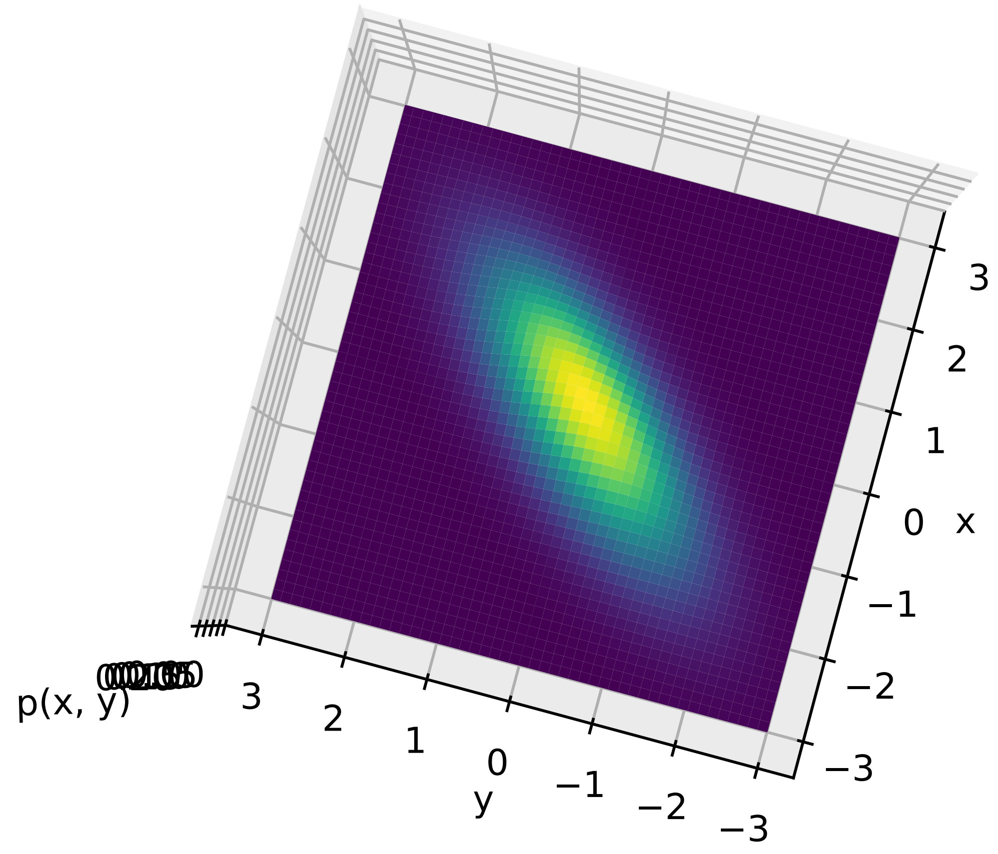
Visually



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



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



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

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

The weighted norm with correlations

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

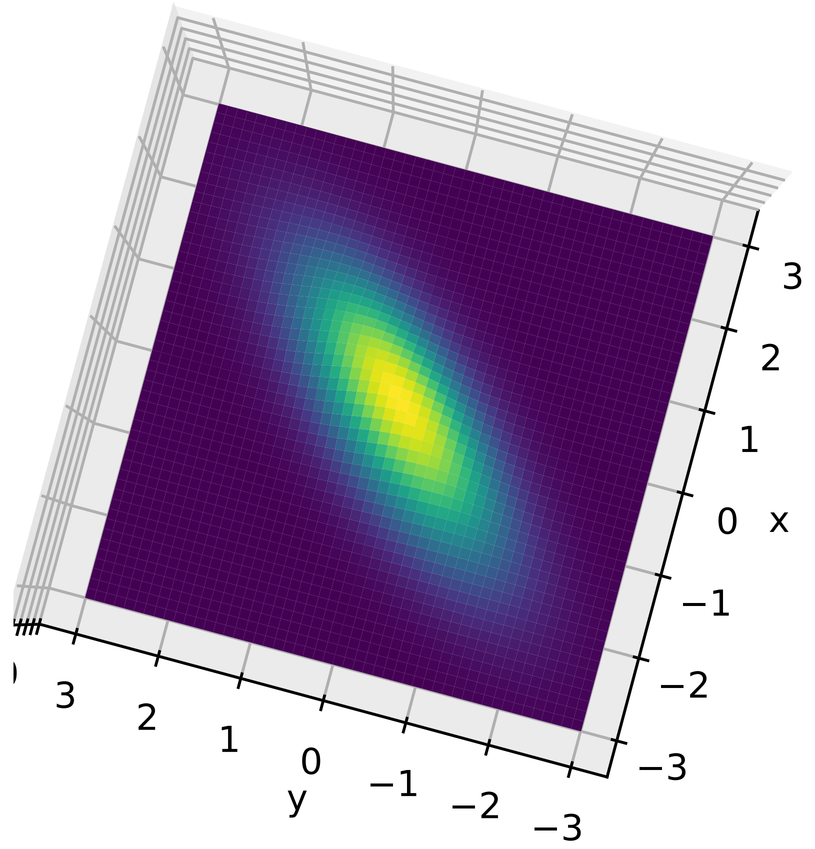
- The weighted norm gives a distance to the mean, for the covariance

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

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

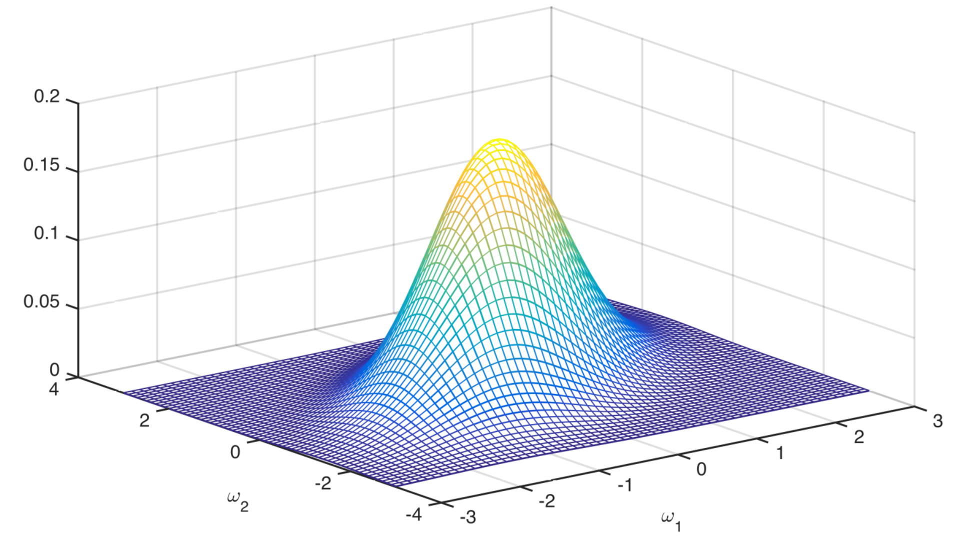
- If e_1 is the opposite sign from e_2, then the distance is larger
(-2.4 * negative number = positive number added to distance)

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



The determinant component

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



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

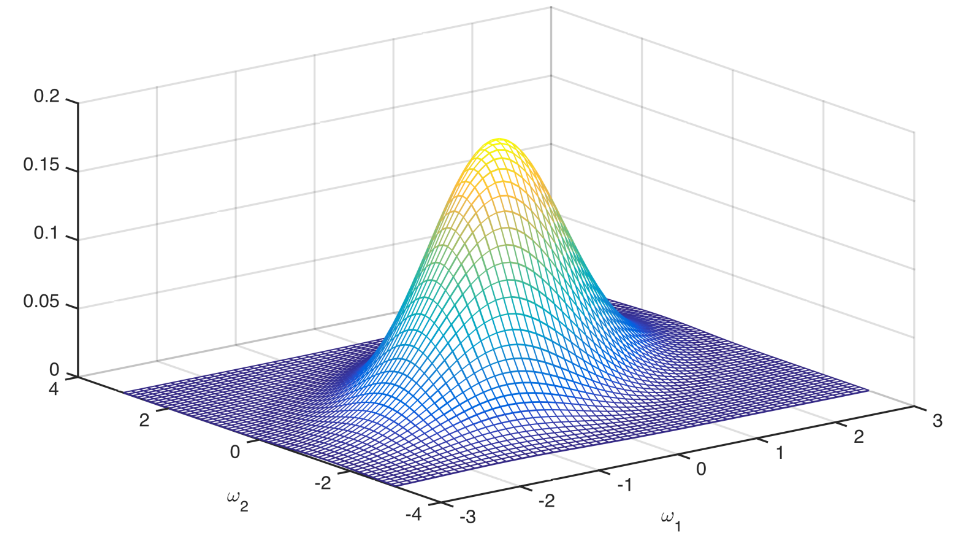
$|\boldsymbol{\Sigma}| = \det(\boldsymbol{\Sigma}) =$ product of singular values

(reflects the magnitude of the covariance)

What is the determinant of this Sigma?

The determinant component

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



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

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

(reflects the magnitude of the covariance)

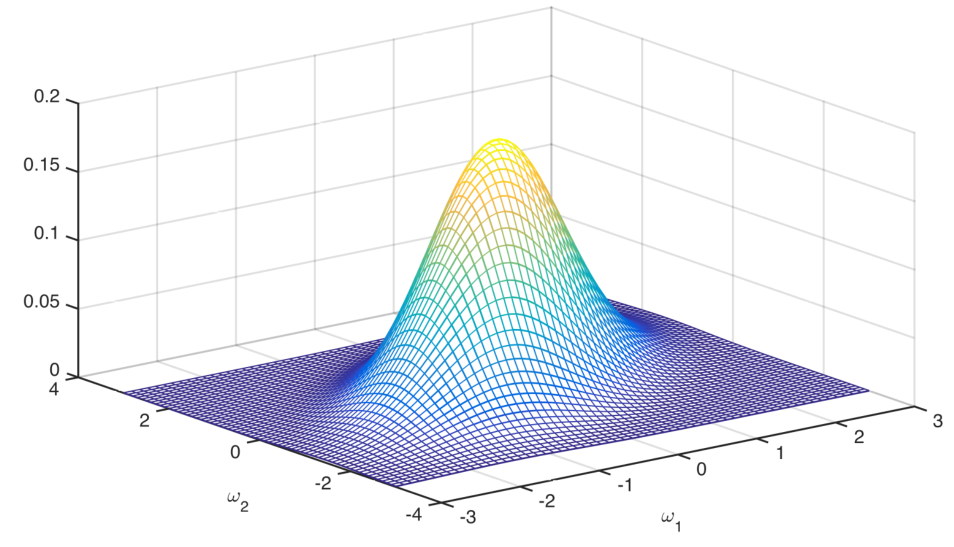
What is the determinant of this other Sigma?

