Approximate inference and learning

Bayesian Statistics and Machine Learning

Dominik Endres

January 21, 2016

Philipps Universität Marburg

Outline

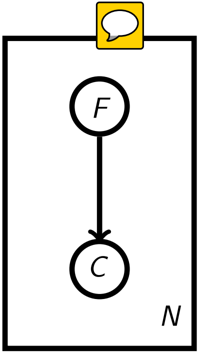
Learning example: how loaded is my coin, and how often ?

You are given a (new?) coin for each toss.

- Possible outcomes $C \in \{h, t\}$.
- Possible coins types: $F \in \{f, l\}$
- You toss N times.
- You observe sequence $S = (t, h, \ldots, t)$.

Questions

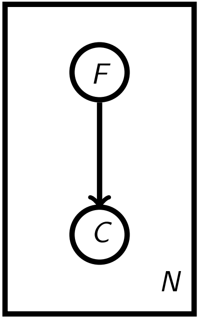
- Is the coin loaded at toss n ? $P(F_n | C_n)$.
- How often is it loaded ?
- How loaded is it ?



Learning example: how loaded is my coin, and how often ?

Reminder:
You are given a (new?) coin for each toss. You toss N times, observing $S = (t, h, \ldots, t)$
Possible outcomes $C_n \in h, t$. Possible coins types: $F_n \in f, l$.

Ques.: How often is the coin loaded?
Ques.: What is $P(F)$ given S?



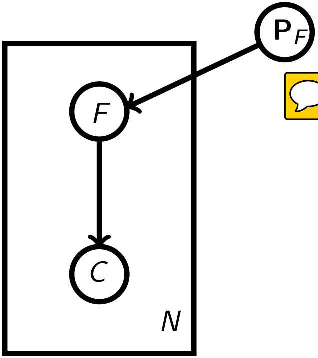
Learning example: how loaded is my coin, and how often ?

Reminder:
You are given a (new?) coin for each toss. You toss N times, observing $S = (t, h, \ldots, t)$
Possible outcomes $C_n \in h, t$. Possible coins types: $F_n \in f, l$.

Ques.: How often is the coin loaded?
Ques.: What is $P(F)$ given S?
$\Rightarrow$ we are *uncertain* about $P(F)$.
$\Rightarrow$ represent $P(F)$ as a random variable.

$$P(F) = \begin{cases} F = f : P_f \\ F = l : P_l \end{cases}$$

$\mathbf{P}_F = (P_f, P_l)$ such that $P_f + P_l = 1$



Learning example: how loaded is my coin, and how often ?

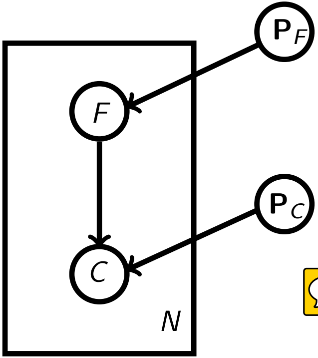
Reminder:
You are given a (new?) coin for each toss. You toss N times, observing $S = (t, h, \ldots, t)$
Possible outcomes $C_n \in h, t$. Possible coins types: $F_n \in f, l$.

Ques.: How often is the coin loaded?
Ques.: What is $P(F)$ given S?
$\Rightarrow$ we are *uncertain* about $P(F)$.
$\Rightarrow$ represent $P(F)$ as a random variable.

$$P(F) = \begin{cases} F = f : P_f \\ F = l : P_l \end{cases}$$

$\mathbf{P}_F = (P_f, P_l)$ such that $P_f + P_l = 1$

Likewise $\mathbf{P}_C = (P_h, P_t)$, $P_h + P_t = 1$



Learning example: how loaded is my coin, and how often ?

Reminder:
You are given a (new?) coin for each toss. You toss N times, observing $S = (t, h, \ldots, t)$
Possible outcomes $C_n \in h, t$. Possible coins types: $F_n \in f, l$.

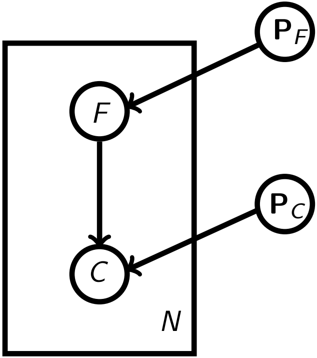
Ques.: How often is the coin loaded?
Ques.: What is $P(F)$ given S?
$\Rightarrow$ we are *uncertain* about $P(F)$.
$\Rightarrow$ represent $P(F)$ as a random variable.

$$P(F) = \begin{cases} F = f : P_f \\ F = l : P_l \end{cases}$$

$\mathbf{P}_F = (P_f, P_l)$ such that $P_f + P_l = 1$

Likewise $\mathbf{P}_C = (P_h, P_t)$, $P_h + P_t = 1$

Reminder:	C_n: data
	F_n: hidden/latent vars.
	$\mathbf{P}_F$,$\mathbf{P}_C$: parameters.

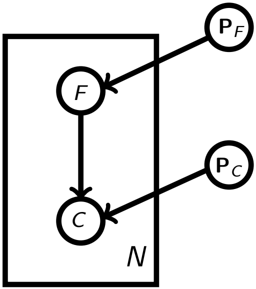


Evaluating the parameter posterior

Reminder:
You are given a (new?) coin for each toss. You toss N times, observing $S = (t, h, \ldots, t)$
Possible outcomes $C_n \in h, t$. Possible coins types: $F_n \in f, l$.
$\mathbf{P}_F$ and $\mathbf{P}_C$ parameterize the distributions of F and $P(C|F = l)$.

Ques.: What is $P(F)$ given S?

Ques.: What is $P(\mathbf{P}_F|S)$?

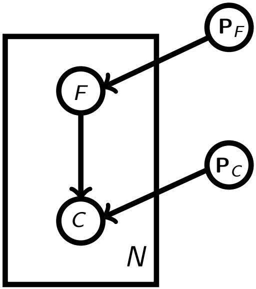


Evaluating the parameter posterior

Reminder:
You are given a (new?) coin for each toss. You toss N times, observing $S = (t, h, \dots, t)$
Possible outcomes $C_n \in h, t$. Possible coins types: $F_n \in f, l$.
$\mathbf{P}_F$ and $\mathbf{P}_C$ parameterize the distributions of F and $P(C|F = l)$.

Ques.: What is $P(F)$ given S?
Ques.: What is $P(\mathbf{P}_F|S)$?
Can we use conditioning:

$$P(\mathbf{P}_F|S) = \frac{P(S|\mathbf{P}_F)P(\mathbf{P}_F)}{P(S)}$$



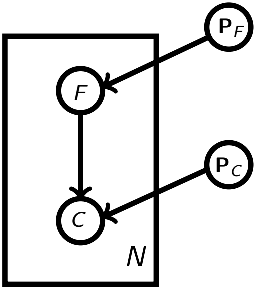
Evaluating the parameter posterior

Reminder:
You are given a (new?) coin for each toss. You toss N times, observing $S = (t, h, \ldots, t)$
Possible outcomes $C_n \in h, t$. Possible coins types: $F_n \in f, l$.
$\mathbf{P}_F$ and $\mathbf{P}_C$ parameterize the distributions of F and $P(C|F = l)$.

Ques.: What is $P(F)$ given S?
Ques.: What is $P(\mathbf{P}_F|S)$?
Can we use conditioning:

$$P(\mathbf{P}_F|S) = \frac{P(S|\mathbf{P}_F)P(\mathbf{P}_F)}{P(S)}$$

But: $\mathbf{P}_F$ is a *continuous* random variable.
What is $P(\mathbf{P}_F)$? Does it exist?



Digression: continuous random variables

Question: What is $P(X)$, where $X \in \mathbb{R}$ is a continuous RV?

Digression: continuous random variables

Question: What is $P(X)$, where $X \in \mathbb{R}$ is a continuous RV?

We know: **Probability distribution**

Definition: Let Y be a (discrete) random variable with range Z. A *probability distribution* is a function $P : Z \to [0, 1]$ such that $\sum_{y \in Z} P(Y = y) = 1$.

Digression: continuous random variables

Question: What is $P(X)$, where $X \in \mathbb{R}$ is a continuous RV?

We know: **Probability distribution**

Definition: Let Y be a (discrete) random variable with range Z. A *probability distribution* is a function $P : Z \to [0, 1]$ such that $\sum_{y \in Z} P(Y = y) = 1$.

Probability density (informal definition)

Definition: Let X be a (continuous) random variable whose range is an interval $Z \subseteq \mathbb{R}$. A *probability density* $p(X)$ is a function $p : Z \to \mathbb{R}$ such that

1. $p(X) \geq 0$
2. $P(x_0 \leq X < x_0 + dx) = p(x_0)\, dx$
3. $\int_Z dx\, p(x) = 1$

Digression: continuous random variables

Question: What is $P(X)$, where $X \in \mathbb{R}$ is a continuous RV?

We know: **Probability distribution**

Definition: Let Y be a (discrete) random variable with range Z. A *probability distribution* is a function $P : Z \to [0, 1]$ such that $\sum_{y \in Z} P(Y = y) = 1$.

