Vegan fleas, movie ratings, and the EM algorithm

Carlos Cotrini

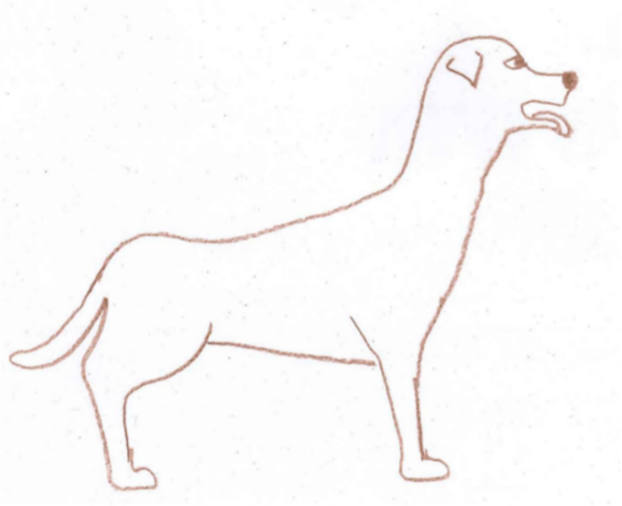
Department of Computer Science
ETH Zürich *ccarlos@inf.ethz.ch*

March 25, 2019

Overview

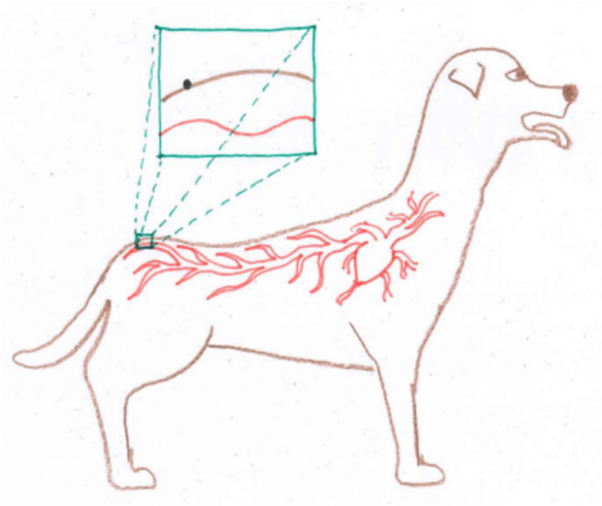
The vegan-flea optimization problem

A two-dimensional dog

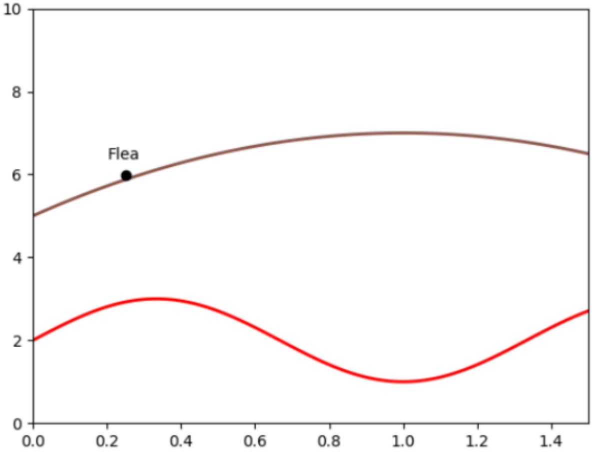


The dog's cardiovascular system

The dog's cardiovascular system



The flea, the dog's skin, and the vessel's upper border



Animation

Formalization

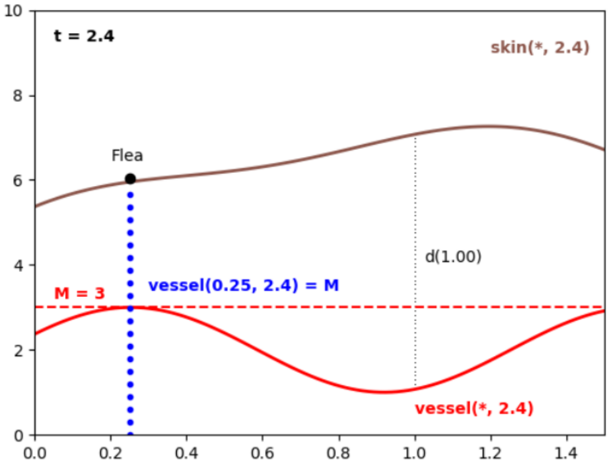
Assumptions

- We assume that for any $x \in [0,1]$ and any two time points $t_1, t_2 \in [0,\infty)$, $skin(x,t_1) - vessel(x,t_1) = skin(x,t_2) - vessel(x,t_2)$.