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

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

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?

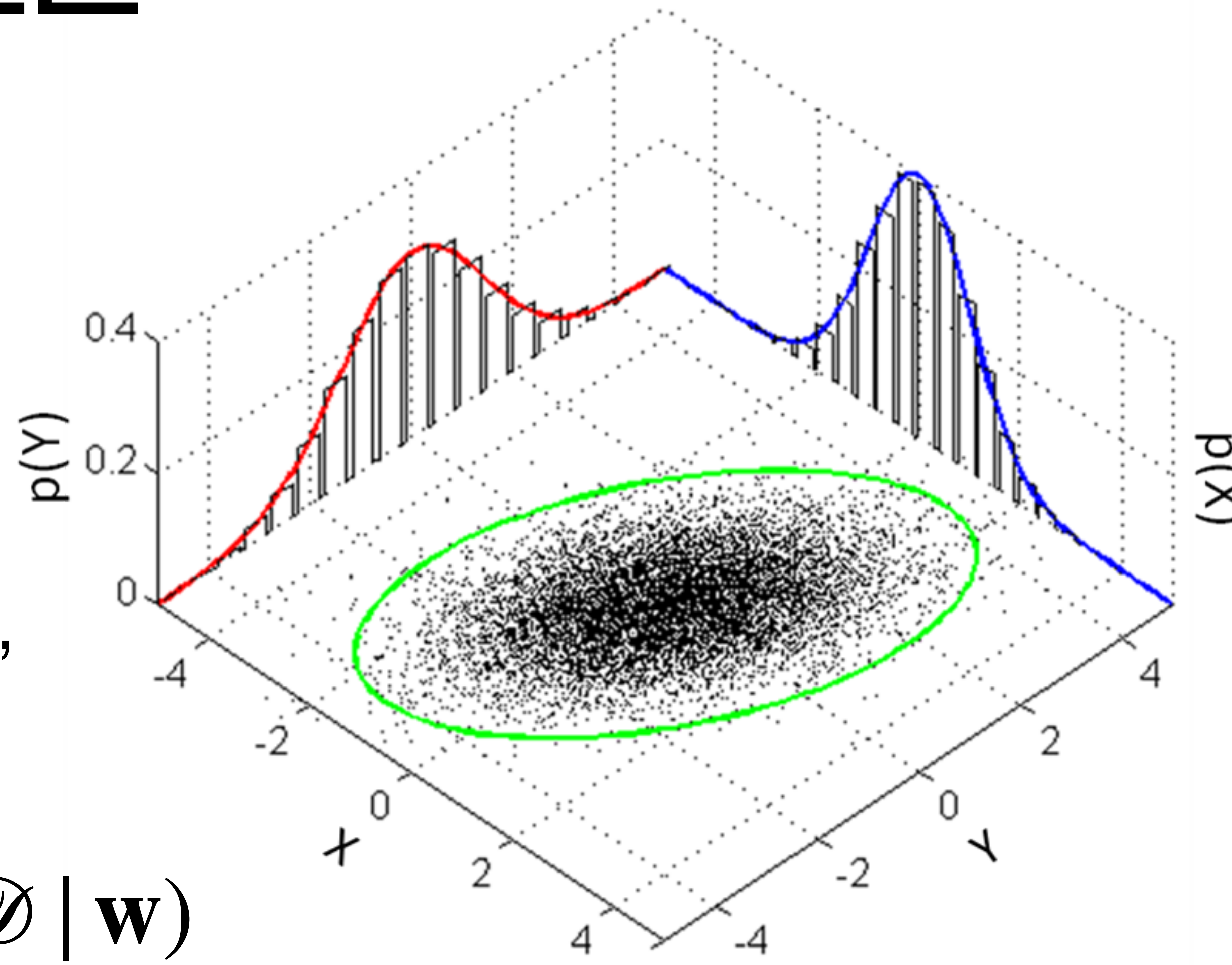
$$\boldsymbol{\Sigma} = \begin{bmatrix} 1.0 & 0.75 \\ 0.75 & 1.0 \end{bmatrix}$$

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

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

Revisiting MLE

- Let us look at MLE for a multivariate Gaussian
- Have a dataset of d-dimensional points $\mathscr{D} = \{\mathbf{x}_i\}_{i=1}^n$
- What is the most likely Gaussian that generated this data, with parameters $\mathbf{w} = (\mu, \mathbf{\Sigma})$?
- Or more precisely, what is the MLE solution $\arg\max_{\mathbf{w}} p(\mathscr{D} \mid \mathbf{w})$
- and what is the MAP solution $\arg\max_{\mathbf{w}} p(\mathbf{w} \mid \mathscr{D})$?



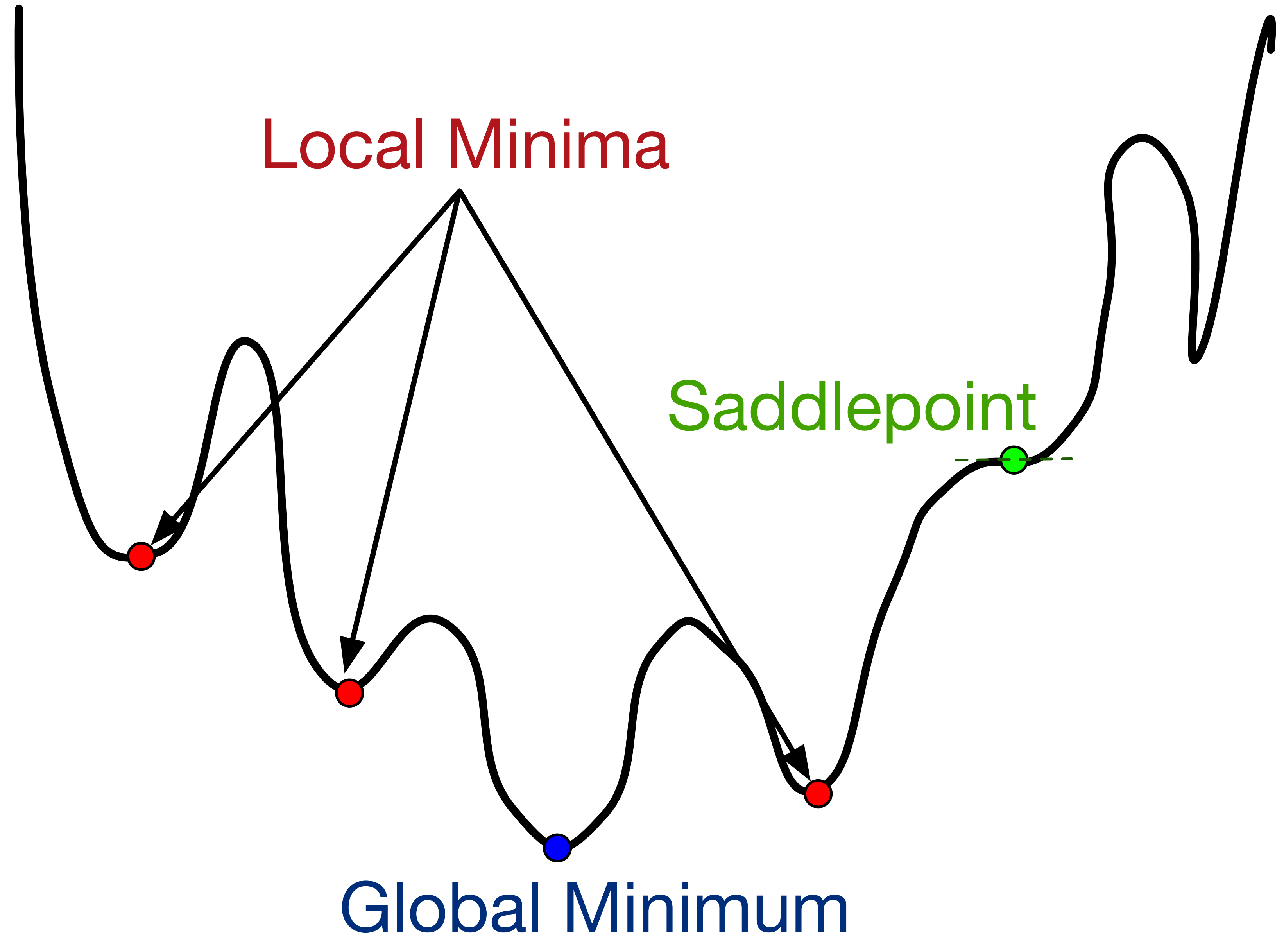
* Image from Wikipedia

Wait, we have a matrix of parameters?

- Gaussian with parameters $\mathbf{w} = (\mu, \boldsymbol{\Sigma})$ means we have $\mathbf{w} = (\mu_1, \mu_2, \ldots, \mu_d, \Sigma_{1,1}, \Sigma_{1,2}, \ldots, \Sigma_{1,d}, \Sigma_{2,1}, \Sigma_{2,2}, \ldots, \Sigma_{2,d}, \ldots, \Sigma_{d,d-1}, \Sigma_{d,d})$
- In other words, we have a vector of parameters of size $d + d^2$
- Our goal is to find $\mathbf{w}$ such that all partial derivatives are zero (at a stationary point)
- Our MLE objective is $-\sum_{i=1}^{n} \ln p(\mathbf{x}_i \mid \mathbf{w})$ so we need $-\frac{\partial}{\partial w_j} \sum_{i=1}^{n} \ln p(\mathbf{x}_i \mid \mathbf{w}) = 0$

Reminder about Stationary Points

- Every minimum of an everywhere-differentiable function $c(w)$ **must** occur at a **stationary point**: A point at which $c'(w) = 0$
- However, not every stationary point is a minimum
- Every stationary point is either:
  - A **local minimum**
  - A **local maximum**
  - A **saddlepoint**