Probability density (informal definition)

Definition: Let X be a (continuous) random variable whose range is an interval $Z \subseteq \mathbb{R}$. A *probability density* $p(X)$ is a function $p : Z \to \mathbb{R}$ such that

1. $p(X) \geq 0$
2. $P(x_0 \leq X < x_0 + dx) = p(x_0)\, dx$
3. $\int_Z dx\, p(x) = 1$

Note: I will use a lowercase $p(Y)$ for a density, an uppercase $P(Y)$ for a distribution.

Probability densities

Probability density (informal definition)

Definition: Let X be a (continuous) random variable whose range is an interval $Z \subseteq \mathbb{R}$. A *probability density* $p(X)$ is a function $p : Z \to \mathbb{R}$ such that

1. $p(X) \geq 0$
2. $P(x_0 \leq X < x_0 + dx) = p(x_0)\, dx$
3. $\int_Z dx\, p(x) = 1$

Notes:

- $p(X)$ does **not** have to be ≤ 1.
- $p(x_0)\, dx$ is the probability that $x \in [x_0, x_0 + dx)$ where dx is infinitesimally small and > 0.
- $P(X \in [a, b]) = \int_a^b dx\, p(x)$

Probability densities

Probability density (informal definition)

Definition: Let X be a (continuous) random variable whose range is an interval $Z \subseteq \mathbb{R}$. A *probability density* $p(X)$ is a function $p : Z \to \mathbb{R}$ such that

1. $p(X) \geq 0$
2. $P(x_0 \leq X < x_0 + dx) = p(x_0)\, dx$
3. $\int_Z dx\, p(x) = 1$

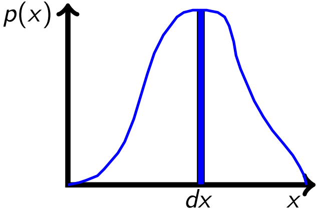
Notes:

- $p(X)$ does **not** have to be ≤ 1.
- $p(x_0)\, dx$ is the probability that $x \in [x_0, x_0 + dx)$ where dx is infinitesimally small and > 0.
- $P(X \in [a, b]) = \int_a^b dx\, p(x)$
- Not all continuous random variables have a density.
- There is a proper measure-theoretic definition of probability densities.

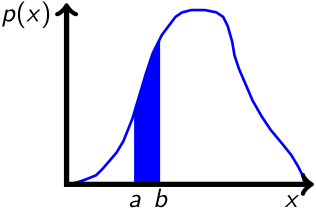
Probability density: graphical interpretation

Probability density: graphical interpretation

$$P(x_0 \leq X < x_0 + dx) = p(x_0)\, dx$$



$$P(X \in [a, b]) = \int_a^b dx\, p(x)$$

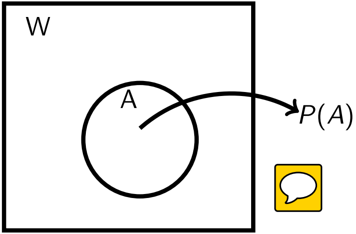


The Radon-Nikodym theorem

Definition: A *probability space* is a tuple $(W, \mathcal{F}, P)$, where $\mathcal{F}$ is a σ-algebra over W and *probability measure* $P : \mathcal{F} \to [0, 1]$, with the properties:

P1 $P(W) = 1$

P2 If $U, V \in \mathcal{F}$ and $U \cap V = \emptyset$, then $P(U \cup V) = P(U) + P(V)$.



- W may have *continuously many* elements, e.g. points in a 2D plane
- $A \in \mathcal{F}$: measurable set

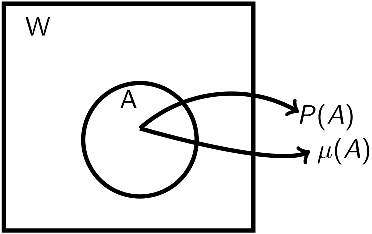
The Radon-Nikodym theorem

Reminder:
Definition: A *probability space* is a tuple $(W, \mathcal{F}, P)$, where $\mathcal{F}$ is a σ-algebra over W and *probability measure* $P : \mathcal{F} \to [0, 1]$, with the properties: (P1) $P(W) = 1$; (P2) If $U, V \in \mathcal{F}$ and $U \cap V = \emptyset$, then $P(U \cup V) = P(U) + P(V)$.

Definition: a *measure* is a function $\mu : \mathcal{F} \to \mathbb{R}_0^+$. The elements of the σ-algebra $\mathcal{F}$ are the measurable sets. μ has the properties

M1 $P(\emptyset) = 0$

M2 If $U, V \in \mathcal{F}$ and $U \cap V = \emptyset$, then $P(U \cup V) = P(U) + P(V)$.



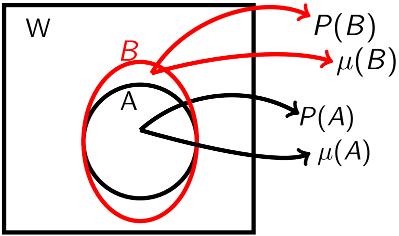
- $A \in \mathcal{F}$: measurable sets
- μ some measure on $\mathcal{F}$, e.g. Lebesgue measure (volume)
- $\mu(A)$: e.g. volume of A

The Radon-Nikodym theorem

Reminder:

Definition: A *probability space* is a tuple $(W, \mathcal{F}, P)$, where $\mathcal{F}$ is a σ-algebra over W and *probability measure* $P : \mathcal{F} \to [0, 1]$, with the properties: (P1) $P(W) = 1$; (P2) If $U, V \in \mathcal{F}$ and $U \cap V = \emptyset$, then $P(U \cup V) = P(U) + P(V)$.

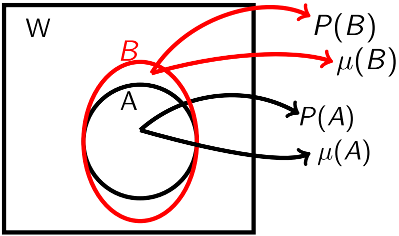
Definition: a *measure* is a function $\mu : \mathcal{F} \to \mathbb{R}_0^+$. The elements of the σ-algebra $\mathcal{F}$ are the measurable sets. μ has the properties (M1) $P(\emptyset) = 0$; (M2) If $U, V \in \mathcal{F}$ and $U \cap V = \emptyset$, then $P(U \cup V) = P(U) + P(V)$.



- $A, B \in \mathcal{F}$: measurable set
- μ Lebesgue measure (volume)
- $\mu(A), \mu(B)$: e.g. volumes of A, B
- $dA = B - A = B \cap \bar{A}$
- $dP = P(B) - P(A) = P(dA)$
- $d\mu = \mu(B) - \mu(A) = \mu(dA)$

The Radon-Nikodym theorem

Reminder:
Definition: A *probability space* is a tuple $(W, \mathcal{F}, P)$, where $\mathcal{F}$ is a σ-algebra over W and *probability measure* $P : \mathcal{F} \rightarrow [0, 1]$, with the properties: (P1) $P(W) = 1$; (P2) If $U, V \in \mathcal{F}$ and $U \cap V = \emptyset$, then $P(U \cup V) = P(U) + P(V)$.
Definition: a *measure* is a function $\mu : \mathcal{F} \rightarrow \mathbb{R}_0^+$. The elements of the σ-algebra $\mathcal{F}$ are the measurable sets. μ has the properties (M1) $P(\emptyset) = 0$; (M2) If $U, V \in \mathcal{F}$ and $U \cap V = \emptyset$, then $P(U \cup V) = P(U) + P(V)$.



- $A, B \in \mathcal{F}$: measurable set
- μ Lebesgue measure (volume)
- $\mu(A), \mu(B)$: e.g. volumes of A, B
- $dA = B - A = B \cap \bar{A}$
- $dP = P(B) - P(A) = P(dA)$
- $d\mu = \mu(B) - \mu(A) = \mu(dA)$

$dP, d\mu \geq 0$ follows from definition of measure
Question: is there a *direct* connection between P and μ ?

The Radon-Nikodym theorem

Reminder:
Definition: A *probability space* is a tuple $(W, \mathcal{F}, P)$, where $\mathcal{F}$ is a σ-algebra over W and *probability measure* $P : \mathcal{F} \to [0, 1]$, with the properties: (P1) $P(W) = 1$; (P2) If $U, V \in \mathcal{F}$ and $U \cap V = \emptyset$, then $P(U \cup V) = P(U) + P(V)$.
Definition: a *measure* is a function $\mu : \mathcal{F} \to \mathbb{R}_0^+$. The elements of the σ-algebra $\mathcal{F}$ are the measurable sets. μ has the properties (M1) $P(\emptyset) = 0$; (M2) If $U, V \in \mathcal{F}$ and $U \cap V = \emptyset$, then $P(U \cup V) = P(U) + P(V)$.

