

TOOLS: INFORMATION THEORY

Information content of a random variable

We consider a random variable X with distribution P .

- P expresses what we know *before* we observe X .
- How much information do we *gain* by observing X ?

That is: By information content of X , we mean the difference in information between knowing P and knowing both P and $X = x$.

To reiterate

For the definition of information, it is useful to think of...

- ...the distribution P as what we *expect to happen*.
- ...the sample outcome $X = x$ as what *actually happens*.

Heuristic motivation

Suppose we sample $X = x$ from a distribution P .

- If $P(x)$ is large: Small surprise; we have not gained much additional information.
- If $P(x)$ is small: We have gained more information.

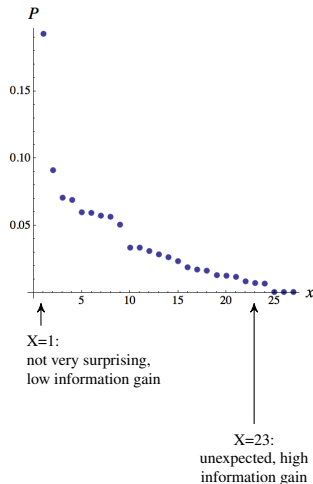
Conclusions

- The information in $X = x$ increases with $\frac{1}{P(x)}$.
- Intuitively, the information gain in two unrelated observations should be additive, so $\frac{1}{P(x)}$ itself is not a useful measure of information.

Definition

The **information** in observing $X = x$ under P is

$$J_P(x) := \log \frac{1}{P(x)} = -\log P(x) .$$



SHANNON'S ENTROPY

We first consider discrete random variables, i.e. $\mathcal{X}$ is finite or countably infinite.

Definition

Let X be a discrete random variable with distribution P . The *expected* information in a draw from P ,

$$\mathbf{H}[X] := \mathbb{E}_P[J_P(X)] = - \sum_{x \in \mathcal{X}} P(x) \log P(x)$$

is called the **Shannon entropy** of X , or the **entropy** for short.

We use the notation $\mathbf{H}[X]$ and $\mathbf{H}(P)$ interchangeably.

Basic Properties

1. The entropy is non-negative:

$$\mathbf{H}[X] \geq 0$$

2. $\mathbf{H}(P) = 0$ means there is no uncertainty in P :

$$\mathbf{H}(P) = 0 \quad \Leftrightarrow \quad P(x_0) = 1 \text{ for some } x_0 \in \mathcal{X} .$$

3. If $\mathcal{X}$ is finite with d elements, the distribution with the largest entropy is the uniform distribution U_d , with

$$\mathbf{H}(U_d) = \log d$$

4. $\mathbf{H}(P)$ is concave as a function of P .

ALTERNATIVE DERIVATION

Suppose we define *some* measure $\mathcal{H}[X]$ of information in X . Instead of a definition, we postulate a number of properties (axioms) that a meaningful measure should satisfy.

Axioms

- We should be able to "remove the joint information" in X and Y from Y by conditioning:

$$\text{(Axiom I)} \quad \mathcal{H}[X, Y] = \mathcal{H}[X] + \mathcal{H}[Y|X]$$

That implies in particular that the information of independent variables is additive:

$$X \perp\!\!\!\perp Y \quad \Rightarrow \quad \mathcal{H}[X, Y] = \mathcal{H}[X] + \mathcal{H}[Y]$$

- If we make a small change to P , then $\mathcal{H}(P)$ should not "jump". That is:

$$\text{(Axiom II)} \quad \mathcal{H}(P) \text{ should be continuous as a function of } P.$$

- If we increase d , the uncertainty in U_d increases; hence, the information gained by sampling should be higher for $d + 1$ than for d :

$$\text{(Axiom III)} \quad \mathcal{H}(U_d) < \mathcal{H}(U_{d+1})$$

Theorem

If a real-valued function $\mathcal{H}$ on distributions on a finite set $\mathcal{X}$ satisfies Axioms I–III, then

$$\mathcal{H}(P) = c \cdot \mathbf{H}(P) \quad \text{for all } P,$$

for some constant $c \in \mathbb{R}_+$. (The constant is the same for all P .)

Over the years, about a dozen different axioms for information measures have been proposed. Loosely speaking, any meaningful combination of two or three of these axioms leads to the same result (i.e. determines the entropy up to scaling).

EXAMPLE: CODING

Suppose we would like to compress a text document (lossless compression).

Huffman Coding

Here is a simple but efficient coding scheme:

1. Given a text, determine the frequency with which each word occurs.
2. Assign short code words to words that occur often, long code words to words that are rare.

This idea (with a specific algorithm for finding determining the code words) is called **Huffman coding**. If all we are allowed to do is to replace text words by code words, this compression method is optimal.

Information-theoretic problems

Suppose we know the distribution P of words in texts. Then we can ask:

1. What is the *expected* compression rate for a random document?
2. Does our encoder achieve the optimal expected rate for P ?

The Source Coding Theorem (Shannon)

Suppose we are given a distribution P on words or symbols and sample a string (a sequence of categorical variables) $X_1, \dots, X_n$ iid from P . Then for every $\varepsilon > 0$, there is a lossless encoder for which

$$\mathbf{H}(P) \leq \mathbb{E} \left[\frac{1}{n} \cdot \text{length}(\text{encoding}(X_1, \dots, X_n)) \right] < \mathbf{H}(P) + \varepsilon$$

for sufficiently large n .