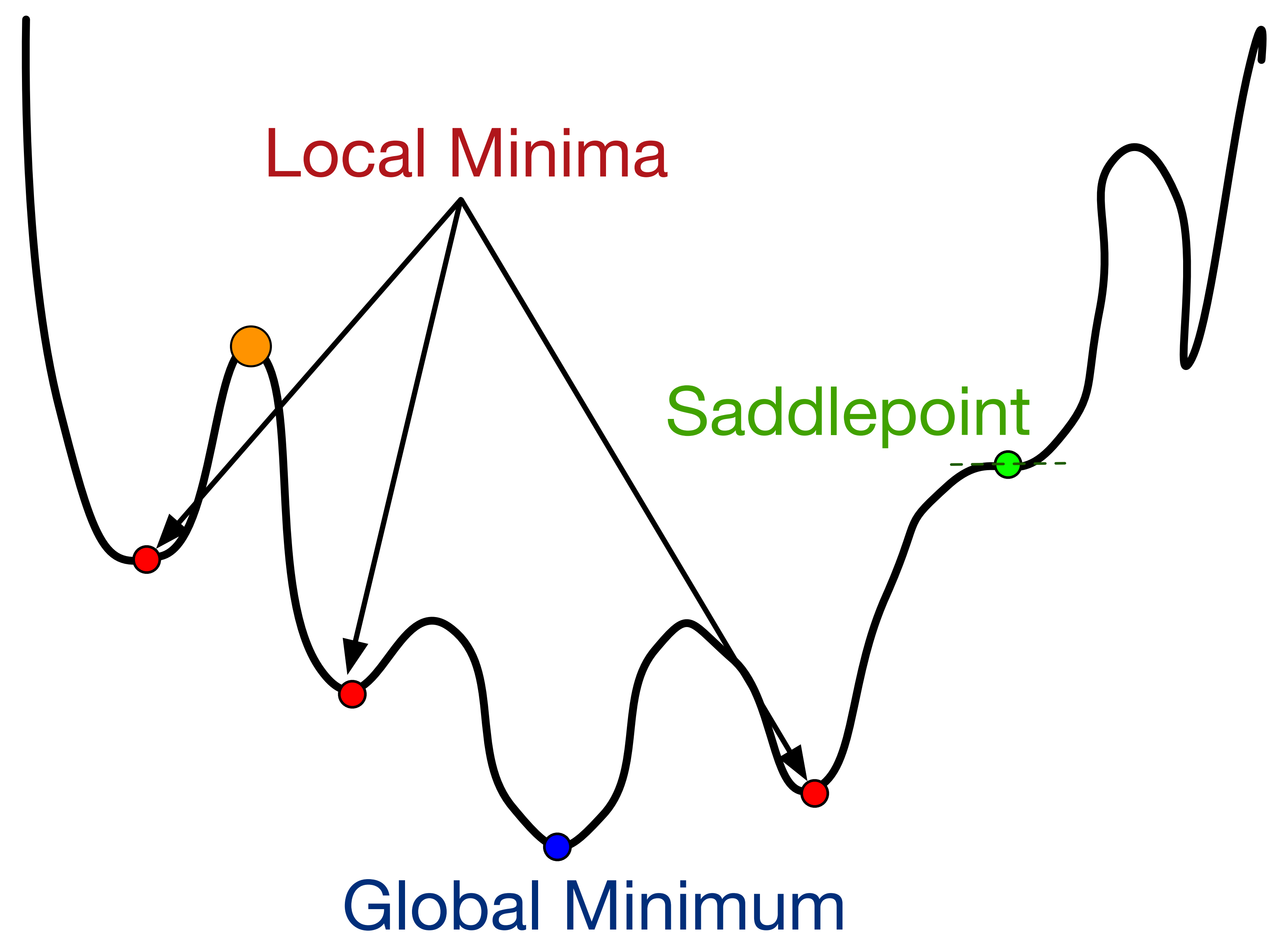


- The **global minimum** is either a local minimum (or a boundary point)

Let's assume for now that w is unconstrained (i.e, $w \in \mathbb{R}$ rather than $w \geq 0$ or $w \in [0,1]$)

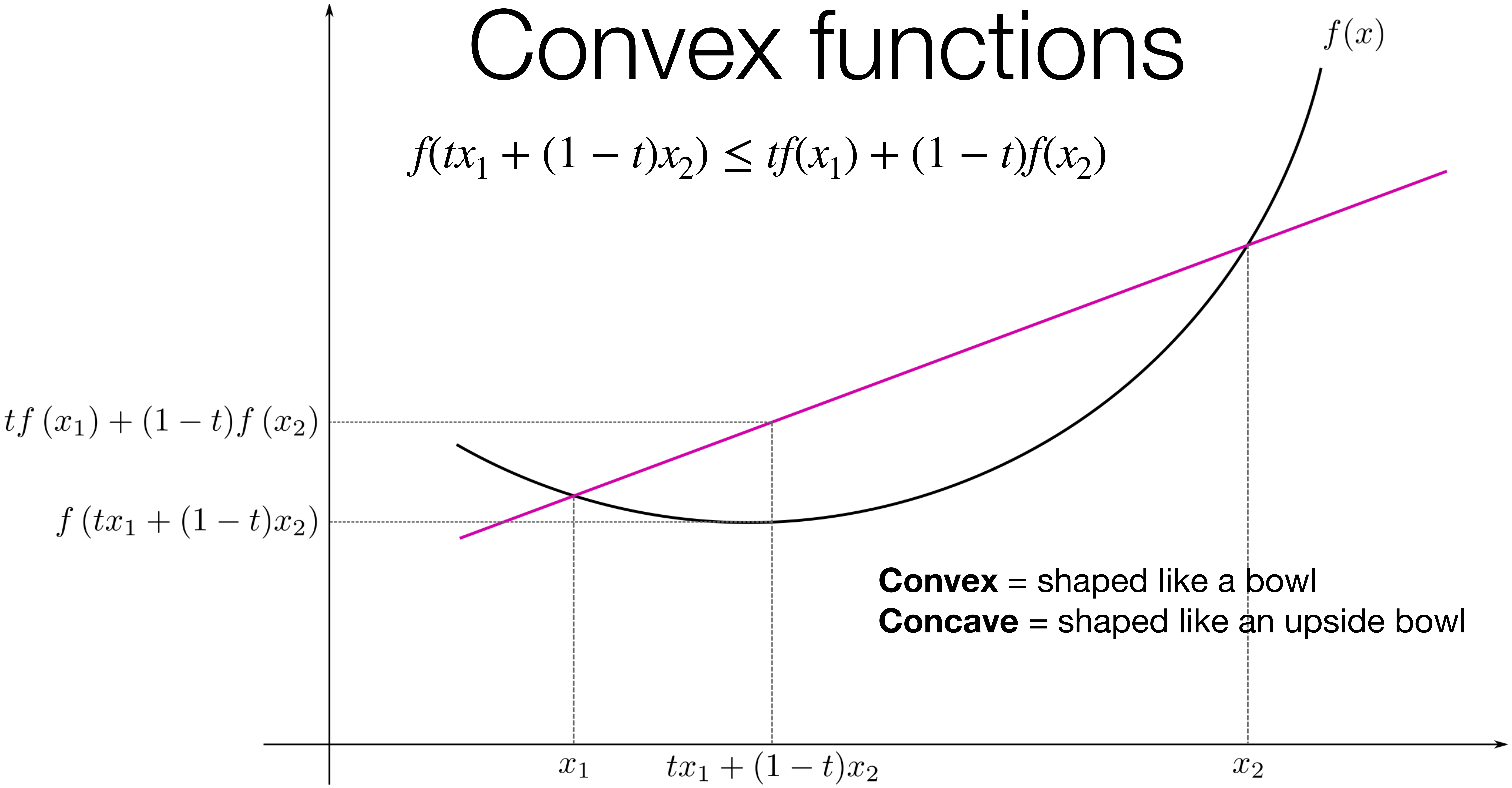
Identifying the type of the stationary point