Radon-Nikodym theorem: given a measurable space $(W, \mathcal{F})$ and measures P, μ where P is *absolutely continuous* with respect to μ ($\mu(A) = 0 \Rightarrow P(A) = 0$), there exists (measurable) function $f : W \to \mathbb{R}_0^+$ such that for all $A \in \mathcal{F}$, $a \in A$

$$P(A) = \int_A f(a) d\mu$$

where f is called the *Radon-Nikodym derivate* of P w.r.t. μ and is denoted as $\frac{dP}{d\mu}$

- If P is a probability measure, then f is a *probability density*, which we will denote with (lowercase) p
- 1D-Continuous random variables are functions $X : W \to \mathbb{R}$

Probability densities of multivariate random variables

Probability density (informal generalization)

Definition: Let X be a continuous, multivariate random variable whose range is a volume $Z \subseteq \mathbb{R}^n$. A *probability density* is a function $p : Z \to \mathbb{R}$ such that

1. $p(\mathbf{x}) \geq 0$
2. $P(X \in d\mathbf{x}(\mathbf{x}_0)) = p(\mathbf{x}_0)\, d\mathbf{x}$
3. $\int_Z d\mathbf{x}\, p(\mathbf{x}) = 1$

Notes:

- For our purposes, think of Z as a n-dimensional cuboid.

Conditioning, marginalization and chain rules for densities

Key properties of probability distributions also apply to densities.
Let X and Y be continuous random variables. Then:

- Joint density of X, Y: $p(X, Y)$.
- Marginal density: $p(X) = \int dy \, p(X, y)$
- Conditional density: $p(X|Y) = \frac{p(X,Y)}{p(X)}$

Conditioning, marginalization and chain rules for densities

Key properties of probability distributions also apply to densities.
Let X and Y be continuous random variables. Then:

- Joint density of X, Y: $p(X, Y)$.
- Marginal density: $p(X) = \int dy\, p(X, y)$
- Conditional density: $p(X|Y) = \frac{p(X,Y)}{p(X)}$
- Bayes' rule: $p(X|Y) = p(Y|X)\frac{p(X)}{p(Y)}$
- Independence btw. X and Y iff $p(X, Y) = p(X)p(Y)$
- Chain rule: $p(X, Y, Z) = p(X|Y, Z)p(Y|Z)p(Z)$.

Conditioning, marginalization and chain rules for densities

Key properties of probability distributions also apply to densities.
Let X and Y be continuous random variables. Then:

- Joint density of X, Y: $p(X, Y)$.
- Marginal density: $p(X) = \int dy\, p(X, y)$
- Conditional density: $p(X|Y) = \frac{p(X,Y)}{p(X)}$
- Bayes' rule: $p(X|Y) = p(Y|X)\frac{p(X)}{p(Y)}$
- Independence btw. X and Y iff $p(X, Y) = p(X)p(Y)$
- Chain rule: $p(X, Y, Z) = p(X|Y, Z)p(Y|Z)p(Z)$.

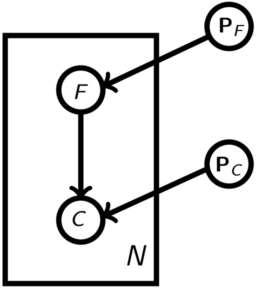
Note: densities and probability distribution can appear together in one expression.

Evaluating the parameter posterior, contd.

Reminder:
You are given a (new?) coin for each toss. You toss N times, observing $S = (t, h, \ldots, t)$
Possible outcomes $C_n \in h, t$. Possible coins types: $F_n \in f, l$.
$\mathbf{P}_F$ and $\mathbf{P}_C$ parameterize the distributions of F and $P(C|F = l)$. If $X \in Z$ is continuous, then $p(X)$ is a density, if $p(X) \geq 0$ and $\int_Z d\vec{x}\, p(\vec{x}) = 1$.

Ques.: What is $P(F)$ given S?

Ques.: Density $p(\mathbf{P}_F|S)$?



Evaluating the parameter posterior, contd.

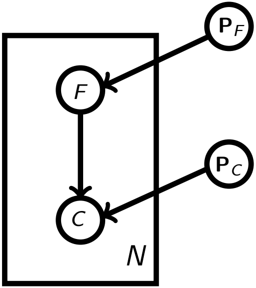
Reminder:
You are given a (new?) coin for each toss. You toss N times, observing $S = (t, h, \ldots, t)$
Possible outcomes $C_n \in h, t$. Possible coins types: $F_n \in f, l$.
$\mathbf{P}_F$ and $\mathbf{P}_C$ parameterize the distributions of F and $P(C|F = l)$. If $X \in Z$ is continuous, then $p(X)$ is a density, if $p(X) \geq 0$ and $\int_Z d\vec{x}\, p(\vec{x}) = 1$.

Ques.: What is $P(F)$ given S?
Ques.: Density $p(\mathbf{P}_F|S)$?
We can use conditioning:

$$p(\mathbf{P}_F|S) = \frac{P(S|\mathbf{P}_F)p(\mathbf{P}_F)}{P(S)}$$

Likewise, for $\mathbf{P}_C$

$$p(\mathbf{P}_C|S) = \frac{P(S|\mathbf{P}_C)p(\mathbf{P}_C)}{P(S)}$$



Evaluating the parameter posterior, contd.

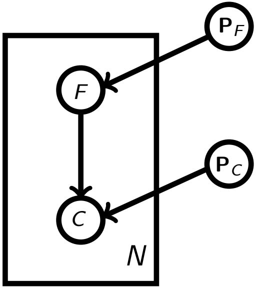
Reminder:
You are given a (new?) coin for each toss. You toss N times, observing $S = (t, h, \ldots, t)$
Possible outcomes $C_n \in h, t$. Possible coins types: $F_n \in f, l$.
$\mathbf{P}_F$ and $\mathbf{P}_C$ parameterize the distributions of F and $P(C|F = l)$. If $X \in Z$ is continuous, then $p(X)$ is a density, if $p(X) \geq 0$ and $\int_Z d\vec{x}\, p(\vec{x}) = 1$.

Ques.: What is $P(F)$ given S?
Ques.: Density $p(\mathbf{P}_F|S)$?
We can use conditioning:

$$p(\mathbf{P}_F|S) = \frac{P(S|\mathbf{P}_F)p(\mathbf{P}_F)}{P(S)}$$

Likewise, for $\mathbf{P}_C$

$$p(\mathbf{P}_C|S) = \frac{P(S|\mathbf{P}_C)p(\mathbf{P}_C)}{P(S)}$$

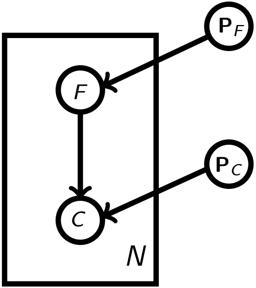


How can we evaluate the marginals? Can we use the sum-product algorithm?

Evaluating the parameter posterior, contd.

Reminder:
You are given a (new?) coin $\in \{f, l\}$ for each toss $\in \{h, t\}$. You toss N times, observing $S = (t, h, \ldots, t)$
$P(F) = \mathbf{P}_F$ and $P(C|F = l) = \mathbf{P}_C$. We want: posterior densitites $p(\mathbf{P}_F|S)$ and $p(\mathbf{P}_C|S)$.

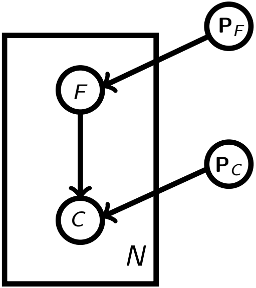
Bayesian network



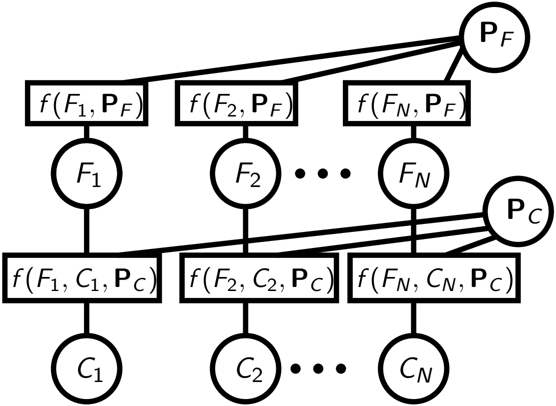
Evaluating the parameter posterior, contd.

Reminder:
You are given a (new?) coin $\in \{f, l\}$ for each toss $\in \{h, t\}$. You toss N times, observing $S = (t, h, \ldots, t)$
$P(F) = \mathbf{P}_F$ and $P(C|F = l) = \mathbf{P}_C$. We want: posterior densitites $p(\mathbf{P}_F|S)$ and $p(\mathbf{P}_C|S)$.

Bayesian network



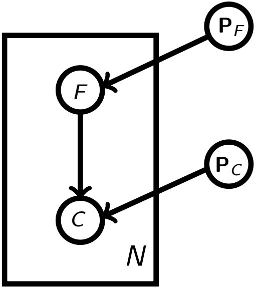
Factor graph



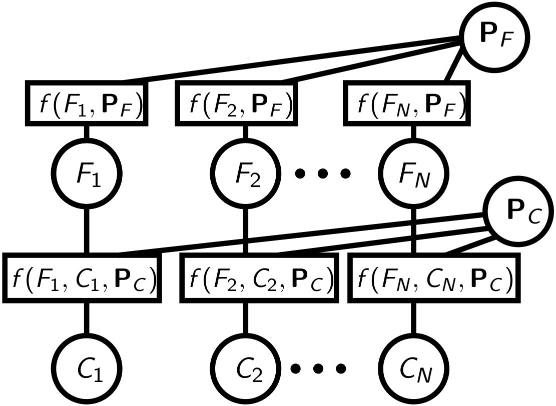
Evaluating the parameter posterior, contd.