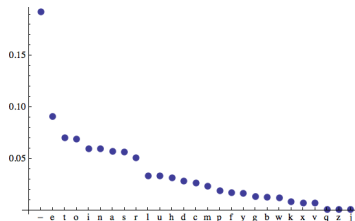
Remarks

- In other words: We can encode the sequence $X_1, \dots, X_n$ without loss of information using $n \mathbf{H}(P)$ bits on average.
- The entropy $\mathbf{H}(P)$ is a lower bound for lossless compression: No code achieves a better (= smaller) rate than $\mathbf{H}(P)$. For fixed-length codes, any rate $< \mathbf{H}(P)$ loses information with probability $\rightarrow 1$ as $n \rightarrow \infty$.

HOW WELL CAN WE COMPRESS ENGLISH TEXT?

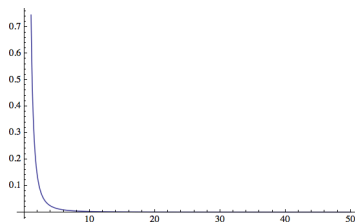
Character-by-character compression

- We can compress text by splitting the text into characters and assigning a code to each character.
- An empirical estimate of character distribution is shown on the right. The entropy is 4.11 bit/character.
- This compression is not very effective: There are 27 characters and $2^4 < 27 \leq 2^5$, hence we can trivially encode with 5 bits/character.



Word-by-word compression

- The ranked frequency of words in English is well-approximated by a Zipf distribution with parameter of about 1 (lower plot).
- It is highly concentrated on a few common words. That means it has relatively low entropy.
- Splitting into words instead of characters hence achieves much better compression rates.
- Commonly used lossless compression algorithms (e.g. Lempel-Ziv) split into substrings which are not necessarily words.



KULLBACK-LEIBLER DIVERGENCE

Heuristic motivation

Suppose we wish to compare two distributions P and Q on $\mathcal{X}$.

- The entropy $\mathbf{H}[Q] = \mathbb{E}_Q[J_Q(X)]$ measures how much information gain (in terms of Q) we can *expect* from a random sample from Q .
- Now ask instead: How much information gain in terms of Q can we expect from a random sample drawn from P ? We compute: $\mathbb{E}_P[J_Q(X)]$.
- A measure of difference between P and Q should vanish if $Q = P$. Since $P = Q$ means $\mathbb{E}_P[J_Q(X)] = \mathbf{H}(P)$, which is usually not 0, we have to normalize by subtracting $\mathbf{H}(P)$.

Definition

The **Kullback-Leibler divergence** or the **relative entropy** of P and Q is

$$\mathbf{KL}(P\|Q) := \mathbb{E}_P[J_Q(X)] - \mathbf{H}(P) = \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)}$$

Properties

- Positive definiteness

$$\mathbf{KL}(P\|Q) \geq 0 \quad \text{and} \quad \mathbf{KL}(P\|Q) = 0 \Leftrightarrow P = Q .$$

- $\mathbf{KL}(P\|Q)$ is convex in both P and Q .

ASYMMETRY OF THE KL DIVERGENCE

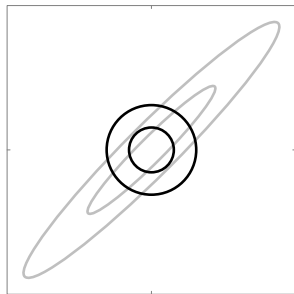
In both the discrete and continuous case:

- The KL divergence is *not* symmetric: $\mathbf{KL}(p||q) \neq \mathbf{KL}(q||p)$ in general.
- It does *not* satisfy a triangle inequality.
- In particular, it is not a metric on probability distributions.

However: The KL divergence exhibits certain geometric properties similar to metrics. In particular, it satisfies a property similar to Pythagoras' theorem (that we will not discuss here).

Asymmetry

- If p puts mass in a region where q almost vanishes, then p/q is very large in that region.
- Since both are normalized, there is some other regions where p/q is small.
- Because of the logarithm, these two do not cancel out.
- For instance, the isotropic Gaussian (solid black contours) on the right has small mass in regions where the eccentric Gaussian (grey) puts mass.



Heuristic Motivation

- Another question we can ask about a pair X, Y of random variables is: How much information do they share?
- In other words: How much does observing X tell us about Y ?
- If X and Y contain no shared information, they are independent, and their joint distribution is $P(x, y) = P(x)P(y)$.
- Idea: Compare the actual joint distribution to the independent case using KL divergence.

Definition

The **mutual information** of X and Y is

$$I[X, Y] := \mathbf{KL}(P(x, y) \| P(x)P(y)) = \sum_{x, y} P(x, y) \log \frac{P(x, y)}{P(x)P(y)}$$

Mutual information characterizes independence

$$I[X, Y] = 0 \quad \Leftrightarrow \quad X \perp\!\!\!\perp Y$$

THE CONTINUOUS CASE

If the sample space $\mathcal{X}$ is uncountable (e.g. $\mathcal{X} = \mathbb{R}$), instead of P and Q we consider densities p and q , we have to substitute integrals for sums.

Differential entropy

The definition of entropy we use for continuous distributions is

$$\mathbf{H}[X] := - \int_{\mathcal{X}} p(x) \log p(x) dx$$

- Since p is a density, we can have $\log p(x) > 0$, and $\mathbf{H}[X]$ can be negative.
- This form of entropy is also called **differential entropy** to distinguish it from the (Shannon) entropy.