- For any $x \in [0,1]$ and any $t \in [0,\infty)$, there is $t' \geq t$ such that $vessel(x,t')$ is a maximum of $vessel(\cdot,t')$.
- For any $t \in [0,\infty)$, the flea can efficiently compute a point x^* that maximizes $skin(\cdot,t)$.
- For any $x \in [0,1]$ and any $t \in [0,\infty)$, the flea can efficiently compute $\hat{t} \geq t$ such that $vessel(x,\hat{t})$ is a maximum of $vessel(\cdot,\hat{t})$.

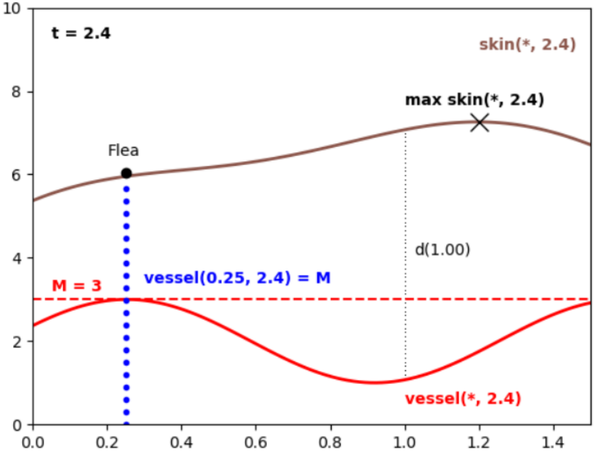
Objective

Can the flea compute x^* such that $d(x^*) \geq d(x_0)$, where x_0 is the flea's current position?

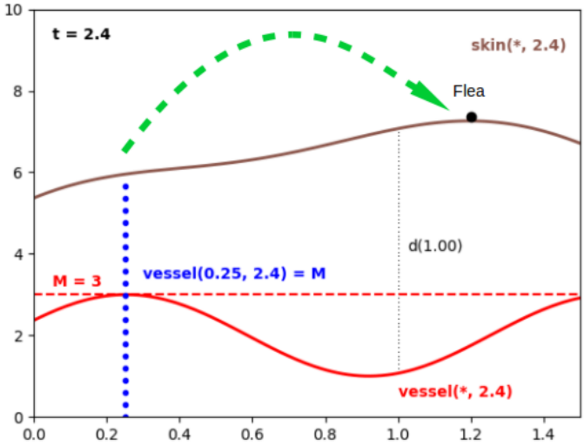
Optimization algorithm



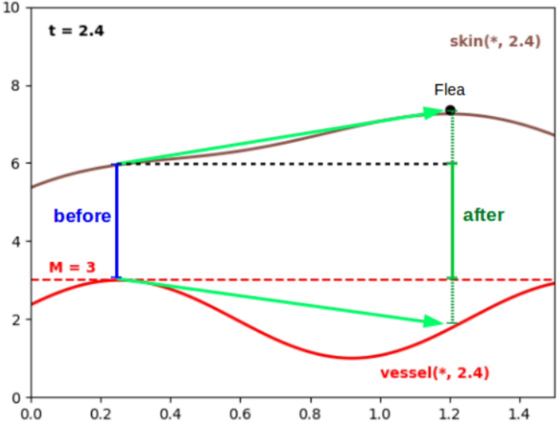
Optimization algorithm



Optimization algorithm



Why does this work?