- If function curved upwards (**convex**) locally, then **local minimum**



Convex functions

$$f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2)$$

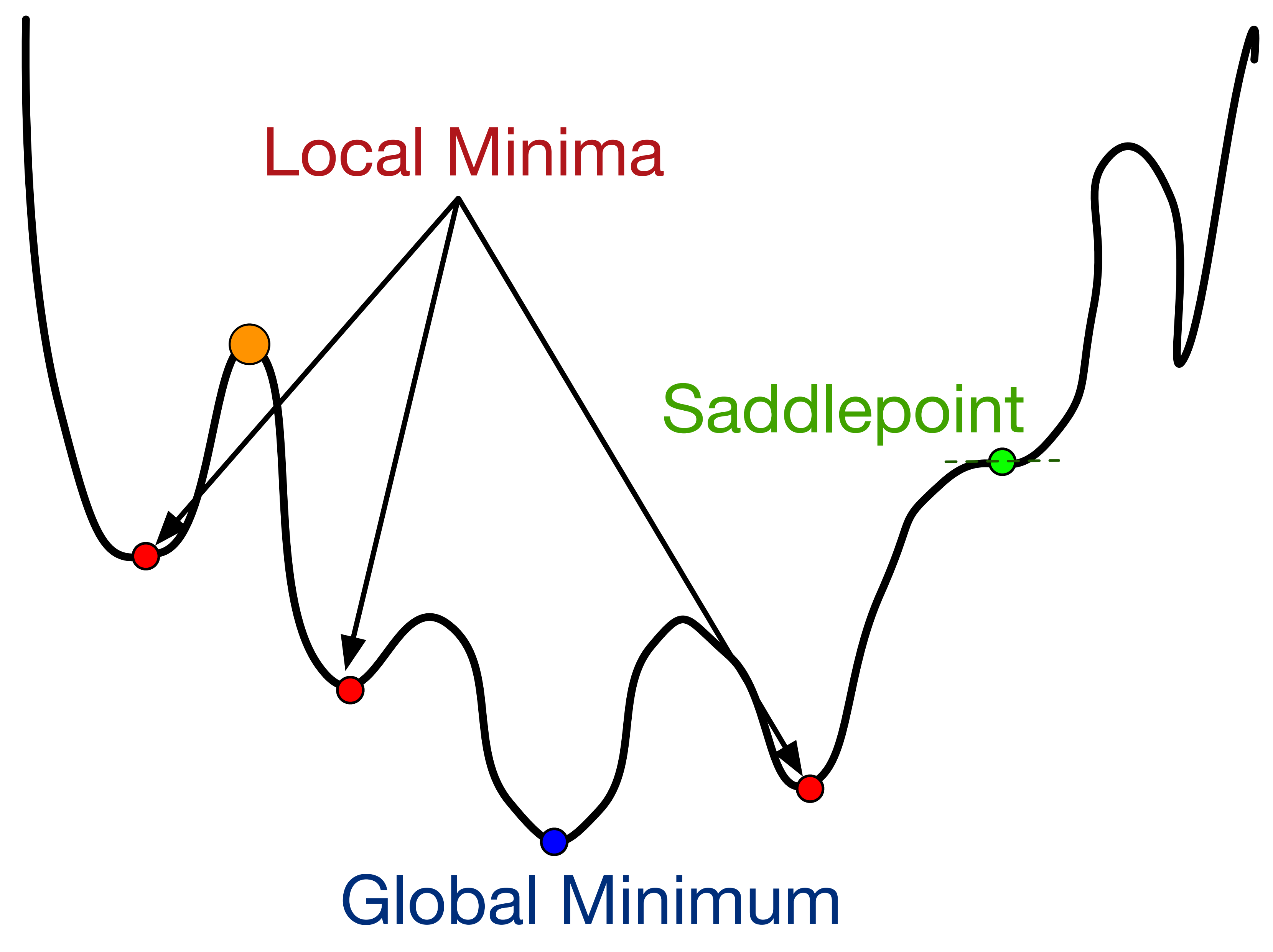


Convex = shaped like a bowl
Concave = shaped like an upside bowl

* from Wikipedia

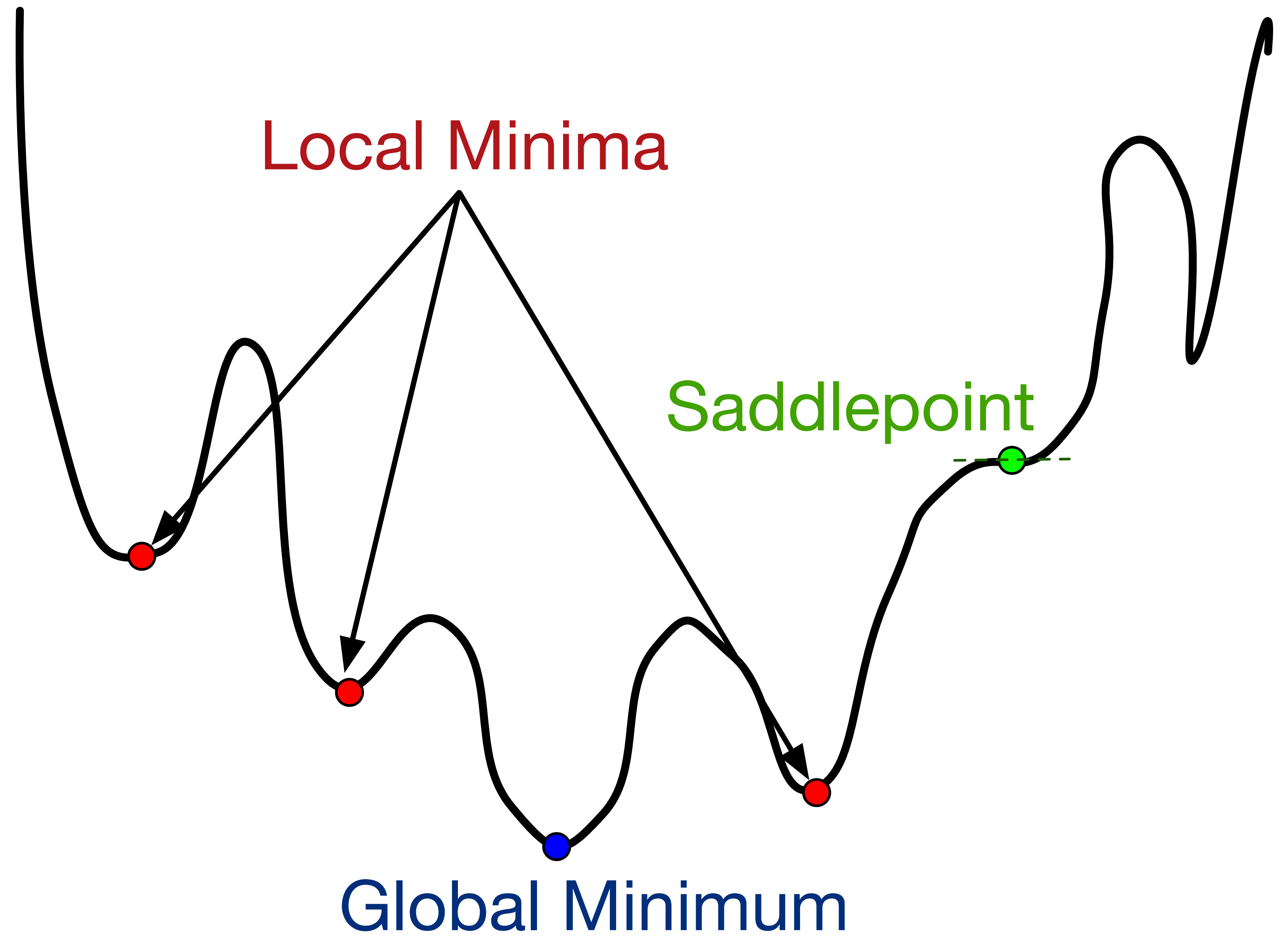
Identifying the type of the stationary point

- If function curved upwards (**convex**) locally, then **local minimum**
- If function curved downwards (**concave**) locally, then **local maximum**
- If function **flat** locally, then might be a **saddlepoint** but could also be a local min or local max
- Locally, cannot distinguish between local min and global min (its a global property of the surface)



Second derivative test

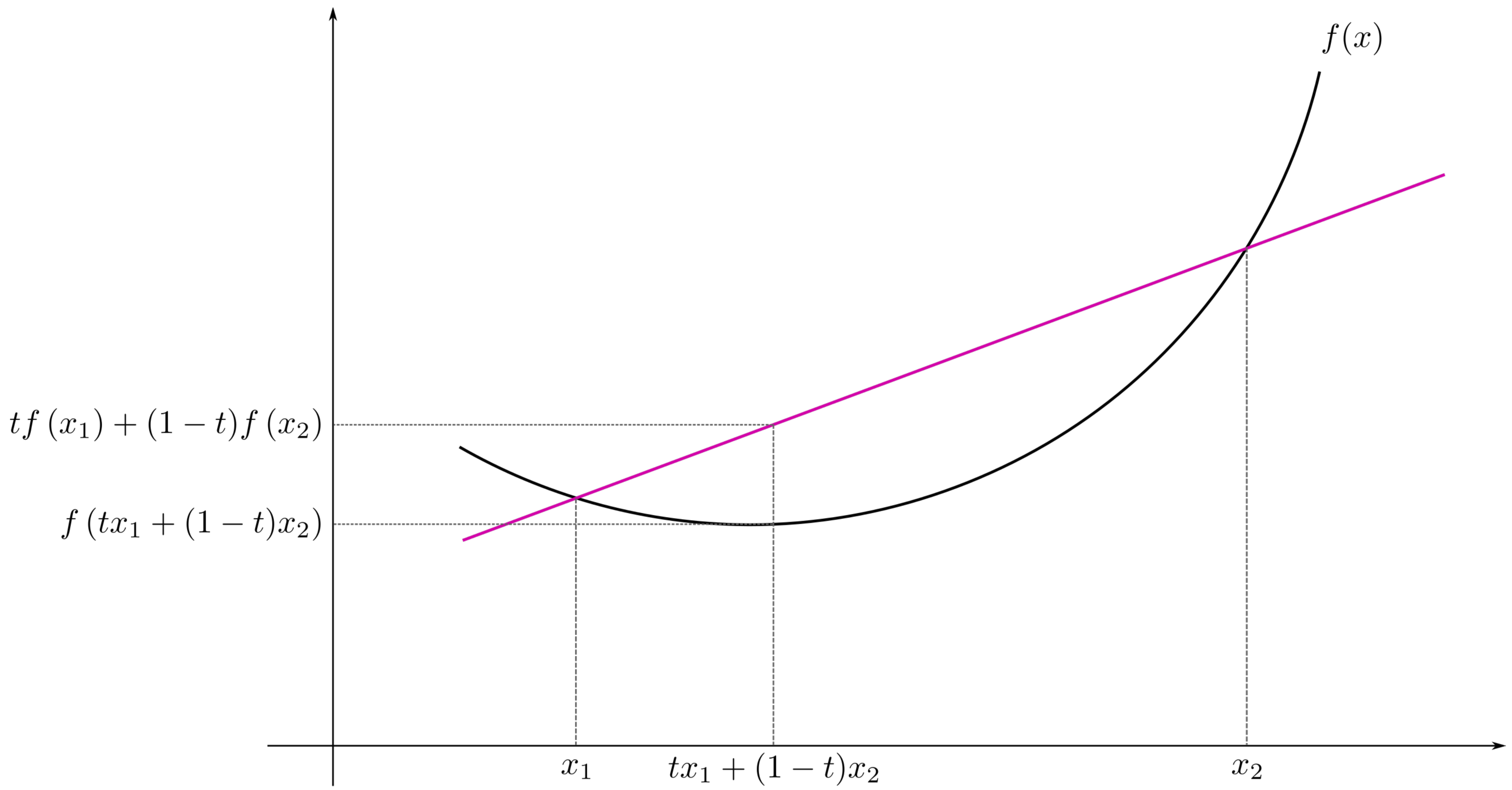
1. If $c''(w_0) > 0$ then w_0 is a local minimum.
2. If $c''(w_0) < 0$ then w_0 is a local maximum.
3. If $c''(w_0) = 0$ then the test is inconclusive: we cannot say which type of stationary point we have and it could be any of the three.



Testing optimality without the second derivative test

Convex functions have a **global** minimum at **every** stationary point

$$c \text{ is convex} \iff c(t\mathbf{w}_1 + (1-t)\mathbf{w}_2) \le tc(\mathbf{w}_1) + (1-t)c(\mathbf{w}_2)$$



Procedure

- Find a stationary point, namely w_0 such that $c'(w_0) = 0$
  - Sometimes we can do this analytically (closed form solution, namely an explicit formula for w_0)
- Reason about if it is optimal
  - Check if your function is convex
  - If you have only one stationary point and it is a local minimum, then it is a global minimum
  - Otherwise, if second derivate test says its a local min, can only say that

Our MLE Objective is Convex with a closed-form solution

- Our MLE objective is $-\sum_{i=1}^{n} \ln p(\mathbf{x}_i \mid \mathbf{w})$ so we need $-\frac{\partial}{\partial w_j} \sum_{i=1}^{n} \ln p(\mathbf{x}_i \mid \mathbf{w}) = 0$
- And $\frac{\partial}{\partial w_j} \sum_{i=1}^{n} \ln p(\mathbf{x}_i \mid \mathbf{w}) = \sum_{i=1}^{n} \frac{\partial}{\partial w_j} \ln p(\mathbf{x}_i \mid \mathbf{w})$
- We can show $-\ln p(\mathbf{x}_i \mid \mathbf{w}) = \frac{d}{2} \ln(2\pi) + \frac{1}{2} \ln |\Sigma| + \frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu})$
- $\frac{\partial}{\partial \mu_1} \ln p(\mathbf{x}_i \mid \mathbf{w}) = 0 + 0 + \frac{\partial}{\partial \mu_1}(\mathbf{x}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}) = \boldsymbol{\Sigma}^{-1}[1, :](\mathbf{x}_i - \mu)$