KL divergence and mutual information

KL and I are defined analogously to the discrete case:

$$\begin{aligned}\mathbf{KL}(p||q) &:= \int_{\mathcal{X}} p(x) \log \frac{p(x)}{q(x)} dx \\ I[X, Y] &:= \int_{\mathcal{X}} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} dx dy\end{aligned}$$

Unlike the entropy, these have the same properties as we have listed in the discrete case.

INFORMATION THEORY AND STATISTICS

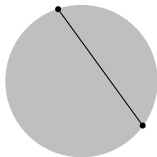
Here are some points that we do not have time to discuss in detail:

- Exponential families are **maximum entropy models**: If $\mathbf{T}$ is a statistic with values in $\mathbb{R}^d$, and $\mathbf{t} \in \mathbb{R}^d$, the distribution that maximizes the entropy under the constraint $\mathbb{E}[\mathbf{T}(X)] = \mathbf{t}$ is an exponential family distribution with sufficient statistic $\mathbf{T}$.
- Maximum likelihood minimizes the KL divergence between empirical distribution and model.
- Variance and covariance regarded as first-order approximations to entropy and mutual information, and the χ^2 -statistic as a second-order approximation to the KL divergence.
- Various methods can be derived by substituting information-theoretic for traditional statistical quantities.
- Example: A dimension-reduction technique called *independent component analysis* can be motivated as (roughly speaking) a PCA-like method which measures independence in terms of mutual information rather than covariance.

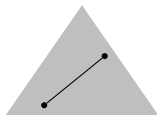
TOOLS: CONVEXITY

Definition

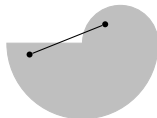
A set $A \subset \mathbb{R}^d$ is called **convex** if, for every two points $\mathbf{x}, \mathbf{y} \in A$, the straight line connecting $\mathbf{x}$ and $\mathbf{y}$ is completely contained in A .



convex



convex



not convex

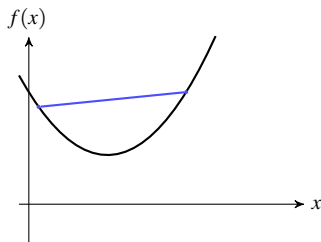
Quantitatively

A is convex if and only if $\lambda \mathbf{x} + (1 - \lambda) \mathbf{y} \in A$ for all $\mathbf{x}, \mathbf{y} \in A$ and $\lambda \in [0, 1]$.

CONVEX FUNCTIONS

A function f is **convex** if every line segment between values of f lies above the graph of f .

Equivalently: If the area above the curve is a convex set.



- Quantitatively: f is convex if

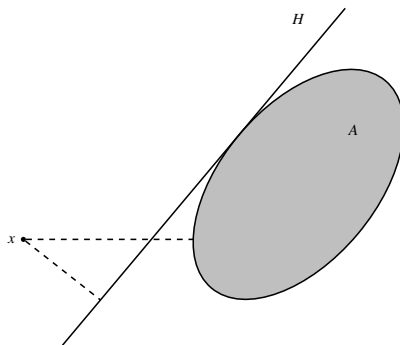
$$\lambda f(\mathbf{x}) + (1 - \lambda)f(\mathbf{y}) \geq f(\lambda \mathbf{x} + (1 - \lambda)\mathbf{y})$$

for all $\mathbf{x}$ and $\mathbf{y}$ in the domain of f and all $\lambda \in [0, 1]$.

- A twice differentiable function is convex if $H_f(\mathbf{x})$ is positive semidefinite for all $\mathbf{x}$ (in $\mathbb{R}$: $f''(x) \geq 0$ for all x).

CONVEX DUALITY

The closest distance between a point $\mathbf{x}$ and a convex set A is the maximum over the distances between $\mathbf{x}$ and all hyperplanes H which separate $\mathbf{x}$ and A .



$$d(\mathbf{x}, A) = \sup_{H \text{ separating}} d(\mathbf{x}, H)$$

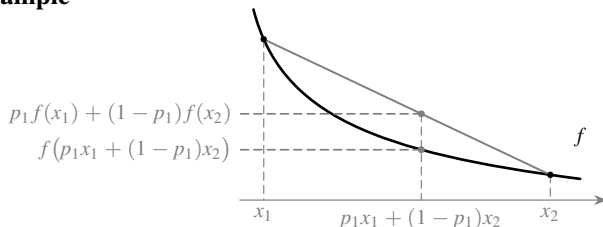
JENSEN'S INEQUALITY

For probability distribution P on $\mathbb{R}^d$ and a convex function $f : \mathbb{R}^d \rightarrow \mathbb{R}$,

$$f(\mathbb{E}_P[X]) \leq \mathbb{E}_P[f(X)] .$$

Equality holds if and only if f is (almost surely) constant or linear on (convex) support of P .

Example



Example: Consider $f(x) = -\log x$ and $p_i \geq 0$ with $\sum p_i = 1$. For points $x_i \in (0, \infty)$, we have

$$-\log\left(\sum_i p_i x_i\right) \leq -\sum_i p_i \log(x_i)$$

APPLICATION: POSITIVE DEFINITENESS OF KL

The KL divergence satisfies $\mathbf{KL}(p||q) \geq 0$, with equality iff $p = q$.

This follows immediately from Jensen's inequality:

$$\mathbf{KL}(p||q) = \mathbb{E}_P \left[-\log \frac{q(X)}{p(X)} \right] \stackrel{\text{Jensen}}{\geq} -\log \mathbb{E}_P \left[\frac{q(X)}{p(X)} \right] \geq -\log 1 = 0 .$$

We have $\mathbb{E}_P[q/p] = 1$ if and only if $p = q$. In this case, $-\log(p/q)$ is constant, so Jensen becomes an equality.