Reminder:
You are given a (new?) coin $\in \{f, l\}$ for each toss $\in \{h, t\}$. You toss N times, observing $S = (t, h, \ldots, t)$
$P(F) = \mathbf{P}_F$ and $P(C|F = l) = \mathbf{P}_C$. We want: posterior densitites $p(\mathbf{P}_F|S)$ and $p(\mathbf{P}_C|S)$.

Bayesian network



Factor graph



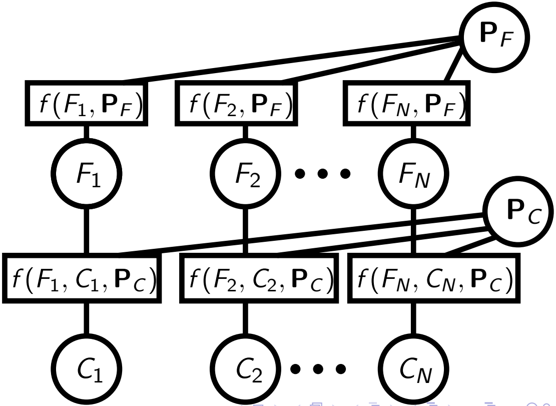
$\Rightarrow$ lots of loops!

We need another approach!

Reminder:
You are given a (new?) coin $\in \{f, l\}$ for each toss $\in \{h, t\}$. You toss N times, observing $S = (t, h, \ldots, t)$
$P(F) = \mathbf{P}_F$ and $P(C|F = l) = \mathbf{P}_C$. We want: posterior densitites $p(\mathbf{P}_F|S)$ and $p(\mathbf{P}_C|S)$.

Prob. 1: loopy graph.

Factor graph



We need another approach!

Reminder:
You are given a (new?) coin $\in \{f, l\}$ for each toss $\in \{h, t\}$. You toss N times, observing $S = (t, h, \dots, t)$
$P(F) = \mathbf{P}_F$ and $P(C|F = l) = \mathbf{P}_C$. We want: posterior densitites $p(\mathbf{P}_F|S)$ and $p(\mathbf{P}_C|S)$.

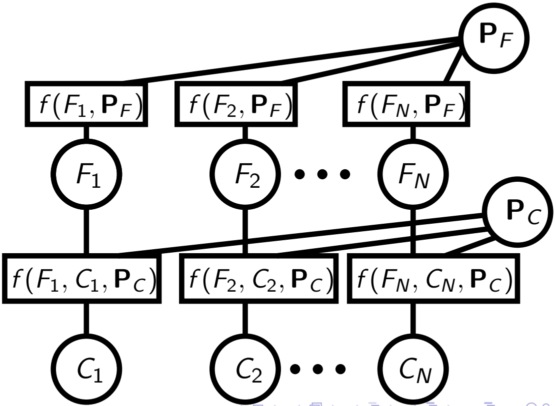
Prob. 1: loopy graph.
Prob. 2: Messages from factors to $\mathbf{P}_C$:

$$\mu_{f(F_1,\mathbf{P}_F)\to\mathbf{P}_F}(\mathbf{P}_F)$$

are *infinitely* long, because $\mathbf{P}_F \in \mathbb{R}^2$.

actually, $\mathbf{P}_F \in [0, 1]$, but still...

Factor graph



We need another approach!

Reminder:
You are given a (new?) coin $\in \{f, l\}$ for each toss $\in \{h, t\}$. You toss N times, observing $S = (t, h, \ldots, t)$
$P(F) = \mathbf{P}_F$ and $P(C|F = l) = \mathbf{P}_C$. We want: posterior densitites $p(\mathbf{P}_F|S)$ and $p(\mathbf{P}_C|S)$.

Prob. 1: loopy graph.
Prob. 2: Messages from factors to $\mathbf{P}_C$:

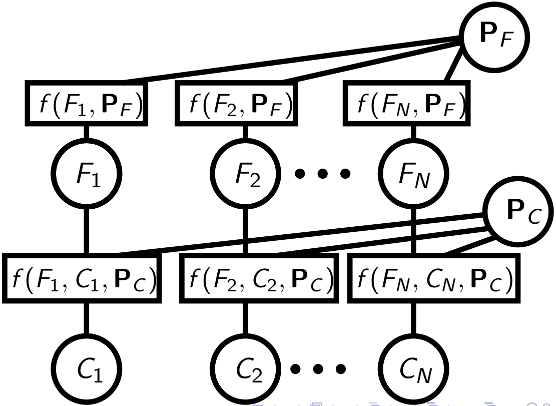
$$\mu_{f(F_1,\mathbf{P}_F)\to\mathbf{P}_F}(\mathbf{P}_F)$$

are *infinitely* long, because $\mathbf{P}_F \in \mathbb{R}^2$.

actually, $\mathbf{P}_F \in [0, 1]$, but still...

$\Rightarrow$ We need a different approach!

Factor graph



Restating the problem: inference as optimization

Reminder:
You are given a (new?) coin $\in \{f, l\}$ for each toss $\in \{h, t\}$. You toss N times, observing $S = (t, h, \ldots, t)$
$P(F) = \mathbf{P}_F$ and $P(C|F = l) = \mathbf{P}_C$. We want: posterior densitites $p(\mathbf{P}_F|S)$ and $p(\mathbf{P}_C|S)$.

We would like : $p(\mathbf{P}_F|S)$ and $p(\mathbf{P}_C|S)$
Direct approach too difficult. Instead, we will

- restate inference as an **optimization problem** with $p(\mathbf{P}_F|S)$ and $p(\mathbf{P}_C|S)$ as solution, and
- constrain the solutions until we can solve it.
- No longer exact, but at least an *approximate solution* may be possible.

Restating the problem: inference as optimization

Reminder:
You are given a (new?) coin $\in \{f, l\}$ for each toss $\in \{h, t\}$. You toss N times, observing $S = (t, h, \dots, t)$
$P(F) = \mathbf{P}_F$ and $P(C|F = l) = \mathbf{P}_C$. We want: posterior densitites $p(\mathbf{P}_F|S)$ and $p(\mathbf{P}_C|S)$.

We would like : $p(\mathbf{P}_F|S)$ and $p(\mathbf{P}_C|S)$
Direct approach too difficult. Instead, we will

- restate inference as an **optimization problem** with $p(\mathbf{P}_F|S)$ and $p(\mathbf{P}_C|S)$ as solution, and
- constrain the solutions until we can solve it.
- No longer exact, but at least an *approximate solution* may be possible.

What should be optimized?

Question: what is a good model for the data?
Answer: marginal probability $P(S)$ is high.
Reason: good explanation for observations, if observations are probable.

Restating the problem: inference as optimization

General optimization problem statement:

$$\text{maximize } P(D) = \sum_H P(D,H)$$

where

- D are observable data.
- H are hidden variables/parameters
- $P(D,H) = P(D|H)P(H)$
- either of D or H could be continuous, in which case we work with densitites (and integrals).

Restating the problem: inference as optimization

General optimization problem statement:

$$\text{maximize } P(D) = \sum_H P(D, H)$$

where

- D are observable data.
- H are hidden variables/parameters
- $P(D, H) = P(D|H)P(H)$
- either of D or H could be continuous, in which case we work with densitites (and integrals).

Problem: Since $P(H|D) = \frac{P(D,H)}{P(D)}$ is too difficult to evaluate directly
$\Rightarrow P(D)$ is too difficult to evaluate directly.

Restating the problem: inference as optimization

General optimization problem statement:

$$\text{maximize } P(D) = \sum_H P(D, H)$$

where

- D are observable data.
- H are hidden variables/parameters
- $P(D, H) = P(D|H)P(H)$
- either of D or H could be continuous, in which case we work with densitites (and integrals).

Problem: Since $P(H|D) = \frac{P(D,H)}{P(D)}$ is too difficult to evaluate directly

$\Rightarrow P(D)$ is too difficult to evaluate directly.

Question 1: What do we maximize instead?

Question 2: With respect to what should $P(D)$ be maximized?

Making the optimization problem tractable

Reminder:
Given: observable data D and hidden variables/parameters H.
We'd like to maximize $P(D) = \sum_H P(D, H) = \sum_H P(D|H)P(H)$.
Either of D or H could be continuous $\Rightarrow$ densities and integrals instead of sums and distributions.

Question 1: if $P(D)$ is too difficult to evaluate, what do we maximize instead?

Answer 1: maximum-a-posteriori, MAP:

$$\text{maximize } P(D, H) = P(D|H)P(H)$$

$\Rightarrow$ ignore all summands in $P(D) = \sum_H P(D|H)P(H)$ except for the largest one.