A movie recommendation system

A simple dataset of movie ratings

	Star wars	Star trek	Titanic	Pretty Woman	007	Mission Impossible
Alice	✓	✓	✗	✗	✗	✗
Bob	✓	✓	✗	✗	✗	✗
Carlos	✗	✗	✓	✓	✗	✗
David	✗	✗	✓	✓	✗	✗
Ellen	✗	✗	✓	✗	✗	✓
Fabian	✗	✗	✓	✓	✗	✗
Gabriel	✓	✓	✗	✗	✓	✓
Hector	✓	✗	✗	✗	✓	✓
Ines	✓	✓	✗	✗	✓	✓
Zelya	✓	✗	✓	✗	✗	✓

A simple dataset of movie ratings

	Star wars	Star trek	Titanic	Pretty Woman	007	Mission Impossible
Alice	✓	✓	✗	✗	✗	✗
Bob	✓	✓	✗	✗	✗	✗
Carlos	✗	✗	✓	✓	✗	✗
David	✗	✗	✓	✓	✗	✗
Ellen	✗	✗	✓	✗	✗	✓
Fabian	✗	✗	✓	✓	✗	✗
Gabriel	✓	✓	✗	✗	✓	✓
Hector	✓	✗	✗	✗	✓	✓
Ines	✓	✓	✗	✗	✓	✓
Zelya	✓	✗	✓	✗	✗	✓

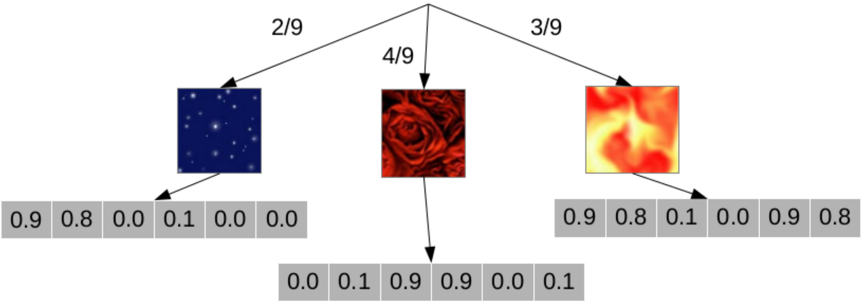
A simple dataset of movie ratings

	Star wars	Star trek	Titanic	Pretty Woman	007	Mission Impossible
Alice	✓	✓	✗	✗	✗	✗
Bob	✓	✓	✗	✗	✗	✗
Carlos	✗	✗	✓	✓	✗	✗
David	✗	✗	✓	✓	✗	✗
Ellen	✗	✗	✓	✗	✗	✓
Fabian	✗	✗	✓	✓	✗	✗
Gabriel	✓	✓	✗	✗	✓	✓
Hector	✓	✗	✗	✗	✓	✓
Ines	✓	✓	✗	✗	✓	✓
Zelya	✓	✗	✓	✗	✗	✓
John	?	✗	?	✓	✗	✗