EXPECTATION MAXIMIZATION

LOG-LIKELIHOODS

- Exponential family models: $p(\mathbf{x}|\theta) = f(\mathbf{x})e^{\theta^\top \mathbf{T}(\mathbf{x})} / Z(\theta)$

$$\ell(\theta) = \theta^\top \sum_i T(\mathbf{x}_i) - n \log Z(\theta) \quad (+ \text{ constants})$$

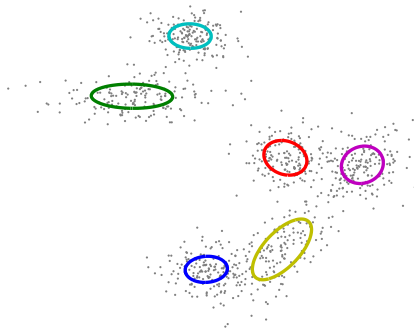
- Concave function.
- Maximum may be closed-form.
- If not, numerical optimization is still generally straightforward.

- Latent variable models: $p(\mathbf{x}|\theta_x, \theta_z) = \underbrace{\int d\mathbf{z} f_x(\mathbf{x}) \frac{e^{\phi(\theta_x, \mathbf{z})^\top \mathbf{T}_x(\mathbf{x})}}{Z_x(\phi(\theta_x, \mathbf{z}))}}_{p(\mathbf{x}|\mathbf{z}, \theta_x)} \underbrace{f_z(\mathbf{z}) \frac{e^{\theta_z^\top \mathbf{T}_z(\mathbf{z})}}{Z_z(\theta_z)}}_{p(\mathbf{z}|\theta_z)}$

$$\ell(\theta_x, \theta_z) = \sum_i \log \int d\mathbf{z} f_x(\mathbf{x}) \frac{e^{\phi(\theta_x, \mathbf{z})^\top \mathbf{T}_x(\mathbf{x})}}{Z_x(\phi(\theta_x, \mathbf{z}))} f_z(\mathbf{z}) \frac{e^{\theta_z^\top \mathbf{T}_z(\mathbf{z})}}{Z_z(\theta_z)}$$

- Usually no closed form optimum.
- Often multiple local maxima.
- Direct numerical optimization may be possible but infrequently easy.

EXAMPLE: MIXTURE OF GAUSSIANS



Data: $\mathcal{D} = \{\mathbf{x}_1 \dots \mathbf{x}_n\}$

Latent process:

$$s_i \stackrel{\text{iid}}{\sim} \text{Disc}[\pi]$$

Component distributions:

$$\mathbf{x}_i \mid (s_i = k) \sim \mathcal{P}_k[\theta_k] = \mathcal{N}(\mu_k, \Sigma_k)$$

Marginal distribution:

$$P(\mathbf{x}_i) = \sum_{k=1}^K \pi_k P_k(\mathbf{x}_i; \theta_k)$$

Log-likelihood:

$$\ell(\{\mu_k\}, \{\Sigma_k\}, \pi) = \sum_{i=1}^n \log \sum_{k=1}^K \frac{\pi_k}{\sqrt{|2\pi\Sigma_k|}} e^{-\frac{1}{2}(\mathbf{x}_i - \mu_k)^\top \Sigma_k^{-1}(\mathbf{x}_i - \mu_k)}$$

THE JOINT-DATA LIKELIHOOD AND EM

- For many models, maximization might be straightforward if $\mathbf{z}$ were not latent, and we could just maximize the joint-data likelihood:

$$\ell(\theta_x, \theta_z) = \sum_i \phi(\theta_x, \mathbf{z}_i)^\top \mathbf{T}_x(\mathbf{x}_i) + \theta_z^\top \sum_i \mathbf{T}_z(\mathbf{z}_i) - \sum_i \log Z_x(\phi(\theta_x, \mathbf{z}_i)) - n \log Z_z(\theta_z)$$

- Conversely, if we knew θ , we might easily compute (the posterior over) the values of $\mathbf{z}$.
- **Idea:** update θ and (the distribution on) $\mathbf{z}$ in alternation, to reach a self-consistent answer.
Will this yield the right answer?
- Typically, it will (as we shall see). This is the **Expectation Maximization (EM)** algorithm.

THE EXPECTATION MAXIMIZATION (EM) ALGORITHM

The EM algorithm (Dempster, Laird & Rubin, 1977; but significant earlier precedents) finds a (local) maximum of a latent variable model likelihood.

Start from arbitrary values of the parameters, and iterate two steps:

E step: Fill in values of latent variables according to posterior given data.

M step: Maximize likelihood as if latent variables were not hidden.

- Decomposes difficult problems into series of tractable steps.
- An alternative to gradient-based iterative methods.
- No learning rate.
- In ML, the E step is called **inference**, and the M step **learning**. In stats, these are often **imputation** and **inference** or **estimation**.
- Not essential for simple models (like MoGs/FA), though often more efficient than alternatives. Crucial for learning in complex settings.
- Provides a framework for **principled approximations**.

THE LOWER BOUND FOR EM – “FREE ENERGY”

Observed data $\mathcal{D} = \{\mathbf{x}_i\}$; Latent variables $\mathcal{Z} = \{\mathbf{z}_i\}$; Parameters $\theta = \{\theta_x, \theta_z\}$.