Making the optimization problem tractable

Reminder:
Given: observable data D and hidden variables/parameters H.
We'd like to maximize $P(D) = \sum_H P(D, H) = \sum_H P(D|H)P(H)$.
Either of D or H could be continuous $\Rightarrow$ densities and integrals instead of sums and distributions.

Question 1: if $P(D)$ is too difficult to evaluate, what do we maximize instead?
Instead of

$$\text{maximize } P(D, H)$$

we can maximize any strictly monotonically increasing function of $P(D, H)$. Popular choice:

$$\log(P(D, H))$$

Why $\log()$**:**

- avoid underflows.
- simplify functional form.
- information-theoretical reasons, minimal codelength of data.

Making the optimization problem tractable

Reminder:
Given: observable data D and hidden variables/parameters H.
We'd like to maximize $P(D) = \sum_H P(D, H) = \sum_H P(D|H)P(H)$.
Either of D or H could be continuous $\Rightarrow$ densities and integrals instead of sums and distributions.

Question 1: if $P(D)$ is too difficult to evaluate, what do we maximize instead?
Answer 2: maximum-likelihood, ML:

$$\text{maximize } P(D|H)$$

$\Rightarrow$ ignore prior $P(H)$ and all summands in $P(D) = \sum_H P(D|H)P(H)$ except for the largest one.

In practical applications:

- maximize $\log(P(D|H))$
- or minimize $-\log(P(D|H))$

Deriving a lower bound on $P(D)$

Reminder:
Given: observable data D and hidden variables/parameters H.
We'd like to maximize $P(D) = \sum_H P(D, H) = \sum_H P(D|H)P(H)$, or alternatively $\log(P(D))$
Either of D or H could be continuous $\Rightarrow$ densities and integrals instead of sums and distributions.

Question 1: if $P(D)$ is too difficult to evaluate, what do we maximize instead?
Answer 3: variational approximation derive a *lower bound* $\mathcal{L}(D)$ on $\log(P(D))$, and maximize this bound.

$$\log(P(D)) \geq \mathcal{L}(D)$$
$$\text{maximize } \mathcal{L}(D)$$

Deriving a lower bound on $P(D)$

Reminder:
Given: observable data D and hidden variables/parameters H.
We'd like to maximize $P(D) = \sum_H P(D, H) = \sum_H P(D|H)P(H)$, or alternatively $\log(P(D))$
Either of D or H could be continuous $\Rightarrow$ densities and integrals instead of sums and distributions.

Question 1: if $P(D)$ is too difficult to evaluate, what do we maximize instead?
Answer 3: **variational approximation** derive a *lower bound* $\mathcal{L}(D)$ on $\log(P(D))$, and maximize this bound.

$$\log(P(D)) \geq \mathcal{L}(D)$$
$$\text{maximize } \mathcal{L}(D)$$

Question 2: With respect to what should $\mathcal{L}(D)$ be maximized?
Answer: We want to compute an approximating distribution to $P(H|D)$, call it $Q(H) \Rightarrow \mathcal{L}(D)$ should also depend on $Q(H)$.

$$\text{maximize } \mathcal{L}(D, Q(H)) \text{ with respect to } Q(H).$$

Deriving a lower bound on $P(D)$

Reminder:
Given: observable data D and hidden variables/parameters H.
We'd like to maximize a lower bound $\mathcal{L}(D, Q(H))$ on $P(D)$: $P(D) \geq \mathcal{L}(D, Q(H))$ with respect to $Q(H)$.
Either of D or H could be continuous $\Rightarrow$ densities and integrals instead of sums and distributions.

(New) optimization problem:

$$\text{maximize} \quad \mathcal{L}(D, Q(H)) \quad \text{with respect to} \quad Q(H).$$

To derive $\mathcal{L}(D, Q(H))$, we need two ingredients:

- Jensen's inequality for convex (or concave) functions
- Kullback-Leibler divergence, or relative entropy.

Jensen's inequality for convex (or concave) functions

Definition: a function $f(x)$ is *convex* over an interval $[a, b]$ if $\forall 0 \leq \lambda \leq 1$ and $\forall x_1, x_2 \in [a, b]$:

$$f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda)f(x_2)$$

Jensen's inequality for convex (or concave) functions

Definition: a function $f(x)$ is *convex* over an interval $[a, b]$ if $\forall 0 \leq \lambda \leq 1$ and $\forall x_1, x_2 \in [a, b]$:

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda)f(x_2)$$

Definition: a function $f(x)$ is *concave* if $-f(x)$ is convex. Then

$$f(\lambda x_1 + (1-\lambda)x_2) \geq \lambda f(x_1) + (1-\lambda)f(x_2)$$

Jensen's inequality for convex (or concave) functions

Definition: a function $f(x)$ is *convex* over an interval $[a, b]$ if $\forall 0 \le \lambda \le 1$ and $\forall x_1, x_2 \in [a, b]$:

$$f(\lambda x_1 + (1-\lambda)x_2) \le \lambda f(x_1) + (1-\lambda)f(x_2)$$

Definition: a function $f(x)$ is *concave* if $-f(x)$ is convex. Then

$$f(\lambda x_1 + (1-\lambda)x_2) \ge \lambda f(x_1) + (1-\lambda)f(x_2)$$

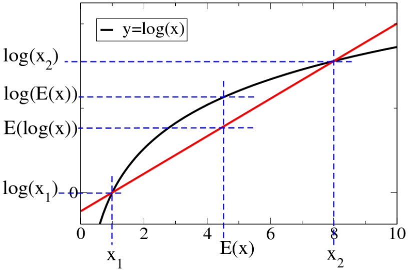
Example: $\log(x)$ is concave.

Let $P(X) = \begin{cases} X = x_1 : \lambda \\ X = x_2 : (1-\lambda) \end{cases}$

Then

$$f(E(X)) \ge E(f(X))$$

.



Jensen's inequality for convex functions

Reminder:
Let $X \in \{x_1, x_2\}$ be a random variable with distribution $P(X)$.
Function $f(x)$ is convex (concave) , if $f(E(X)) \leq (\geq) E(f(X))$.

Jensen's inequality: let $f(x)$ be a convex function, and $X \in \{x_1, \ldots, x_N\}$ be a random variable with distribution $P(X)$.
Then

$$f(E(X)) \leq E(f(X))$$

Proof: by induction over N.
The inequality holds for $N = 2$. Let $P(X = x_i) = p_i$.
Since $\sum_{i=1}^{N-1} p_i = (1 - p_N)$, $P(X = x_i) = \frac{p_i}{1-P_N} = p_i'$ is a probability distribution over $X' \in \{x_1, \ldots, x_{N-1}\}$

Jensen's inequality for convex functions, contd.

Reminder:
Let $X \in \{x_1, x_2\}$ be a random variable with distribution $P(X)$.
Function $f(x)$ is convex (concave) , if $f(E(X)) \leq (\geq) E(f(X))$.
If $P(X = x_i) = p_i$ for $X \in \{x_1, \dots, X_N\}$, then $P(X' = x_i) = p'_i = \frac{p_i}{1-p_N}$ is a distribution over $X \in \{x_1, \dots, X_{N-1}\}$

$$
\begin{aligned}
E(f(X)) = \sum_{i=1}^{N} p_i f(x_i) &= p_N f(x_N) + \sum_{i=1}^{N-1} p_i f(x_i) \\
&= p_N f(x_N) + (1 - p_N) \sum_{i=1}^{N-1} p'_i f(x_i) \\
&\geq p_N f(x_N) + (1 - p_N) f\left(\sum_{i=1}^{N-1} p'_i x_i\right) \\
&\geq f\left(p_N x_N + (1 - p_N) \sum_{i=1}^{N-1} p'_i x_i\right) \\
&= f\left(\sum_{i=1}^{N} x_i\right) = f(E(X)) \quad \square
\end{aligned}
$$

Jensen's inequality for convex functions

Jensen's inequality: let $f(x)$ be a convex function, and $X \in \{x_1, \ldots, x_N\}$ be a random variable with distribution $P(X)$. Then

$$f(E(X)) \leq E(f(X))$$

Moreover, if $f(x)$ is *strictly convex*, then $f(E(X)) = E(f(X))$ implies that X is constant.

Likewise, if $f(x)$ is a concave function, then

$$f(E(X)) \geq E(f(X))$$

Note: Jensen's inequality also holds for continuous random variables!

Kullback-Leibler divergence

Reminder:
Jensen's inequality: let $f(x)$ be a convex (concave) function, and $X \in \{x_1, \ldots, x_N\}$ be a random variable with distribution $P(X)$. Then $f(E(X)) \leq (\geq) E(f(X))$.

Definition: Let $Q(X)$ and $P(X)$ be probability distributions over X. The *Kullback-Leibler divergence* or *relative entropy* is given by

$$D(Q||P) = \sum_X Q(x) \log \left(\frac{Q(x)}{P(x)} \right)$$

Important property: $D(Q||P) \geq 0$ with equality if $Q(X) = P(X)$.
Note: $D(Q||P)$ is not symmetric.

Proof: Kullback-Leibler divergence is non-negative

Reminder:
Jensen's inequality: let $f(x)$ be a convex (concave) function, and $X \in \{x_1, \ldots, x_N\}$ be a random variable with distribution $P(X)$. Then $f(E(X)) \leq (\geq) E(f(X))$.