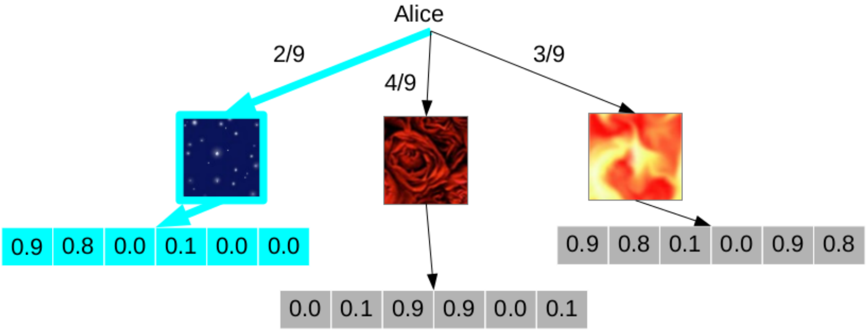
A simple dataset of movie ratings

	Star wars	Star trek	Titanic	Pretty Woman	007	Mission Impossible
Alice	✓	✓	✗	✗	✗	✗
Bob	✓	✓	✗	✗	✗	✗
Carlos	✗	✗	✓	✓	✗	✗
David	✗	✗	✓	✓	✗	✗
Ellen	✗	✗	✓	✗	✗	✓
Fabian	✗	✗	✓	✓	✗	✗
Gabriel	✓	✓	✗	✗	✓	✓
Hector	✓	✗	✗	✗	✓	✓
Ines	✓	✓	✗	✗	✓	✓
Zelya	✓	✗	✓	✗	✗	✓
John	?	✗	?	✓	✗	✗

A probability model for movie ratings

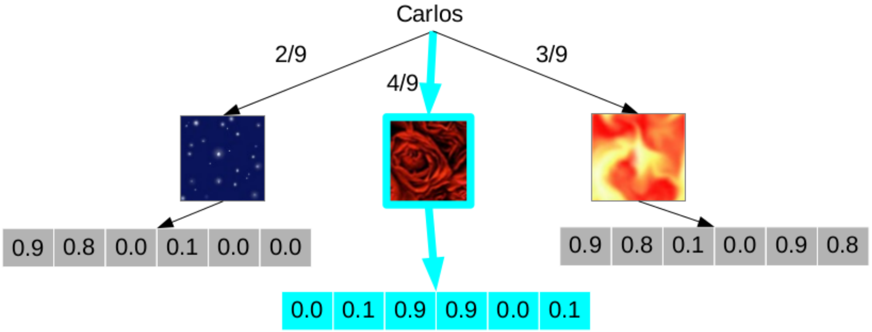


A probability model for movie ratings



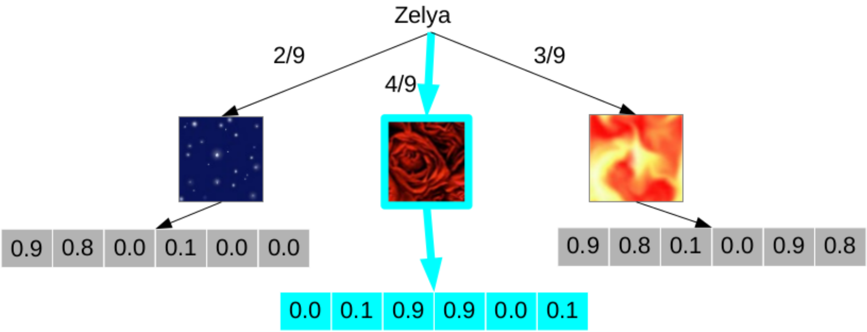
	Star wars	Star trek	Titanic	Pretty Woman	007	Mission Impossible
Alice	✓	✓	✗	✗	✗	✗

A probability model for movie ratings



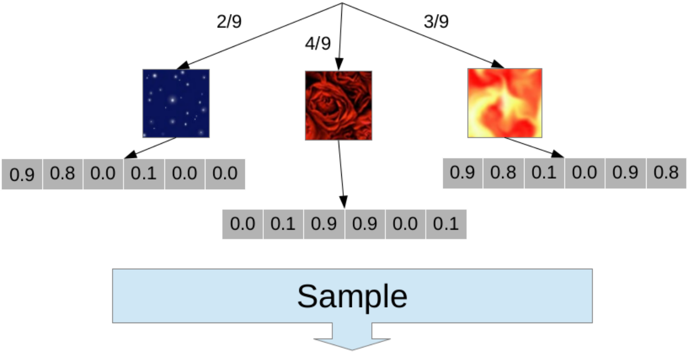
	Star wars	Star trek	Titanic	Pretty Woman	007	Mission Impossible
Carlos	✗	✗	✓	✓	✗	✗