Given without derivation

First row of $\boldsymbol{\Sigma}^{-1}$

Our MLE Objective is Convex with a closed-form solution

- Our MLE objective is $-\sum_{i=1}^{n} \ln p(\mathbf{x}_i \mid \mathbf{w})$ so we need $-\frac{\partial}{\partial w_j} \sum_{i=1}^{n} \ln p(\mathbf{x}_i \mid \mathbf{w}) = 0$
- And $\frac{\partial}{\partial w_j} \sum_{i=1}^{n} \ln p(\mathbf{x}_i \mid \mathbf{w}) = \sum_{i=1}^{n} \frac{\partial}{\partial w_j} \ln p(\mathbf{x}_i \mid \mathbf{w})$
- We can show $-\ln p(\mathbf{x}_i \mid \mathbf{w}) = \frac{d}{2} \ln(2\pi) + \frac{1}{2} \ln |\Sigma| + \frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu})$
- More simply we can write $\frac{\partial}{\partial \boldsymbol{\mu}} \ln p(\mathbf{x}_i \mid \mathbf{w}) = \mathbf{0} + \mathbf{0} + \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu})$

Our MLE Objective is Convex with a closed-form solution

- Our MLE objective is $-\sum_{i=1}^{n} \ln p(\mathbf{x}_i \mid \mathbf{w})$ so we need $-\frac{\partial}{\partial w_j} \sum_{i=1}^{n} \ln p(\mathbf{x}_i \mid \mathbf{w}) = 0$
- $\frac{\partial}{\partial \boldsymbol{\mu}} - \ln p(\mathbf{x}_i \mid \mathbf{w}) = \mathbf{0} + \mathbf{0} + \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}) \qquad \in \mathbb{R}^d$
- Sp $-\frac{\partial}{\partial \boldsymbol{\mu}} \sum_{i=1}^{n} \ln p(\mathbf{x}_i \mid \mathbf{w}) = \sum_{i=1}^{n} \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}) = \boldsymbol{\Sigma}^{-1} \sum_{i=1}^{n} (\mathbf{x}_i - \boldsymbol{\mu}) = \mathbf{0}$
- which occurs if and only if $\sum_{i=1}^{n} \mathbf{x}_i - \boldsymbol{\mu} = \mathbf{0}$, giving us $\boldsymbol{\mu} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i$

Our MLE Objective is Convex with a closed-form solution

- Our MLE objective is $-\sum_{i=1}^{n} \ln p(\mathbf{x}_i \mid \mathbf{w})$ so we need $-\frac{\partial}{\partial w_j} \sum_{i=1}^{n} \ln p(\mathbf{x}_i \mid \mathbf{w}) = 0$
- $\frac{\partial}{\partial \boldsymbol{\mu}} - \ln p(\mathbf{x}_i \mid \mathbf{w}) = \mathbf{0} + \mathbf{0} + \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}) \qquad \in \mathbb{R}^d$
- Sp $-\frac{\partial}{\partial \boldsymbol{\mu}} \sum_{i=1}^{n} \ln p(\mathbf{x}_i \mid \mathbf{w}) = \mathbf{0}$ gives $\boldsymbol{\mu} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i$ (sample mean)
- And $-\frac{\partial}{\partial \boldsymbol{\Sigma}} \sum_{i=1}^{n} \ln p(\mathbf{x}_i \mid \mathbf{w}) = \mathbf{0}$ gives $\boldsymbol{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top$ (sample covariance)

What about the MAP objective?

- Now we have to select a prior on $\mathbf{w} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$. What prior might we pick?
- Can pick a zero-mean Gaussian on $\boldsymbol{\mu}$, with variance indicating how big it can be
- But more complicated for covariance $\boldsymbol{\Sigma}$, because constrained to be positive definite
  - There are such distributions but goes beyond what you need to know for this course
- Once we pick a prior, the steps are similar to MLE
- **Q1:** Intuitively, is there any information you might a priori put on the covariance? What if you know dimensions 1 and 2 are independent variables? Or know they are dependent?
- **Q2:** Why might it help to add a prior?

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

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

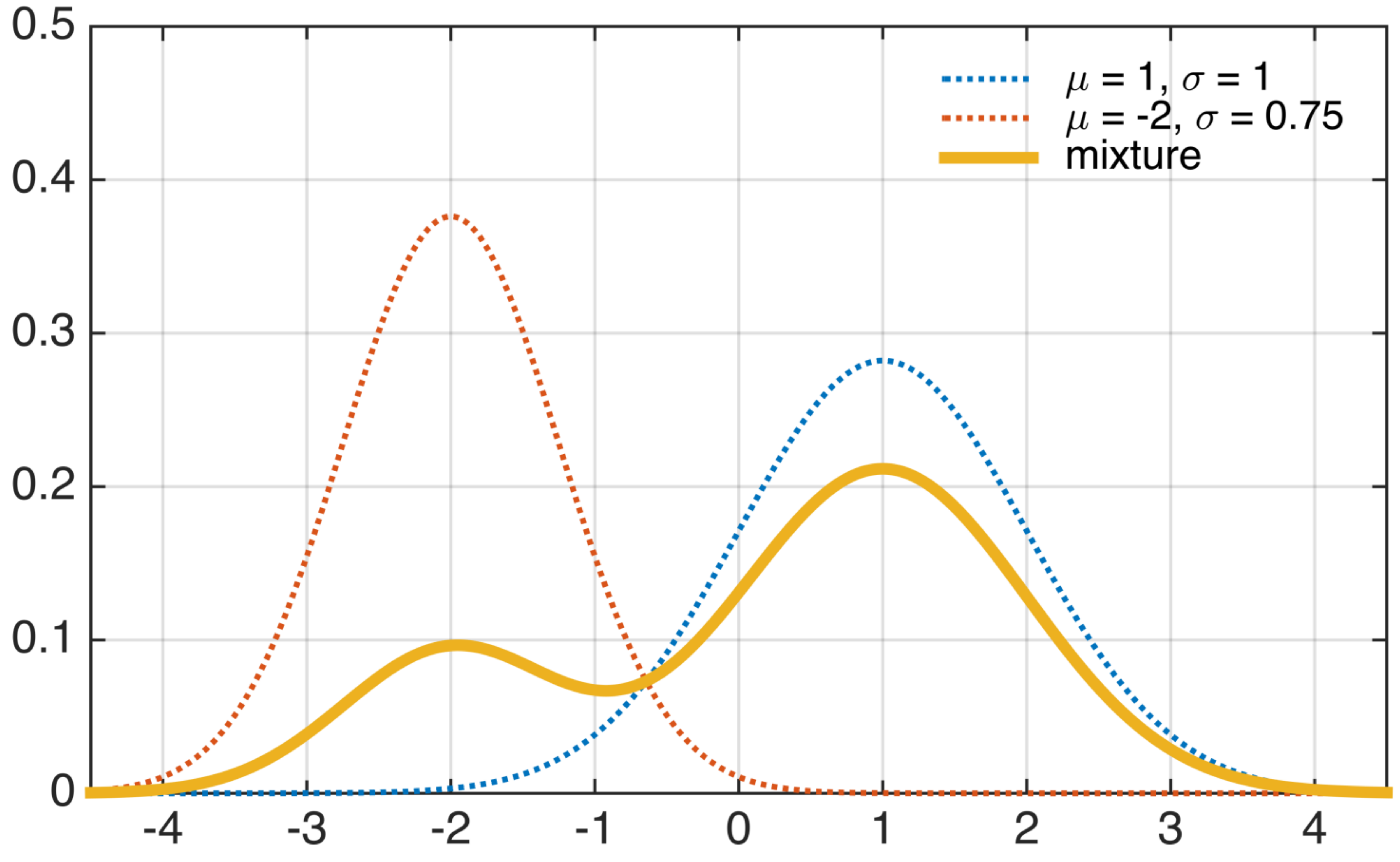
Mixture of $m = 2$ Gaussian distributions:

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

Question: What are the parameters of the distribution p?

p is defined by vector of parameters

$\boldsymbol{\theta} = (w_1, w_2, \mu_1, \mu_2, \sigma_1, \sigma_2)$



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 numbers is nonnegative

Exercise Solution for PMFs

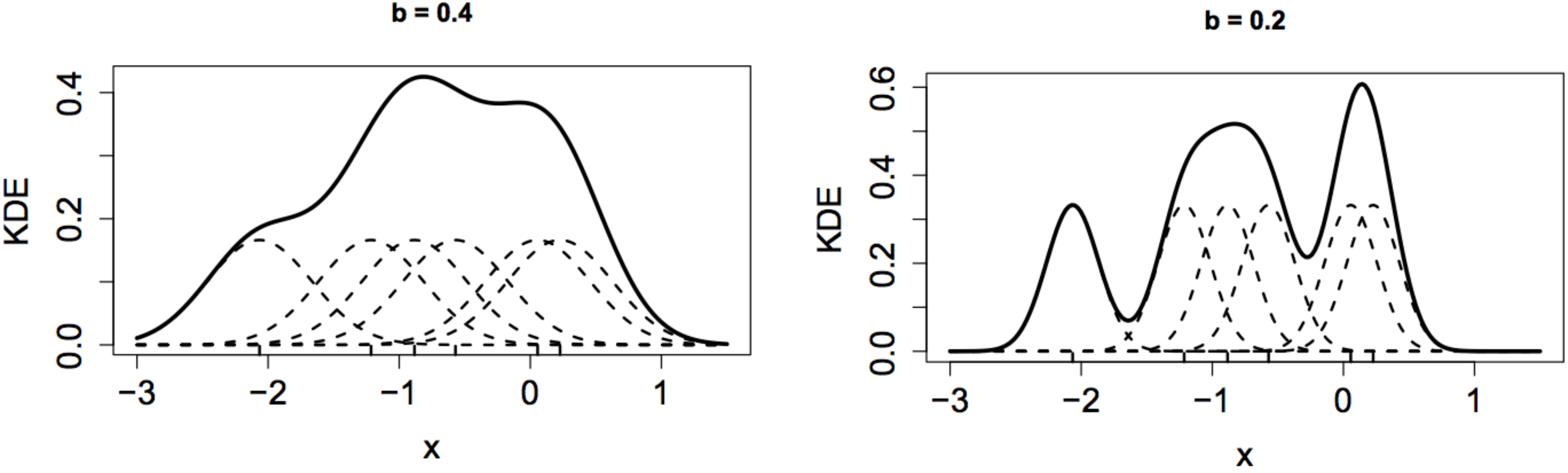
$$
\begin{aligned}
\sum_{x \in \mathcal{X}} p(x) &= \sum_{x \in \mathcal{X}} \sum_{i=1}^{m} w_i p_i(x) \\
&= \sum_{i=1}^{m} \sum_{x \in \mathcal{X}} w_i p_i(x) \\
&= \sum_{i=1}^{m} w_i \underbrace{\sum_{x \in \mathcal{X}} p_i(x)}_{=1} \\
&= \sum_{i=1}^{m} w_i = 1
\end{aligned}
$$