Log-likelihood:
$$\ell(\theta) = \log P(\mathcal{D}|\theta) = \log \int d\mathcal{Z} P(\mathcal{Z}, \mathcal{D}|\theta)$$

By Jensen, any distribution, $q(\mathcal{Z})$, over the latent variables generates a lower bound:

$$\ell(\theta) = \log \int d\mathcal{Z} q(\mathcal{Z}) \frac{P(\mathcal{Z}, \mathcal{D}|\theta)}{q(\mathcal{Z})} \geq \int d\mathcal{Z} q(\mathcal{Z}) \log \frac{P(\mathcal{Z}, \mathcal{D}|\theta)}{q(\mathcal{Z})} \stackrel{\text{def}}{=} \mathcal{F}(q, \theta).$$

Now,

$$\begin{aligned} \int d\mathcal{Z} q(\mathcal{Z}) \log \frac{P(\mathcal{Z}, \mathcal{D}|\theta)}{q(\mathcal{Z})} &= \int d\mathcal{Z} q(\mathcal{Z}) \log P(\mathcal{Z}, \mathcal{D}|\theta) - \int d\mathcal{Z} q(\mathcal{Z}) \log q(\mathcal{Z}) \\ &= \int d\mathcal{Z} q(\mathcal{Z}) \log P(\mathcal{Z}, \mathcal{D}|\theta) + \mathbf{H}(q), \end{aligned}$$

where $\mathbf{H}(q)$ is the entropy of $q(\mathcal{Z})$.

So:

$$\mathcal{F}(q, \theta) = \langle \log P(\mathcal{Z}, \mathcal{D}|\theta) \rangle_{q(\mathcal{Z})} + \mathbf{H}(q)$$

THE E AND M STEPS OF EM

The free-energy lower bound on $\ell(\theta)$ is a function of θ and a distribution q :

$$\mathcal{F}(q, \theta) = \langle \log P(\mathcal{Z}, \mathcal{D} | \theta) \rangle_{q(\mathcal{Z})} + \mathbf{H}(q),$$

The EM steps can be re-written:

- **E step:** optimize $\mathcal{F}(q, \theta)$ wrt distribution over hidden variables holding parameters fixed:

$$q^{(t)}(\mathcal{Z}) := \arg \max_{q(\mathcal{Z})} \mathcal{F}(q(\mathcal{Z}), \theta^{(t-1)}).$$

- **M step:** maximize $\mathcal{F}(q, \theta)$ wrt parameters holding hidden distribution fixed:

$$\theta^{(t)} := \arg \max_{\theta} \mathcal{F}(q^{(t)}(\mathcal{Z}), \theta) = \arg \max_{\theta} \langle \log P(\mathcal{Z}, \mathcal{D} | \theta) \rangle_{q^{(t)}(\mathcal{Z})}$$

The second equality comes from the fact that $\mathbf{H}(q^{(t)}(\mathcal{Z}))$ does not depend directly on θ .

THE E STEP

The free energy can be re-written

$$\begin{aligned}\mathcal{F}(q, \theta) &= \int q(\mathcal{Z}) \log \frac{P(\mathcal{Z}, \mathcal{D}|\theta)}{q(\mathcal{Z})} d\mathcal{Z} \\ &= \int q(\mathcal{Z}) \log \frac{P(\mathcal{Z}|\mathcal{D}, \theta)P(\mathcal{D}|\theta)}{q(\mathcal{Z})} d\mathcal{Z} \\ &= \int q(\mathcal{Z}) \log P(\mathcal{D}|\theta) d\mathcal{Z} + \int q(\mathcal{Z}) \log \frac{P(\mathcal{Z}|\mathcal{D}, \theta)}{q(\mathcal{Z})} d\mathcal{Z} \\ &= \ell(\theta) - \mathbf{KL}(q(\mathcal{Z})\|P(\mathcal{Z}|\mathcal{D}, \theta))\end{aligned}$$

The second term is the Kullback-Leibler divergence.

This means that, for fixed θ , $\mathcal{F}$ is bounded above by ℓ , and achieves that bound when $\mathbf{KL}(q(\mathcal{Z})\|P(\mathcal{Z}|\mathcal{D}, \theta)) = 0$.

But $\mathbf{KL}(q\|p)$ is zero if and only if $q = p$.

So, the E step sets

$$q^{(t)}(\mathcal{Z}) = P(\mathcal{Z}|\mathcal{D}, \theta^{(t-1)}) \quad [\text{inference / imputation}]$$

and, after an E step, the free energy **equals** the likelihood.

EM NEVER DECREASES THE LIKELIHOOD

The E and M steps together never decrease the log likelihood:

$$\ell(\theta^{(t-1)}) \underset{\text{E step}}{=} \mathcal{F}(q^{(t)}, \theta^{(t-1)}) \underset{\text{M step}}{\leq} \mathcal{F}(q^{(t)}, \theta^{(t)}) \underset{\text{Jensen}}{\leq} \ell(\theta^{(t)}),$$

- The E step brings the free energy to the likelihood.
- The M-step maximizes the free energy wrt θ .
- $\mathcal{F} \leq \ell$ by Jensen – or, equivalently, from the non-negativity of KL

If the M-step is executed so that $\theta^{(t)} \neq \theta^{(t-1)}$ iff $\mathcal{F}$ increases, then the overall EM iteration will step to a new value of θ iff the likelihood increases.

Can also show that fixed points of EM (generally) correspond to maxima of the likelihood (see appendices).