A probability model for movie ratings



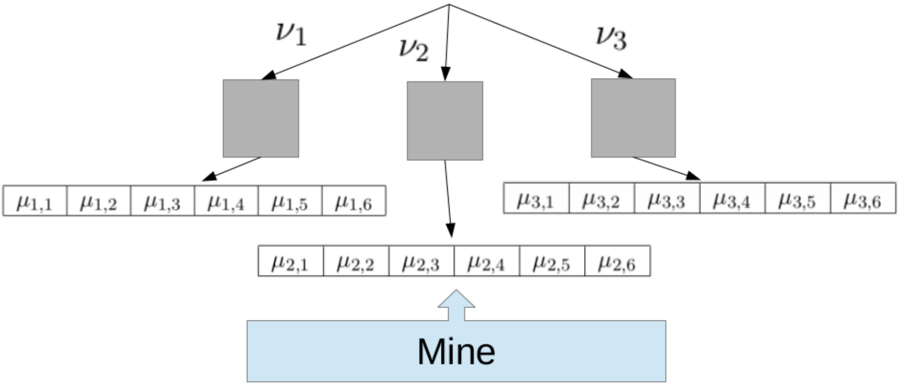
	Star wars	Star trek	Titanic	Pretty Woman	007	Mission Impossible
Zelya	✗	✓	✗	✓	✗	✓

A probability model for movie ratings



	Star wars	Star trek	Titanic	La La Land	007	Mission Impossible
Alice	✔	✔	✖	✖	✖	✖
Bob	✔	✔	✖	✖	✖	✖
Carlos	✖	✖	✔	✔	✖	✖
David	✖	✖	✔	✔	✖	✖
Ellen	✖	✖	✔	✖	✖	✔
Fabian	✖	✖	✔	✔	✖	✖
Gabriel	✔	✔	✖	✖	✔	✔
Hector	✔	✖	✖	✖	✔	✔
Ines	✔	✔	✖	✖	✔	✔
Zelya	✔	✖	✔	✖	✖	✔

A probability model for movie ratings



	Star wars	Star trek	Titanic	La La Land	007	Mission Impossible
Alice	✔	✔	✘	✘	✘	✘
Bob	✔	✔	✘	✘	✘	✘
Carlos	✘	✘	✔	✔	✘	✘
David	✘	✘	✔	✔	✘	✘
Ellen	✘	✘	✔	✘	✘	✔
Fabian	✘	✘	✔	✔	✘	✘
Gabriel	✔	✔	✘	✘	✔	✔
Hector	✔	✘	✘	✘	✔	✔
Ines	✔	✔	✘	✘	✔	✔
Zelya	✔	✘	✔	✘	✘	✔

A probability model for movie ratings

	Star wars	Star trek	Titanic	Pretty Woman	007	Mission Impossible
Alice	✓	✓	✗	✗	✗	✗
Bob	✓	✓	✗	✗	✗	✗
Carlos	✗	✗	✓	✓	✗	✗
David	✗	✗	✓	✓	✗	✗
Ellen	✗	✗	✓	✗	✗	✓
Fabian	✗	✗	✓	✓	✗	✗
Gabriel	✓	✓	✗	✗	✓	✓
Hector	✓	✗	✗	✗	✓	✓
Ines	✓	✓	✗	✗	✓	✓
Zelya	✓	✗	✓	✗	✗	✓
John	?	✗	?	✓	✗	✗

Notation

- $X = (x_{i,j})_{i \leq N, j \leq D}$. Here, $x_{i,j} \in \{0, 1\}$ indicates whether person i liked movie j or not.
- $\bar{\mu} = (\mu_{k,j})_{k \leq K, j \leq D}$. Here, $\mu_{k,j} \in [0, 1]$ denotes the probability that someone in category k likes movie j.
- $\bar{\nu} = (\nu_k)_{k \leq K}$. Here, $\nu_k \in [0, 1]$ denotes the probability that a person belongs to category k.
- $\bar{z} = (z(i))_{i \leq N}$. Here, $z(i) \in \{0, \ldots, K\}$ indicates person i's category.