Important property: $D(Q||P) \geq 0$ with equality if $Q(X) = P(X)$.
Proof:

$$
\begin{aligned}
-D(Q||P) &= -\sum_X Q(x) \log\left(\frac{Q(x)}{P(x)}\right) = \sum_X Q(X) \log\left(\frac{P(x)}{Q(x)}\right) \\
&= E_Q\left(\log\left(\frac{P(X)}{Q(X)}\right)\right) \leq \log\left(E_Q\left(\frac{P(X)}{Q(X)}\right)\right) \\
&= \log\left(\sum_X P(x)\right) = \log(1) = 0 \quad \square
\end{aligned}
$$

Note: $D(P||Q) = 0$ if (and only if) $P(X) = Q(X)$.

Deriving a lower bound on $P(D)$

Reminder:

Given: observable data D and hidden variables/parameters H.

We'd like to maximize a lower bound $\mathcal{L}(D, Q(H))$ on $P(D)$: $P(D) \geq \mathcal{L}(D, Q(H))$ with respect to $Q(H)$.

Kullback-Leibler divergence: $D(Q||P) = E_Q\left(\log\left(\frac{Q(X)}{P(X)}\right)\right)$

Jensen's inequality: $\log(E(X)) \geq E(\log(X))$

We derive a lower bound $\mathcal{L}(D, Q(H))$ on $\log(P(D))$ via:

$$
\begin{aligned}
\log(P(D)) &= \log\left(\sum_H P(D,H)\right) = \log\left(\sum_H \frac{Q(H)}{Q(H)} P(D,H)\right) \\
&= \log\left(\sum_H Q(H)\frac{P(D,H)}{Q(H)}\right) = \log\left(E_{Q(H)}\left(\frac{P(D,H)}{Q(H)}\right)\right) \\
&\geq E_{Q(H)}\left(\log\left(\frac{P(D,H)}{Q(H)}\right)\right) = \sum_H Q(H)\left(\log\left(\frac{P(D,H)}{Q(H)}\right)\right) \\
&=: \mathcal{L}(D, Q(H))
\end{aligned}
$$

Deriving a lower bound on $P(D)$

Reminder:
Given: observable data D and hidden variables/parameters H.
We'd like to maximize a lower bound $\mathcal{L}(D, Q(H))$ on $P(D)$: $P(D) \geq \mathcal{L}(D, Q(H))$ with respect to $Q(H)$.
Kullback-Leibler divergence: $D(Q||P) = E_Q\left(\log\left(\frac{Q(X)}{P(X)}\right)\right)$
Jensen's inequality: $\log(E(X)) \geq E(\log(X))$

We derive a lower bound $\mathcal{L}(D, Q(H))$ on $\log(P(D))$ via:

$$\begin{aligned}
\log(P(D)) &= \log\left(\sum_H P(D,H)\right) = \log\left(\sum_H \frac{Q(H)}{Q(H)} P(D,H)\right) \\
&= \log\left(\sum_H Q(H) \frac{P(D,H)}{Q(H)}\right) = \log\left(E_{Q(H)}\left(\frac{P(D,H)}{Q(H)}\right)\right) \\
&\geq E_{Q(H)}\left(\log\left(\frac{P(D,H)}{Q(H)}\right)\right) = \sum_H Q(H)\left(\log\left(\frac{P(D,H)}{Q(H)}\right)\right) \\
&=: \mathcal{L}(D, Q(H))
\end{aligned}$$

Those are the **key steps** in constructing the lower bound. Works with densities, too!

When is $\mathcal{L}(D, Q(H))$ tight?

Reminder:

Given: observable data D and hidden variables/parameters H.

We'd like to maximize a lower bound $\mathcal{L}(D, Q(H))$ on $P(D)$: $P(D) \geq \mathcal{L}(D, Q(H))$ with respect to $Q(H)$.

Kullback-Leibler divergence: $D(Q||P) = E_Q\left(\log\left(\frac{Q(X)}{P(X)}\right)\right)$. Jensen's inequality: $\log(E(X)) \geq E(\log(X))$

Lower bound $\mathcal{L}(D, Q(H)) = E_{Q(H)}\left(\log\left(\frac{P(D,H)}{Q(H)}\right)\right) \leq \log(P(D))$

Ques.: When is $\mathcal{L}(D, Q(H)) = \log(P(D))$?

When is $\mathcal{L}(D, Q(H))$ tight?

Reminder:
Given: observable data D and hidden variables/parameters H.
We'd like to maximize a lower bound $\mathcal{L}(D, Q(H))$ on $P(D)$: $P(D) \geq \mathcal{L}(D, Q(H))$ with respect to $Q(H)$.
Kullback-Leibler divergence: $D(Q||P) = E_Q\left(\log\left(\frac{Q(X)}{P(X)}\right)\right)$. Jensen's inequality: $\log(E(X)) \geq E(\log(X))$
Lower bound $\mathcal{L}(D, Q(H)) = E_{Q(H)}\left(\log\left(\frac{P(D,H)}{Q(H)}\right)\right) \leq \log(P(D))$

Ques.: When is $\mathcal{L}(D, Q(H)) = \log(P(D))$?
Answer: $P(D, H) = P(H|D)P(D)$. Thus, if $Q(H) = P(H|D)$:

$$\begin{aligned}
\mathcal{L}(D, Q(H)) &= E_{Q(H)}\left(\log\left(\frac{P(H|D)P(D)}{Q(H)}\right)\right) \\
&= E_{Q(H)}\left(\log\left(\frac{P(H|D)P(D)}{P(H|D)}\right)\right) \\
&= E_{Q(H)}\left(\log\left(P(D)\right)\right) = \log(P(D))
\end{aligned}$$

$\Rightarrow$ the bound is tight if (and only if) $Q(H) = P(H|D)$, i.e. when the approximation is exact.

Reminder:
Given: observable data D and hidden variables/parameters H.
We'd like to maximize a lower bound $\mathcal{L}(D, Q(H))$ on $P(D)$: $P(D) \geq \mathcal{L}(D, Q(H))$ with respect to $Q(H)$.
Kullback-Leibler divergence: $D(Q||P) = E_Q\left(\log\left(\frac{Q(X)}{P(X)}\right)\right)$. Jensen's inequality: $\log(E(X)) \geq E(\log(X))$
Lower bound $\mathcal{L}(D, Q(H)) = E_{Q(H)}\left(\log\left(\frac{P(D,H)}{Q(H)}\right)\right) \leq \log(P(D))$

Consequence of $\mathcal{L}(D, Q(H)) \leq \log(P(D))$:

no overfitting!

Reminder:

Given: observable data D and hidden variables/parameters H.

We'd like to maximize a lower bound $\mathcal{L}(D, Q(H))$ on $P(D)$: $P(D) \geq \mathcal{L}(D, Q(H))$ with respect to $Q(H)$.

Kullback-Leibler divergence: $D(Q||P) = E_Q\left(\log\left(\frac{Q(X)}{P(X)}\right)\right)$. Jensen's inequality: $\log(E(X)) \geq E(\log(X))$

Lower bound $\mathcal{L}(D, Q(H)) = E_{Q(H)}\left(\log\left(\frac{P(D,H)}{Q(H)}\right)\right) \leq \log(P(D))$

Consequence of $\mathcal{L}(D, Q(H)) \leq \log(P(D))$:

no overfitting!

Question: does that mean we don't have to cross-validate?

Answer: no. Need to check how "underfitted" the solution is.

An interpretation of $\mathcal{L}(D, Q(H))$

Reminder:
Given: observable data D and hidden variables/parameters H.
We'd like to maximize a lower bound $\mathcal{L}(D, Q(H))$ on $P(D)$: $P(D) \geq \mathcal{L}(D, Q(H))$ with respect to $Q(H)$.
Kullback-Leibler divergence: $D(Q||P) = E_Q\left(\log\left(\frac{Q(X)}{P(X)}\right)\right)$. Jensen's inequality: $\log(E(X)) \geq E(\log(X))$
Lower bound $\mathcal{L}(D, Q(H)) = E_{Q(H)}\left(\log\left(\frac{P(D,H)}{Q(H)}\right)\right) \leq \log(P(D))$

Question: can $\mathcal{L}(D, Q(H))$ be interpreted?
Answer 1: Note that $P(D, H) = P(D|H)P(H)$. Thus

$$\begin{aligned}
\mathcal{L}(D, Q(H)) &= E_{Q(H)}\left(\log\left(\frac{P(D,H)}{Q(H)}\right)\right) = E_{Q(H)}\left(\log\left(\frac{P(D|H)P(H)}{Q(H)}\right)\right) \\
&= E_{Q(H)}\left(\log(P(D|H)) + \log\left(\frac{P(H)}{Q(H)}\right)\right) \\
&= E_{Q(H)}(\log(P(D|H))) - E_{Q(H)}\left(\log\left(\frac{Q(H)}{P(H)}\right)\right) \\
&= E_{Q(H)}(\log(P(D|H))) - D(Q(H)||P(H))
\end{aligned}$$

An interpretation of $\mathcal{L}(D, Q(H))$, contd.