Exercise Solution for PDFs

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

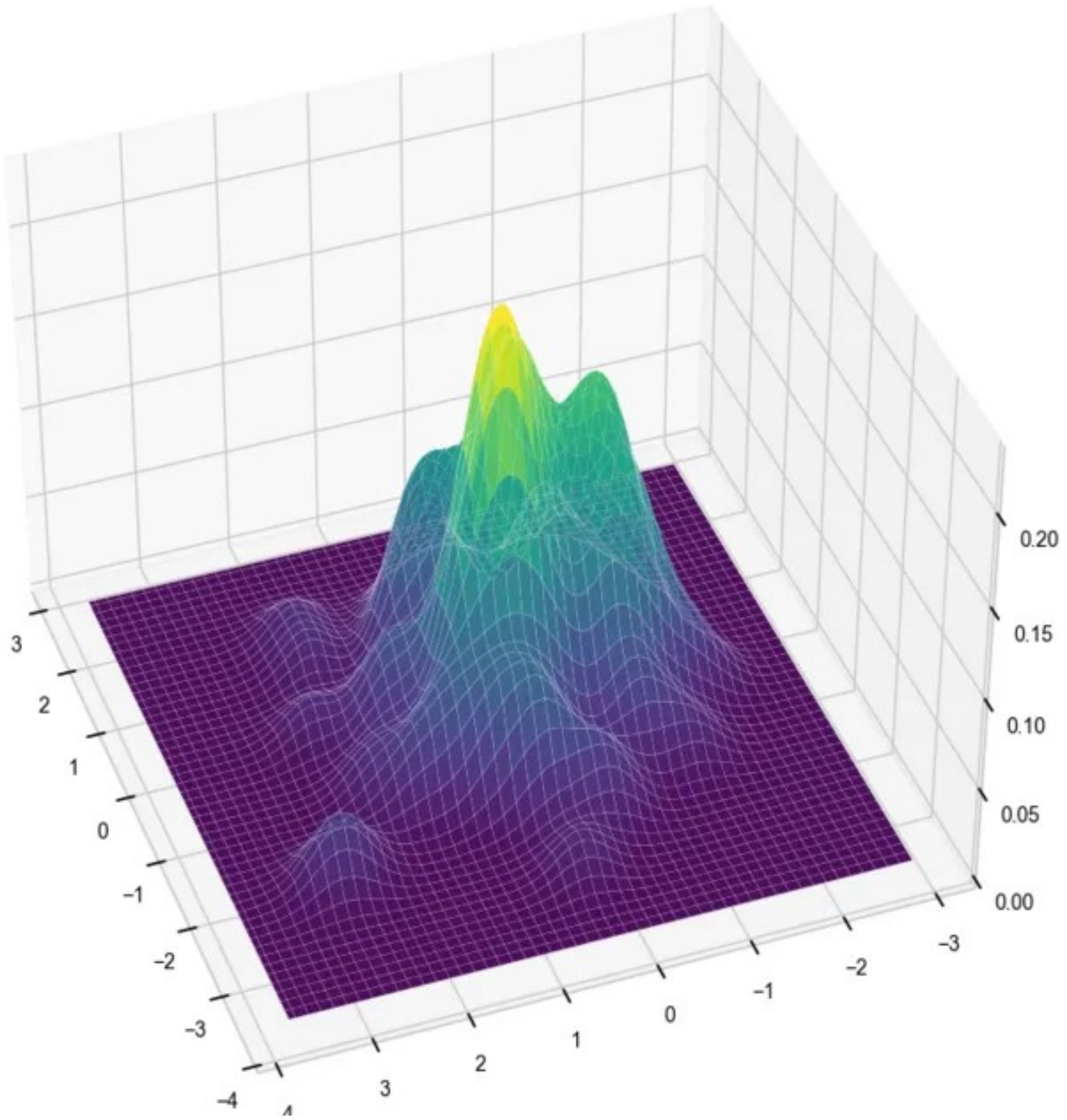
$$
\begin{aligned}
\int_{\mathcal{X}} p(x) dx &= \int_{\mathcal{X}} \sum_{i=1}^{m} w_i p_i(x) dx \\
&= \sum_{i=1}^{m} \int_{\mathcal{X}} w_i p_i(x) dx \\
&= \sum_{i=1}^{m} w_i \underbrace{\int_{\mathcal{X}} p_i(x) dx}_{=1} \\
&= \sum_{i=1}^{m} w_i = 1
\end{aligned}
$$

Mixture Can Produce Complex Distributions



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

And multivariate mixtures too



* Image from https://towardsdatascience.com/the-math-behind-kernel-density-estimation-5deca75cba38

Parameters for multivariate mixture

- What if we wanted a mixture of 5 components for a multivariate RV of dimension d?
- Then we can have a mixture over multivariate Gaussians of dimension d
- The parameters are $\boldsymbol{\theta} = (w_1, w_2, w_3, w_4, w_5, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\mu}_3, \boldsymbol{\mu}_4, \boldsymbol{\mu}_5, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \boldsymbol{\Sigma}_3, \boldsymbol{\Sigma}_4, \boldsymbol{\Sigma}_5)$

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?

Ordered, discrete targets

- Imagine we have a dataset of pairs $(\mathbf{x}, y)$ where $\mathbf{x}$ are features about a call center and y are the number of calls received in one hour. We have $y \in \{0,1,2,3,\ldots\}$
- We can model this using a Poisson distribution
- Recall the PMF for a Poisson $p(y) = \lambda^y \exp(-\lambda)/y!$

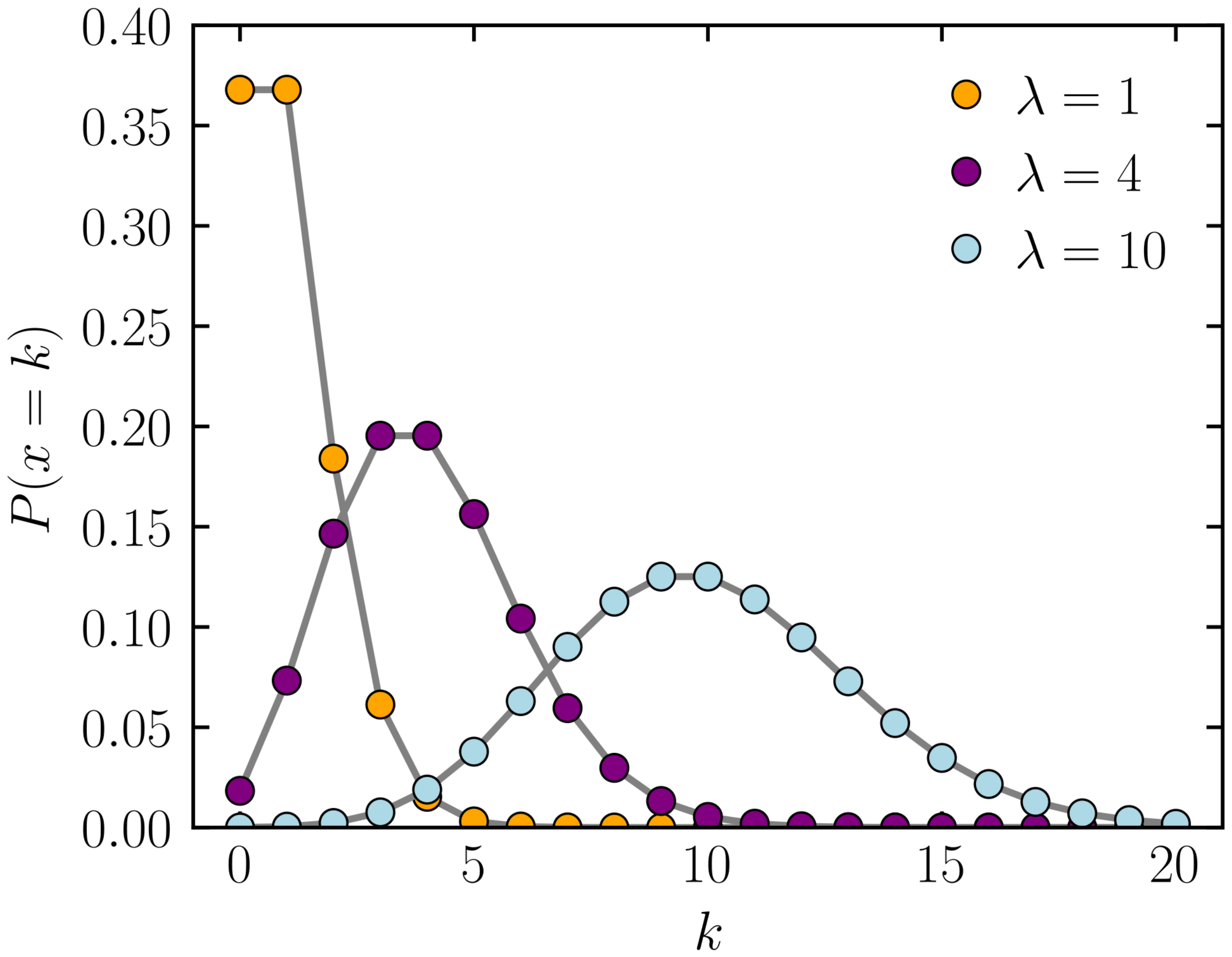
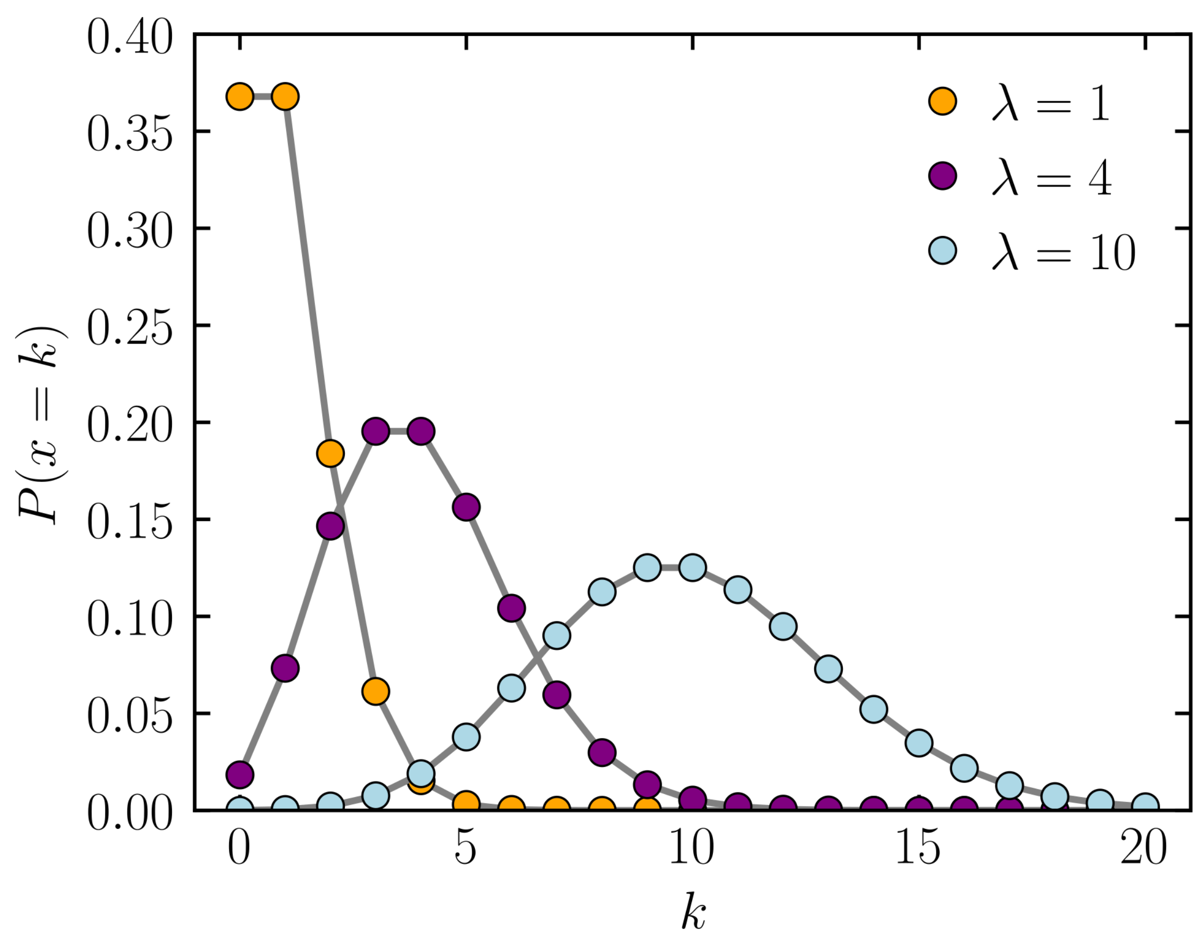