How to mine a probability model from X?

Maximum-likelihood approach: Solve the following problem.

$$\arg\max_{\bar{\mu},\bar{\nu}} \quad \log p\left(X \mid \bar{\mu},\bar{\nu}\right).$$

$$s.t. \quad \sum_{k\leq K} \nu_k = 1.$$

- Incomplete-data log likelihood.
- Complete-data log likelihood.

$$\log p\left(X,\bar{z} \mid \bar{\mu},\bar{\nu}\right).$$

How to mine a probability model from X?

Maximum-likelihood approach: Solve the following problem.

$$\arg\max_{\bar{\mu},\bar{\nu}} \quad \boxed{\sum_{i\leq N} \log\left(\sum_{z(i)} \nu_{z(i)} \prod_{j\leq D} \mu_{z(i),j}^{x_{i,j}} (1-\mu_{z(i),j})^{1-x_{i,j}}\right).}$$

$$s.t. \quad \sum_{k\leq K} \nu_k = 1.$$

- Incomplete-data log likelihood.
- Complete-data log likelihood.

$$\boxed{\sum_{i\leq N} \log \nu_{z(i)} + \sum_{j\leq D} x_{i,j} \log \mu_{z(i),j} + (1-x_{i,j}) \log\left(1-\mu_{z(i),j}\right).}$$

The dilemma

We are between a problem we want to solve, but we don't know how, and a problem we know how to solve but we don't want to solve. Let's try to connect them.

Connecting incomplete-data and complete-data log likelihoods

Let $\theta = (\bar{\mu}, \bar{\nu})$
How can we connect $\log p(X \mid \theta)$ and $\log p(X, \bar{z} \mid \theta)$?
We can start with the following:

$$p(\bar{z} \mid X, \theta) = \frac{p(X, \bar{z} \mid \theta)}{p(X \mid \theta)}.$$

Connecting incomplete-data and complete-data log likelihoods

Let $\theta = (\bar{\mu}, \bar{\nu})$
How can we connect $\log p(X \mid \theta)$ and $\log p(X, \bar{z} \mid \theta)$?
We can start with the following:

$$p(\bar{z} \mid X, \theta) = \frac{p(X, \bar{z} \mid \theta)}{p(X \mid \theta)}.$$

From here, we can derive that:

$$\log p(X \mid \theta) = \log p(X, \bar{z} \mid \theta) - \log p(\bar{z} \mid X, \theta).$$

Connecting incomplete-data and complete-data log likelihoods

Let $\theta = (\bar{\mu}, \bar{\nu})$
How can we connect $\log p(X \mid \theta)$ and $\log p(X, \bar{z} \mid \theta)$?
We can start with the following:

$$p(\bar{z} \mid X, \theta) = \frac{p(X, \bar{z} \mid \theta)}{p(X \mid \theta)}.$$

From here, we can derive that:

$$\log p(X \mid \theta) = \log p(X, \bar{z} \mid \theta) - \log p(\bar{z} \mid X, \theta).$$

But we don't know the value of $\bar{z}$.

Take expectations on both sides with respect to $\bar{z}$, using some pdf $\tilde{p}\left(\bar{z}\right)$ for $\bar{z}$.

$$\int \tilde{p}\left(\bar{z}\right) \log p(X \mid \theta) d\bar{z} = \int \tilde{p}\left(\bar{z}\right) \log p(X, \bar{z} \mid \theta) d\bar{z} - \int \tilde{p}\left(\bar{z}\right) \log p(\bar{z} \mid X, \theta) d\bar{z}.$$

Since $\log p(X \mid \theta)$ does not depend on $\bar{z}$, we get

$$\log p(X \mid \theta) = \int \tilde{p}\,(\bar{z}) \log p(X, \bar{z} \mid \theta) d\bar{z} - \int \tilde{p}\,(\bar{z}) \log p(\bar{z} \mid X, \theta) d\bar{z}.$$

In other words,

$$\log p(X \mid \theta) = \mathbb{E}_{\tilde{p}(\bar{z})} \log p(X, \bar{z} \mid \theta) - \mathbb{E}_{\tilde{p}(\bar{z})} \log p(\bar{z} \mid X, \theta).$$

Does this look familiar?