Reminder:
Given: observable data D and hidden variables/parameters H.
We'd like to maximize a lower bound $\mathcal{L}(D, Q(H))$ on $P(D)$: $P(D) \geq \mathcal{L}(D, Q(H))$ with respect to $Q(H)$.
Kullback-Leibler divergence: $D(Q||P) = E_Q \left(\log \left(\frac{Q(X)}{P(X)} \right) \right)$. Jensen's inequality: $\log(E(X)) \geq E(\log(X))$
Lower bound $\mathcal{L}(D, Q(H)) = E_{Q(H)} \left(\log \left(\frac{P(D,H)}{Q(H)} \right) \right) \leq \log(P(D))$

We found:

$$\mathcal{L}(D, Q(H)) = \underbrace{E_{Q(H)}(\log(P(D|H)))}_{\text{log-likelihood of data D}} - \underbrace{D(Q(H)||P(H))}_{\text{divergence from prior}}$$

Maximizing $\mathcal{L}(D, Q(H))$ therefore means:

- find a good explanation for D (large log-likelihood), and
- maintain prior beliefs as much as possible.

Another interpretation of $\mathcal{L}(D, Q(H))$

Reminder:

Given: observable data D and hidden variables/parameters H.

We'd like to maximize a lower bound $\mathcal{L}(D, Q(H))$ on $P(D)$: $P(D) \geq \mathcal{L}(D, Q(H))$ with respect to $Q(H)$.

Kullback-Leibler divergence: $D(Q||P) = E_Q\left(\log\left(\frac{Q(X)}{P(X)}\right)\right)$. Jensen's inequality: $\log(E(X)) \geq E(\log(X))$

Lower bound $\mathcal{L}(D, Q(H)) = E_{Q(H)}\left(\log\left(\frac{P(D,H)}{Q(H)}\right)\right) \leq \log(P(D))$

Question: can $\mathcal{L}(D, Q(H))$ be interpreted?

Answer 2:

$$
\begin{aligned}
\mathcal{L}(D, Q(H)) &= E_{Q(H)}\left(\log\left(\frac{P(D,H)}{Q(H)}\right)\right) \\
&= E_{Q(H)}\left(\log(P(D,H)) - \log(Q(H))\right) \\
&= \underbrace{E_{Q(H)}\left(\log(P(D,H))\right)}_{=:-U(D,Q(H))} - \underbrace{E_{Q(H)}\left(\log(Q(H))\right)}_{-S(Q(H))} \\
&= -U(D, Q(H)) + S(Q(H))
\end{aligned}
$$

$U(D, Q(H))$: expected 'energy' or 'cost' of H under $Q(H)$

$S(Q(H))$: Shannon entropy (uncertainty) of H under $Q(H)$.

c.f. [Shannon 1948, Friston et al. 2012]

Another interpretation of $\mathcal{L}(D, Q(H))$, contd.

Reminder:
Given: observable data D and hidden variables/parameters H.
We'd like to maximize a lower bound $\mathcal{L}(D, Q(H))$ on $P(D)$: $P(D) \geq \mathcal{L}(D, Q(H))$ with respect to $Q(H)$.
Kullback-Leibler divergence: $D(Q||P) = E_Q\left(\log\left(\frac{Q(X)}{P(X)}\right)\right)$. Jensen's inequality: $\log(E(X)) \geq E(\log(X))$
Lower bound $\mathcal{L}(D, Q(H)) = E_{Q(H)}\left(\log\left(\frac{P(D,H)}{Q(H)}\right)\right) \leq \log(P(D))$

We found:

$$\mathcal{L}(D, Q(H)) = -U(D, Q(H)) + S(Q(H))$$

Maximizing $\mathcal{L}(D, Q(H))$ therefore means:

- minimize expected cost/energy, and
- maximize posterior uncertainty about H

Another interpretation of $\mathcal{L}(D, Q(H))$, contd.

Reminder:
Given: observable data D and hidden variables/parameters H.
We'd like to maximize a lower bound $\mathcal{L}(D, Q(H))$ on $P(D)$: $P(D) \geq \mathcal{L}(D, Q(H))$ with respect to $Q(H)$.
Kullback-Leibler divergence: $D(Q||P) = E_Q\left(\log\left(\frac{Q(X)}{P(X)}\right)\right)$. Jensen's inequality: $\log(E(X)) \geq E(\log(X))$
Lower bound $\mathcal{L}(D, Q(H)) = E_{Q(H)}\left(\log\left(\frac{P(D,H)}{Q(H)}\right)\right) \leq \log(P(D))$

We found:

$$\mathcal{L}(D, Q(H)) = -U(D, Q(H)) + S(Q(H))$$

Maximizing $\mathcal{L}(D, Q(H))$ therefore means:

- minimize expected cost/energy, and
- maximize posterior uncertainty about H

Note: formal relationship with Helmholtz free energy $F = U - TS$ in thermal physics: if $T = 1$, $\mathcal{L}(D, Q(H)) = -F$.
Hence, maximizing $\mathcal{L}(D, Q(H)) \Leftrightarrow$ minimizing F.

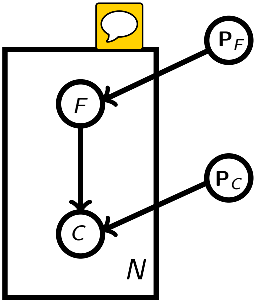
Back to the example: how loaded is my coin, and how often?

Reminder:
Lower bound $\mathcal{L}(D, Q(H)) = E_{Q(H)}\left(\log\left(\frac{P(D,H)}{Q(H)}\right)\right) \leq \log(P(D))$

To construct $\mathcal{L}(D, Q(H))$, we need

- Joint density: $p(C_i, F_i, \mathbf{P}_F, \mathbf{P}_C)$
- Approximating density: $q(C_i, F_i, \mathbf{P}_F, \mathbf{P}_C)$

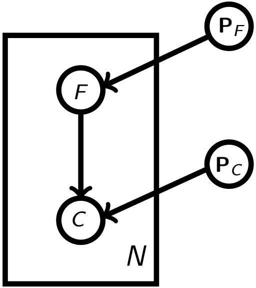
$$p(C_i, F_i, \mathbf{P}_F, \mathbf{P}_C) =$$
$$\left[\prod_{i=1}^{N} P(C_i|F_i, \mathbf{P}_C)P(F_i|\mathbf{P}_F)\right] p(\mathbf{P}_F)p(\mathbf{P}_C)$$



Breaking the loops

Reminder:
Lower bound $\mathcal{L}(D, Q(H)) = E_{Q(H)}\left(\log\left(\frac{P(D,H)}{Q(H)}\right)\right) \leq \log(P(D))$
$p(C_i, F_i, \mathbf{P}_F, \mathbf{P}_C) = \left[\prod_{i=1}^{N} P(C_i|F_i, \mathbf{P}_C)P(F_i|\mathbf{P}_F)\right] p(\mathbf{P}_F)p(\mathbf{P}_C)$

Question: how to choose $q(C_i, F_i, \mathbf{P}_F, \mathbf{P}_C)$?
Hard part are the loops.

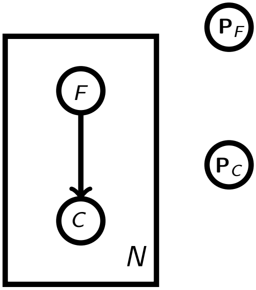


Breaking the loops

Reminder:
Lower bound $\mathcal{L}\left(D, Q\left(H\right)\right) = E_{Q(H)}\left(\log\left(\frac{P(D,H)}{Q(H)}\right)\right) \leq \log(P(D))$
$p(C_i, F_i, \mathbf{P}_F, \mathbf{P}_C) = \left[\prod_{i=1}^{N} P(C_i|F_i, \mathbf{P}_C)P(F_i|\mathbf{P}_F)\right] p(\mathbf{P}_F)p(\mathbf{P}_C)$

Question: how to choose $q(C_i, F_i, \mathbf{P}_F, \mathbf{P}_C)$?
Hard part are the loops.
Let's break them:

$$q(F_i, \mathbf{P}_F, \mathbf{P}_C) = \left[\prod_{i=1}^{N} Q_i(F_i)\right] q(\mathbf{P}_F)q(\mathbf{P}_C)$$



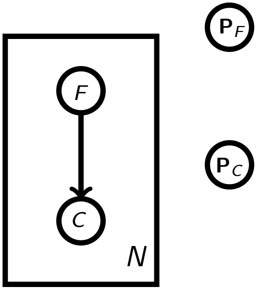
Breaking the loops

Reminder:
Lower bound $\mathcal{L}(D, Q(H)) = E_{Q(H)}\left(\log\left(\frac{P(D,H)}{Q(H)}\right)\right) \leq \log(P(D))$
$p(C_i, F_i, \mathbf{P}_F, \mathbf{P}_C) = \left[\prod_{i=1}^{N} P(C_i|F_i, \mathbf{P}_C)P(F_i|\mathbf{P}_F)\right] p(\mathbf{P}_F)p(\mathbf{P}_C)$

Question: how to choose $q(C_i, F_i, \mathbf{P}_F, \mathbf{P}_C)$?
Hard part are the loops.
Let's break them:

$$q(F_i, \mathbf{P}_F, \mathbf{P}_C) = \left[\prod_{i=1}^{N} Q_i(F_i)\right] q(\mathbf{P}_F)q(\mathbf{P}_C)$$

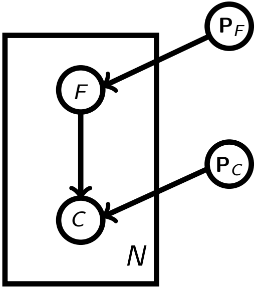
But this looks like the parameters $\mathbf{P}_F$, $\mathbf{P}_C$ are disconnected from the data C_i!