*Image from Wikipedia

Ordered, discrete targets

- Imagine we have a dataset of pairs $(\mathbf{x}, y)$ where $\mathbf{x}$ are features about a call center and y are the number of calls received in one hour. We have $y \in \{0,1,2,3,\ldots\}$
- We can model this using a conditional Poisson distribution $p(y \mid \mathbf{x}) = \lambda(\mathbf{x})^y \exp(-\lambda(x))/y!$
- Why would we choose to do this instead of using a categorical? How would you use a categorical?



* Later we'll see Poisson regression

Contrasting Poisson & Categorical

Poisson

$\boldsymbol{\phi}(\mathbf{x})$

$\mathbf{x} \in \mathbb{R}^d$ → $\lambda(\mathbf{x})$

Categorical

Need to assume a max number of calls,
Say 100

$\boldsymbol{\phi}(\mathbf{x})$

$\mathbf{x} \in \mathbb{R}^d$ →

$\alpha_1(\mathbf{x}) = \hat{p}(y = (1,0,\ldots,0) \,|\, \mathbf{x})$

$\alpha_2(\mathbf{x}) = \hat{p}(y = (0,1,0,\ldots,0) \,|\, \mathbf{x})$

.
.
.

$\alpha_{100}(\mathbf{x}) = \hat{p}(y = (0,0,\ldots,0,1) \,|\, \mathbf{x})$

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

- Uncorrelated RVs (i.e., 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:** $X \sim$ 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 $\mathbf{X}$, $H(\mathbf{X}) \leq \frac{d}{2}(\ln 2\pi + 1 + \ln \det \mathbf{\Sigma})$
  - 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