- An **iterative** algorithm that finds (local) maxima of the likelihood of a latent variable model.

$$\ell(\theta) = \log P(\mathcal{D}|\theta) = \log \int d\mathcal{Z} P(\mathcal{D}|\mathcal{Z}, \theta) P(\mathcal{Z}|\theta)$$

- Increases a **variational lower bound** on the likelihood by coordinate ascent.

$$\mathcal{F}(q, \theta) = \langle \log P(\mathcal{Z}, \mathcal{D}|\theta) \rangle_{q(\mathcal{Z})} + \mathbf{H}(q) = \ell(\theta) - \mathbf{KL}(q(\mathcal{Z}) \| P(\mathcal{Z}|\mathcal{D}, \theta)) \leq \ell(\theta)$$

- **E step:**

$$q^{(t)}(\mathcal{Z}) := \arg \max_{q(\mathcal{Z})} \mathcal{F}(q(\mathcal{Z}), \theta^{(t-1)}) = P(\mathcal{Z}|\mathcal{D}, \theta^{(t-1)})$$

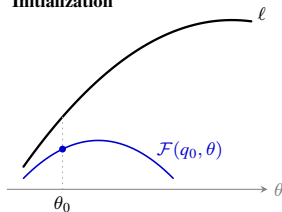
- **M step:**

$$\theta^{(t)} := \arg \max_{\theta} \mathcal{F}(q^{(t)}(\mathcal{Z}), \theta) = \arg \max_{\theta} \langle \log P(\mathcal{Z}, \mathcal{D}|\theta) \rangle_{q^{(t)}(\mathcal{Z})}$$

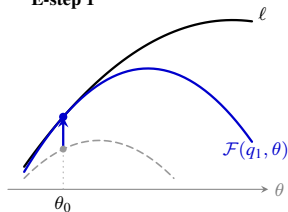
- After E-step $\mathcal{F}(q, \theta) = \ell(\theta) \Rightarrow$ maximum of free-energy is maximum of likelihood.

FREE ENERGY AND LOG-LIKELIHOOD

Initialization

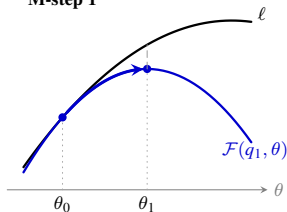


E-step 1



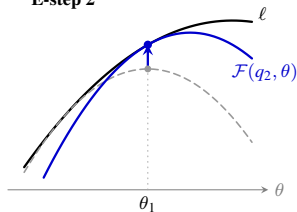
Choose q_1 to touch ℓ at θ_0 .

M-step 1



Maximize $\mathcal{F}(q_1, \cdot)$ get θ_1 .

E-step 2



Choose q_2 to touch ℓ at θ_1 .

PARTIAL M STEPS AND PARTIAL E STEPS

Partial M steps: The proof holds even if we just *increase* $\mathcal{F}$ wrt θ rather than maximize. (Dempster, Laird and Rubin (1977) call this the generalized EM, or GEM, algorithm).

In fact, immediately after an E step

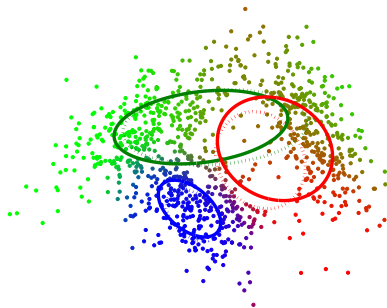
$$\left. \frac{\partial}{\partial \theta} \right|_{\theta^{(t-1)}} \langle \log P(\mathcal{D}, \mathcal{Z}|\theta) \rangle_{q^{(t)}(\mathcal{Z})[=P(\mathcal{Z}|\mathcal{D}, \theta^{(t-1)})]} = \left. \frac{\partial}{\partial \theta} \right|_{\theta^{(t-1)}} \log P(\mathcal{D}|\theta)$$

[cf. mixture gradients in latent variable models section.]

This means the E-step (inference) can be used to construct other gradient-based optimization schemes (e.g. “Expectation Conjugate Gradient”, Salakhutdinov et al. *ICML* 2003).

Partial E steps: We can also just *increase* $\mathcal{F}$ wrt to some of the q s.

For example, sparse or online versions of the EM algorithm would compute the posterior for a subset of the data points or as the data arrives, respectively. One might also update the posterior over a subset of the hidden variables, while holding others fixed...



- Evaluate responsibilities

$$r_{ik} = \frac{P_k(\mathbf{x}_i)\pi_k}{\sum_{k'} P_{k'}(\mathbf{x}_i)\pi_{k'}}$$

- Update parameters

$$\mu_k \leftarrow \frac{\sum_i r_{ik} \mathbf{x}_i}{\sum_i r_{ik}}$$

$$\Sigma_k \leftarrow \frac{\sum_i r_{ik} (\mathbf{x}_i - \mu_k)(\mathbf{x}_i - \mu_k)^\top}{\sum_i r_{ik}}$$

$$\pi_k \leftarrow \frac{\sum_i r_{ik}}{n}$$

THE GAUSSIAN MIXTURE MODEL (E-STEP)

In a univariate Gaussian mixture model, the density of a data point x is:

$$p(x|\theta) = \sum_{k=1}^K p(s = k|\theta)p(x|s = k, \theta) \propto \sum_{k=1}^K \frac{\pi_k}{\sigma_k} \exp \left\{ -\frac{1}{2\sigma_k^2} (x - \mu_k)^2 \right\},$$

where θ is the collection of parameters: means μ_k , variances σ_k^2 and mixing proportions $\pi_k = p(s = k|\theta)$.