In other words,

$$\log p(X \mid \theta) = \mathbb{E}_{\tilde{p}(\bar{z})} \log p(X, \bar{z} \mid \theta) - \mathbb{E}_{\tilde{p}(\bar{z})} \log p(\bar{z} \mid X, \theta).$$

Does this look familiar?

$$d(\theta) = skin\left(\theta, \tilde{p}\right) - vessel\left(\theta, \tilde{p}\right).$$

In other words,

$$\log p(X \mid \theta) = \mathbb{E}_{\tilde{p}(\bar{z})} \log p(X, \bar{z} \mid \theta) - \mathbb{E}_{\tilde{p}(\bar{z})} \log p(\bar{z} \mid X, \theta).$$

Does this look familiar?

$$d(\theta) = skin\,(\theta, \tilde{p}) - vessel\,(\theta, \tilde{p})\,.$$

- Like a vegan flea, we want to maximize the value for θ that maximizes the distance between $\mathbb{E}_{\tilde{p}(\bar{z})} \log p(X, \bar{z} \mid \theta)$ and $\mathbb{E}_{\tilde{p}(\bar{z})} \log p(\bar{z} \mid X, \theta)$!

In other words,

$$\log p(X \mid \theta) = \mathbb{E}_{\tilde{p}(\bar{z})} \log p(X, \bar{z} \mid \theta) - \mathbb{E}_{\tilde{p}(\bar{z})} \log p(\bar{z} \mid X, \theta).$$

Does this look familiar?

$$d(\theta) = skin\left(\theta, \tilde{p}\right) - vessel\left(\theta, \tilde{p}\right).$$

- Like a vegan flea, we want to maximize the value for θ that maximizes the distance between $\mathbb{E}_{\tilde{p}(\bar{z})} \log p(X, \bar{z} \mid \theta)$ and $\mathbb{E}_{\tilde{p}(\bar{z})} \log p(\bar{z} \mid X, \theta)$!
- It turns out that all assumptions hold!

In other words,

$$\log p(X \mid \theta) = \mathbb{E}_{\tilde{p}(\bar{z})} \log p(X, \bar{z} \mid \theta) - \mathbb{E}_{\tilde{p}(\bar{z})} \log p(\bar{z} \mid X, \theta).$$

Does this look familiar?

$$d(\theta) = skin\,(\theta, \tilde{p}) - vessel\,(\theta, \tilde{p})\,.$$

- Like a vegan flea, we want to maximize the value for θ that maximizes the distance between $\mathbb{E}_{\tilde{p}(\bar{z})} \log p(X, \bar{z} \mid \theta)$ and $\mathbb{E}_{\tilde{p}(\bar{z})} \log p(\bar{z} \mid X, \theta)$!
- It turns out that all assumptions hold!
- We can apply our optimization algorithm to approximately maximize $\log p(X \mid \theta)$ with respect to θ.

The EM-algorithm

Prerequisites

A1 It is efficient to calculate $p\left(\bar{z} \mid X, \theta\right)$, for any θ.

A2 It is efficient to compute $\arg\max_{\theta} \mathbb{E}_{p(\bar{z}|X,\theta^o)}\left[\log\, p\left(X, \bar{z} \mid \theta\right)\right]$, for any θ_0.

EM-algorithm

Init Initialize θ^o with random values.

E-step Compute $p\left(\bar{z} \mid X, \theta^o\right)$.

M-step $\theta \leftarrow \arg\max_{\theta} \mathbb{E}_{p(\bar{z}|X,\theta^o)}\left[\log\, p\left(X, \bar{z} \mid \theta\right)\right]$.

Repeat If θ^o and θ are close enough, finish; otherwise, set $\theta^o \leftarrow \theta$ and go to [E-step].