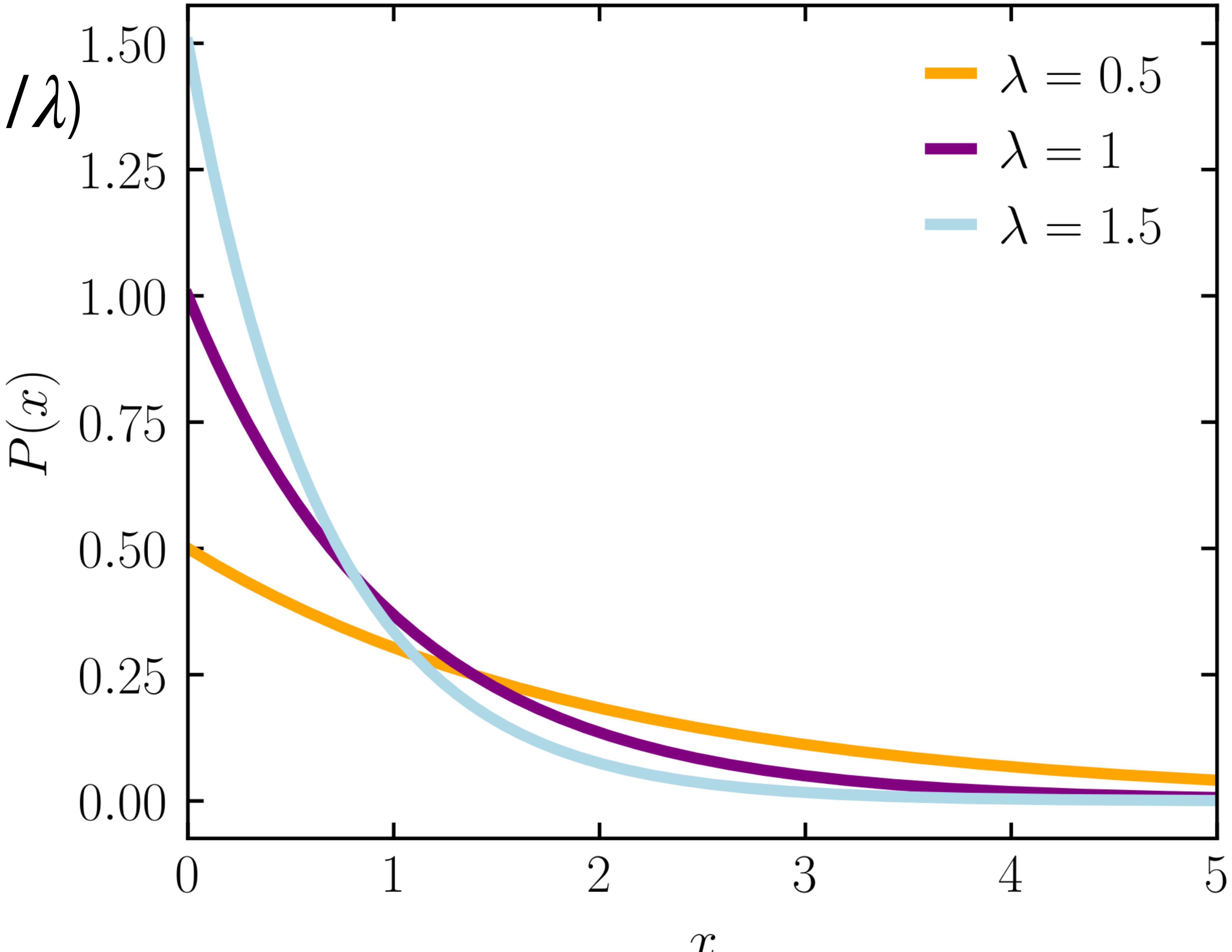
entropy = $-ln0.5 + 1 \approx 1.7$

variance = $1/0.5^2 = 4$

lambda = 1.5

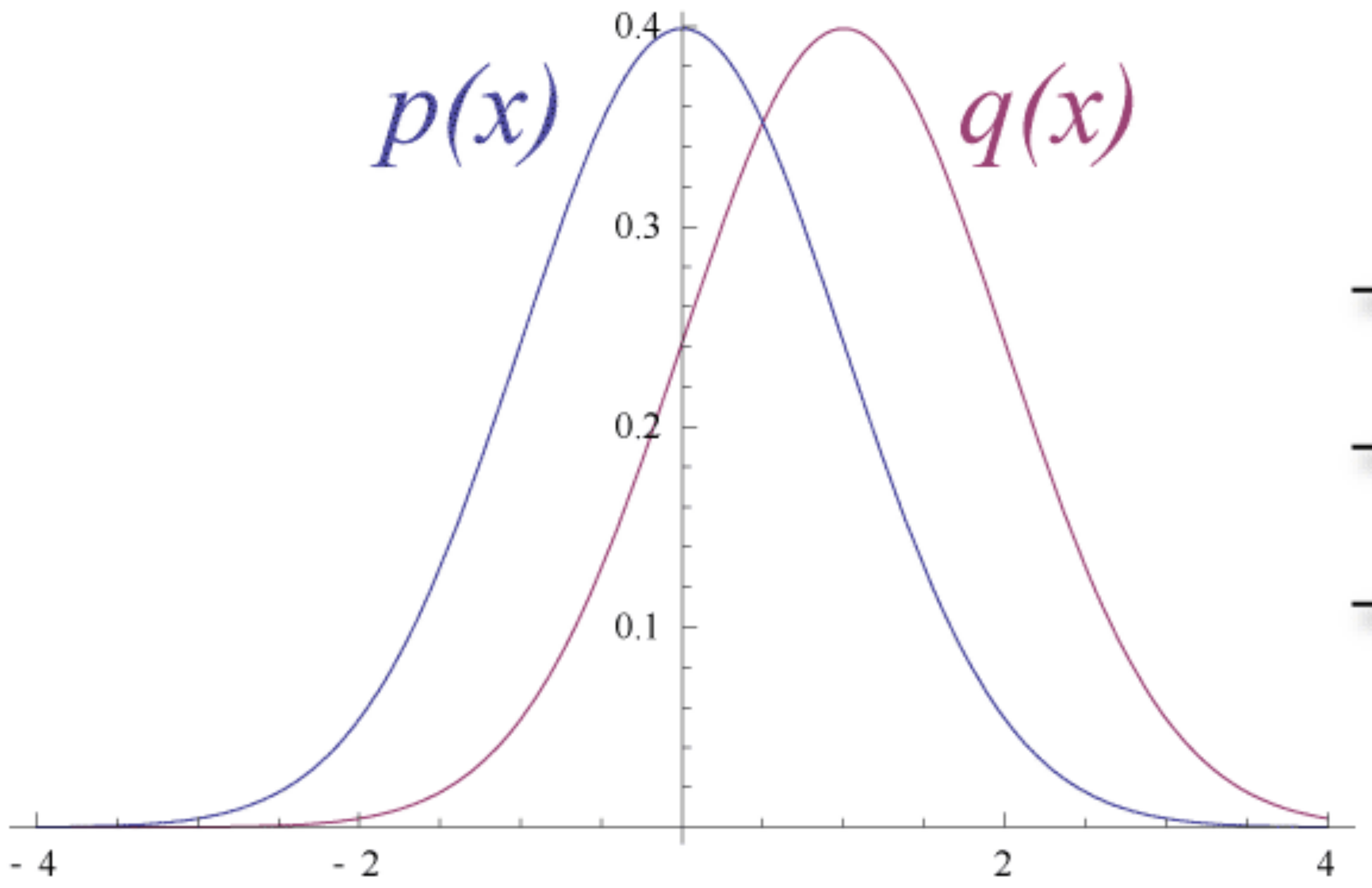
entropy = $-ln1.5 + 1 \approx 0.6$

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

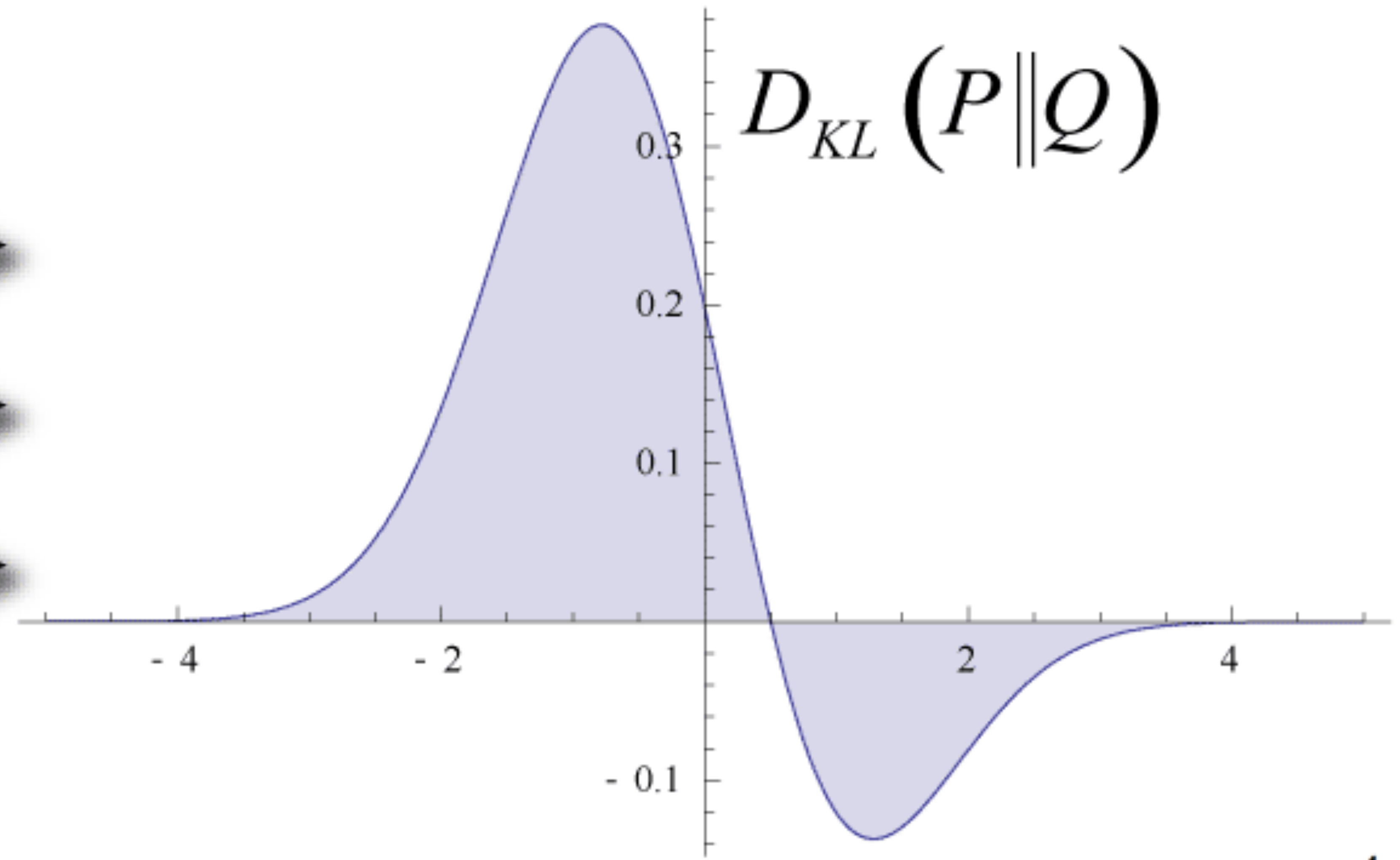


KL Divergence

* Images from Wikipedia



Original Gaussian PDF's



KL Area to be Integrated

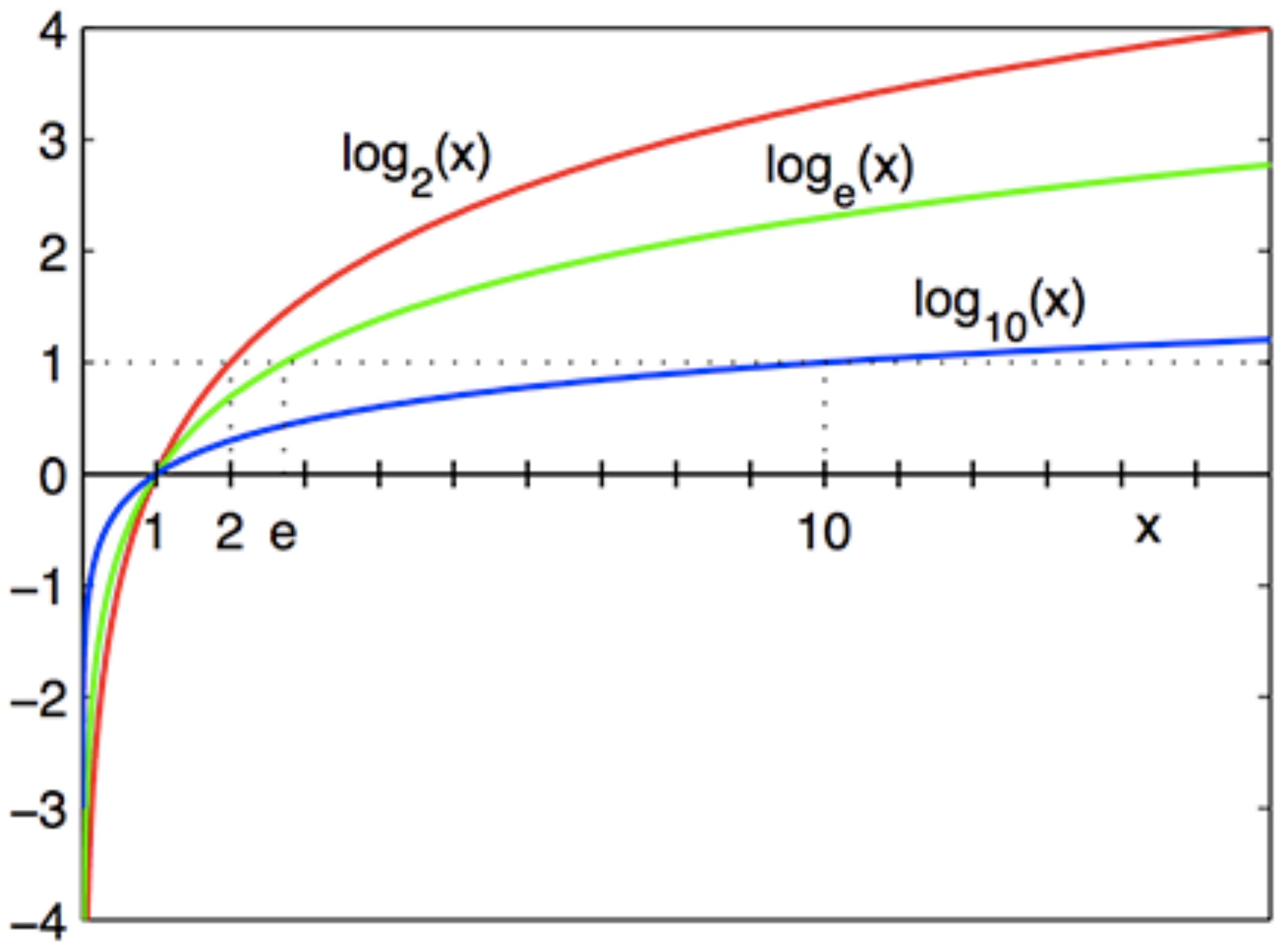
$$\mathrm{KL}(p||q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}$$

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 $D_{\mathrm{KL}}(p||q) \geq 0$ and
- $D_{\mathrm{KL}}(p||q) = 0$ if and only if (iff) $p = q$



Revisiting Our Example

- **Example:** $X \sim$ Uniform$\{-2, -1, 0, 1, 2\}$, $Y = X^2$
  - $\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0 - 0\mathbb{E}[Y] = 0$
- $\mathscr{X} = \{-2, -1, 0, 1, 2\}$ and $\mathscr{Y} = \{0, 1, 4\}$
- $p(x, y) = 0$ 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$(p || p_x p_y) = \sum_{(x,y) \in \mathscr{X} \times \mathscr{Y}} p(x, y) \log \frac{p(x, y)}{p_x(x) p_y(y)}$

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 $p_y(0) = 1/5, p_y(1) = 2/5, p_y(4) = 2/5$
- $$\begin{aligned} \text{KL}(p \,||\, p_x p_y) &= \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} p(x, y) \log \frac{p(x, y)}{p_x(x) p_y(y)} \\ &= \sum_{x \in \mathcal{X}, y = x^2} \frac{1}{5} \log \frac{1/5}{1/5 p_y(y)} \\ &= \frac{1}{5} \sum_{x \in \mathcal{X}, y = x^2} \log \frac{1}{p_y(y)} \\ &= \frac{1}{5} [\log \frac{1}{1/5} + 4 \log \frac{1}{2/5}] = \frac{1}{5} [\log 5 + 4 \log \frac{5}{2}] \approx 1.05 \neq 0 \end{aligned}$$

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 (bigger n) we get $\frac{1}{n}\sum_{i=1}^{n} - \ln p_\theta(x_i) \approx - \mathbb{E}[\ln p_\theta(X)]$ and so closer to minimizing the KL to the true distribution

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
- $\arg\min_\theta D_{\text{KL}}(p_0 || p_\theta) = \arg\min_\theta - \mathbb{E}[\ln p_\theta(X)]$
- **Question1**: Imagine our class of models are Gaussian, $\theta = (\mu, \sigma^2)$, and the true distribution is Gaussian. Is there a p_θ that can get zero $D_{\text{KL}}(p_0 || p_\theta)$?
- **Question2**: What if our class of models are Gaussian, but p_θ is a mixture model?