The hidden variable s_i indicates which component generated observation x_i .

The E-step computes the posterior for s_i given the current parameters:

$$q(s_i) = p(s_i|x_i, \theta) \propto p(x_i|s_i, \theta)p(s_i|\theta)$$
$$r_{ik} \stackrel{\text{def}}{=} q(s_i = k) \propto \frac{\pi_k}{\sigma_k} \exp \left\{ -\frac{1}{2\sigma_k^2} (x_i - \mu_k)^2 \right\} \quad (\text{responsibilities}) \quad \leftarrow \langle \delta_{s_i=k} \rangle_q$$

with the normalization such that $\sum_k r_{ik} = 1$.

THE GAUSSIAN MIXTURE MODEL (M-STEP)

In the M-step we optimize the sum (since s is discrete):

$$\begin{aligned} E &= \langle \log p(x, s | \theta) \rangle_{q(s)} = \sum q(s) \log[p(s | \theta) p(x | s, \theta)] \\ &= \sum_{i,k} r_{ik} \left[\log \pi_k - \log \sigma_k - \frac{1}{\sigma_k^2} (x_i - \mu_k)^2 \right]. \end{aligned}$$

Optimum is found by setting the partial derivatives of E to zero:

$$\begin{aligned} \frac{\partial}{\partial \mu_k} E &= \sum_i r_{ik} \frac{(x_i - \mu_k)}{2\sigma_k^2} = 0 \Rightarrow \mu_k = \frac{\sum_i r_{ik} x_i}{\sum_i r_{ik}}, \\ \frac{\partial}{\partial \sigma_k} E &= \sum_i r_{ik} \left[-\frac{1}{\sigma_k} + \frac{(x_i - \mu_k)^2}{\sigma_k^3} \right] = 0 \Rightarrow \sigma_k^2 = \frac{\sum_i r_{ik} (x_i - \mu_k)^2}{\sum_i r_{ik}}, \\ \frac{\partial}{\partial \pi_k} E &= \sum_i r_{ik} \frac{1}{\pi_k}, \quad \frac{\partial E}{\partial \pi_k} + \lambda = 0 \Rightarrow \pi_k = \frac{1}{n} \sum_i r_{ik}. \end{aligned}$$

(Here, λ is a Lagrange multiplier ensuring that the mixing proportions sum to unity. We discuss Lagrange multipliers in detail later.)

EM FOR EXPONENTIAL FAMILIES

EM is often applied to models whose **joint** over $\xi = (\mathbf{z}, \mathbf{x})$ has exponential-family form:

$$p(\xi|\theta) = f(\xi) \exp\{\theta^\top \mathbf{T}(\xi)\} / Z(\theta)$$

(with $Z(\theta) = \int f(\xi) \exp\{\theta^\top \mathbf{T}(\xi)\} d\xi$) but whose marginal $p(\mathbf{x}) \notin \text{ExpFam}$.

The free energy dependence on θ is given by:

$$\begin{aligned}\mathcal{F}(q, \theta) &= \int q(\mathbf{z}) \log p(\mathbf{z}, \mathbf{x}|\theta) d\mathbf{z} + \mathbf{H}(q) \\ &= \int q(\mathbf{z}) [\theta^\top \mathbf{T}(\mathbf{z}, \mathbf{x}) - \log Z(\theta)] d\mathbf{z} + \text{const wrt } \theta \\ &= \theta^\top \langle \mathbf{T}(\mathbf{z}, \mathbf{x}) \rangle_{q(\mathbf{z})} - \log Z(\theta) + \text{const wrt } \theta\end{aligned}$$

So, in the **E step** all we need to compute are the **expected sufficient statistics** under q .

We also have:

$$\begin{aligned}\frac{\partial}{\partial \theta} \log Z(\theta) &= \frac{1}{Z(\theta)} \frac{\partial}{\partial \theta} Z(\theta) = \frac{1}{Z(\theta)} \frac{\partial}{\partial \theta} \int f(\xi) \exp\{\theta^\top \mathbf{T}(\xi)\} \\ &= \int \underbrace{\frac{1}{Z(\theta)} f(\xi) \exp\{\theta^\top \mathbf{T}(\xi)\}}_{p(\xi|\theta)} \cdot \mathbf{T}(\xi) = \langle \mathbf{T}(\xi) \rangle_{p(\xi|\theta)}\end{aligned}$$

Thus, the **M step** solves:
$$\frac{\partial \mathcal{F}}{\partial \theta} = \langle \mathbf{T}(\mathbf{z}, \mathbf{x}) \rangle_{q(\mathbf{z})} - \langle \mathbf{T}(\mathbf{z}, \mathbf{x}) \rangle_{p(\mathbf{z}, \mathbf{x}|\theta)} = 0$$

EM FOR EXPONENTIAL FAMILY MIXTURES

To derive EM formally for models with discrete latents (including mixtures) it is useful to introduce an **indicator** vector $\mathbf{s}$ in place of the discrete s .

$$s_i = k \quad \Leftrightarrow \quad \mathbf{s}_i = [0, 0, \dots, \underbrace{1}_{k\text{th position}}, \dots, 0]$$

Collecting the K component distributions' natural params into a matrix $\Theta = [\theta_k]$:

$$\log P(\mathcal{D}, \mathcal{S}) = \sum_i \left[(\log \pi)^\top \mathbf{s}_i + \mathbf{s}_i^\top \Theta^\top T(\mathbf{x}_i) - \mathbf{s}_i^\top \log \mathbf{Z}(\Theta) \right] + \text{const}$$

where $\log \mathbf{Z}(\Theta)$ collects the log-normalizers for all components into a K -element vector. Then, the expected sufficient statistics (E-step) are:

$$\begin{aligned} \sum_i \langle \mathbf{s}_i \rangle_q & \quad (\text{responsibilities } r_{ik}) \\ \sum_i T(\mathbf{x}_i) \langle \mathbf{s}_i^\top \rangle_q & \quad (\text{responsibility-weighted sufficient stats}) \end{aligned}$$

And maximization of the expected log-joint (M-step) gives:

$$\begin{aligned} \pi^{(t+1)} & \propto \sum_i \langle \mathbf{s}_i \rangle_q \\ \langle T(\mathbf{x}) | \theta_k^{(t+1)} \rangle & = \left(\sum_i T(\mathbf{x}_i) \langle [\mathbf{s}_i]_k \rangle_q \right) / \left(\sum_i \langle [\mathbf{s}_i]_k \rangle_q \right) \end{aligned}$$

What if we have a prior?

$$p(\xi|\theta) = f(\xi) \exp\{\theta^\top \mathbf{T}(\xi)\} / Z(\theta) \qquad p(\theta) = F(\nu, \tau) \exp\{\theta^\top \tau\} / Z(\theta)^\nu$$

Augment the free energy by adding the log prior:

$$\begin{aligned} \mathcal{F}_{\text{MAP}}(q, \theta) &= \int q(\mathcal{Z}) \log p(\mathcal{Z}, \mathcal{D}, \theta) d\mathcal{Z} + \mathbf{H}(q) \leq \log P(\mathcal{D}|\theta) + \log P(\theta) \\ &= \int q(\mathcal{Z}) \left[\theta^\top \left(\sum_i \mathbf{T}(\xi_i) + \tau \right) - (n + \nu) \log Z(\theta) \right] d\mathcal{Z} + \text{const wrt } \theta \\ &= \theta^\top \left(\left\langle \sum_i \mathbf{T}(\xi_i) \right\rangle_{q(\mathbf{z})} + \tau \right) - (n + \nu) \log Z(\theta) + \text{const wrt } \theta \end{aligned}$$

So, the expected sufficient statistics in the E step are unchanged.

Thus, after an E-step the augmented free-energy **equals** the log-joint, and so free-energy maxima are log-joint maxima (i.e. MAP values).

Can we find posteriors? Only approximately – cf. COMP0085 (“Variational Bayes”).

FIXED POINTS OF EM ARE STATIONARY POINTS IN ℓ

Let a fixed point of EM occur with parameter θ^* . Then:

$$\left. \frac{\partial}{\partial \theta} \langle \log P(\mathcal{Z}, \mathcal{D} | \theta) \rangle_{P(\mathcal{Z}|\mathcal{D}, \theta^*)} \right|_{\theta^*} = 0$$

Now,

$$\begin{aligned} \ell(\theta) &= \log P(\mathcal{D}|\theta) = \langle \log P(\mathcal{D}|\theta) \rangle_{P(\mathcal{Z}|\mathcal{D}, \theta^*)} \\ &= \left\langle \log \frac{P(\mathcal{Z}, \mathcal{D}|\theta)}{P(\mathcal{Z}|\mathcal{D}, \theta^*)} \right\rangle_{P(\mathcal{Z}|\mathcal{D}, \theta^*)} \\ &= \langle \log P(\mathcal{Z}, \mathcal{D}|\theta) \rangle_{P(\mathcal{Z}|\mathcal{D}, \theta^*)} - \langle \log P(\mathcal{Z}|\mathcal{D}, \theta^*) \rangle_{P(\mathcal{Z}|\mathcal{D}, \theta^*)} \end{aligned}$$

so,

$$\frac{d}{d\theta} \ell(\theta) = \frac{d}{d\theta} \langle \log P(\mathcal{Z}, \mathcal{D}|\theta) \rangle_{P(\mathcal{Z}|\mathcal{D}, \theta^*)} - \frac{d}{d\theta} \langle \log P(\mathcal{Z}|\mathcal{D}, \theta^*) \rangle_{P(\mathcal{Z}|\mathcal{D}, \theta^*)}$$

The second term is 0 at θ^* if the derivative exists (minimum of $\mathbf{KL}(\cdot||\cdot)$), and thus:

$$\left. \frac{d}{d\theta} \ell(\theta) \right|_{\theta^*} = \left. \frac{d}{d\theta} \langle \log P(\mathcal{Z}, \mathcal{D}|\theta) \rangle_{P(\mathcal{Z}|\mathcal{D}, \theta^*)} \right|_{\theta^*} = 0$$

If EM converges, it converges to a stationary point of $\ell(\theta)$.

MAXIMA IN $\mathcal{F}$ CORRESPOND TO MAXIMA IN ℓ

Suppose (q^*, θ^*) is a joint local maximum of $\mathcal{F}$. We have

$$\ell(\theta) = \max_q \mathcal{F}(q, \theta) = \mathcal{F}(P(\mathcal{Z}|\mathcal{D}, \theta), \theta) .$$

It follows that

$$\mathcal{F}(q^*, \theta^*) = \max_{\theta} (\max_q \mathcal{F}(q, \theta)) = \max_{\theta} \ell(\theta) .$$

If (q^*, θ^*) is a (local) maximum of $\mathcal{F}$, then θ^* is a (local) maximum of ℓ .