Connecting the parameters to the data via $\mathcal{L}(D, Q(H))$

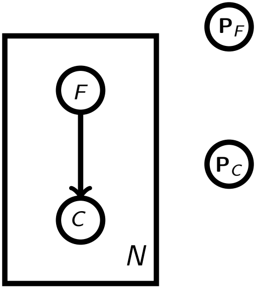
Reminder:
Lower bound $\mathcal{L}(D, Q(H)) = E_{Q(H)}\left(\log\left(\frac{P(D,H)}{Q(H)}\right)\right) \le \log(P(D))$

$p(\mathbf{C}, \mathbf{F}, \mathbf{P}_F, \mathbf{P}_C)$



$\mathcal{L}(\mathbf{C}, \prod_i Q_i(F_i)q(\mathbf{P}_C)q(\mathbf{P}_F))$

$\prod_i Q_i(F_i)q(\mathbf{P}_F)q(\mathbf{P}_C)$



$\Rightarrow$ the bound contains *both* the exact joint density, and the approximating one. Parameters **are connected** to the data.

Computing the bound

Reminder:
Lower bound $\mathcal{L}\left(D, Q\left(H\right)\right) = E_{Q(H)}\left(\log\left(\frac{P(D,H)}{Q(H)}\right)\right) \leq \log(P(D))$
Exact joint density $p(C_i, F_i, \mathbf{P}_F, \mathbf{P}_C) = \left[\prod_{i=1}^{N} P(C_i|F_i, \mathbf{P}_C)P(F_i|\mathbf{P}_F)\right] p(\mathbf{P}_F)p(\mathbf{P}_C)$
Approximating density $\left[\prod_{i=1}^{N} Q_i(F_i)\right] q(\mathbf{P}_F)q(\mathbf{P}_C)$

In our example (E_q is expectation w.r.t approximating density):

$$\begin{aligned}
\mathcal{L} &= E_q\left[\log\left(\frac{\left[\prod_{i=1}^{N} P(C_i|F_i,\mathbf{P}_C)P(F_i|\mathbf{P}_F)\right] p(\mathbf{P}_F)p(\mathbf{P}_C)}{\left[\prod_{i=1}^{N} Q_i(F_i)\right] q(\mathbf{P}_F)q(\mathbf{P}_C)}\right)\right] \\
&= \sum_{i=1}^{N} E_q\left[\log\left(\frac{P(C_i|F_i,\mathbf{P}_C)P(F_i|\mathbf{P}_F)}{Q_i(F_i)}\right)\right] \\
&\quad - E_q\left[\log\left(\frac{q(\mathbf{P}_F)}{p(\mathbf{P}_F)}\right)\right] - E_q\left[\log\left(\frac{q(\mathbf{P}_C)}{p(\mathbf{P}_C)}\right)\right]
\end{aligned}$$

Computing the bound, contd.

Reminder:
Approximating density $\left[\prod_{i=1}^{N} Q_i(F_i)\right] q(\mathbf{P}_F) q(\mathbf{P}_C)$

$$E_q\left[\log\left(\frac{q(\mathbf{P}_F)}{p(\mathbf{P}_F)}\right)\right] = E_{q(\mathbf{P}_F)}\left[\log\left(\frac{q(\mathbf{P}_F)}{p(\mathbf{P}_F)}\right)\right] = D(q(\mathbf{P}_F)||p(\mathbf{P}_F))$$

$$E_q\left[\log\left(\frac{q(\mathbf{P}_C)}{p(\mathbf{P}_C)}\right)\right] = E_{q(\mathbf{P}_C)}\left[\log\left(\frac{q(\mathbf{P}_C)}{p(\mathbf{P}_C)}\right)\right] = D(q(\mathbf{P}_C)||p(\mathbf{P}_C))$$

Computing the bound, contd.

Reminder:
Approximating density $\left[\prod_{i=1}^{N} Q_i(F_i)\right] q(\mathbf{P}_F)q(\mathbf{P}_C)$

$$
\begin{aligned}
L_i &= E_q\left[\log\left(\frac{P(C_i|F_i,\mathbf{P}_C)P(F_i|\mathbf{P}_F)}{Q_i(F_i)}\right)\right] \\
&= E_q\left[\log\left(P(C_i|F_i,\mathbf{P}_C)P(F_i|\mathbf{P}_F)\right)\right] - E_q\left[\log(Q_i(F_i))\right] \\
&= E_{Q_i(F_i)}\left[E_{q(\mathbf{P}_F)q(\mathbf{P}_C)}\left[\log\left(P(C_i|F_i,\mathbf{P}_C)P(F_i|\mathbf{P}_F)\right)\right] - \log(Q_i(F_i))\right] \\
&=: E_{Q_i(F_i)}\left[\log(U_i(F_i)) - \log(Q_i(F_i))\right]
\end{aligned}
$$

Computing the bound, contd.

Reminder:
Approximating density $\left[\prod_{i=1}^{N} Q_i(F_i)\right] q(\mathbf{P}_F)q(\mathbf{P}_C)$

$$
\begin{aligned}
L_i &= E_q\left[\log\left(\frac{P(C_i|F_i,\mathbf{P}_C)P(F_i|\mathbf{P}_F)}{Q_i(F_i)}\right)\right] \\
&= E_q\left[\log\left(P(C_i|F_i,\mathbf{P}_C)P(F_i|\mathbf{P}_F)\right)\right] - E_q\left[\log(Q_i(F_i))\right] \\
&= E_{Q_i(F_i)}\left[E_{q(\mathbf{P}_F)q(\mathbf{P}_C)}\left[\log\left(P(C_i|F_i,\mathbf{P}_C)P(F_i|\mathbf{P}_F)\right)\right] - \log(Q_i(F_i))\right] \\
&=: E_{Q_i(F_i)}\left[\log(U_i(F_i)) - \log(Q_i(F_i))\right]
\end{aligned}
$$

Let $Z_i = \sum_{F_i} U_i(F_i)$. Then $\tilde{Q}_i(F_i) = \frac{U_i(F_i)}{Z_i}$ is a probability distribution over F_i. Thus

Computing the bound, contd.

Reminder:
Approximating density $\left[\prod_{i=1}^{N} Q_i(F_i)\right] q(\mathbf{P}_F)q(\mathbf{P}_C)$

$$
\begin{aligned}
L_i &= E_q\left[\log\left(\frac{P(C_i|F_i,\mathbf{P}_C)P(F_i|\mathbf{P}_F)}{Q_i(F_i)}\right)\right] \\
&= E_q\left[\log\left(P(C_i|F_i,\mathbf{P}_C)P(F_i|\mathbf{P}_F)\right)\right] - E_q\left[\log(Q_i(F_i))\right] \\
&= E_{Q_i(F_i)}\left[E_{q(\mathbf{P}_F)q(\mathbf{P}_C)}\left[\log\left(P(C_i|F_i,\mathbf{P}_C)P(F_i|\mathbf{P}_F)\right)\right] - \log(Q_i(F_i))\right] \\
&=: E_{Q_i(F_i)}\left[\log(U_i(F_i)) - \log(Q_i(F_i))\right]
\end{aligned}
$$

Let $Z_i = \sum_{F_i} U_i(F_i)$. Then $\tilde{Q}_i(F_i) = \frac{U_i(F_i)}{Z_i}$ is a probability distribution over F_i. Thus

$$
\begin{aligned}
L_i &= E_{Q_i(F_i)}\left[\log(Z_i) + \log(\tilde{Q}_i(F_i)) - \log(Q_i(F_i))\right] \\
&= \log(Z_i) - E_{Q_i(F_i)}\left[\log\left(\frac{\log(Q_i(F_i))}{\log(\tilde{Q}_i(F_i))}\right)\right] \\
&= \log(Z_i) - D(Q_i(F_i)||\tilde{Q}_i(F_i))
\end{aligned}
$$

Putting it all together

Thus, we find for the bound

$$\mathcal{L} = \sum_{i=1}^{N} \Big(\log(Z_i) - D(Q_i(F_i)||\tilde{Q}_i(F_i))\Big) \\ -D(q(\mathbf{P}_F)||p(\mathbf{P}_F)) - D(q(\mathbf{P}_C)||p(\mathbf{P}_C))$$

1. For fixed $q(\mathbf{P}_F)$ and $q(\mathbf{P}_C)$, $\mathcal{L}$ is maximized by setting $Q_i(F_i) = \tilde{Q}_i(F_i)$.
2. $\mathcal{L}$ can be increased further by fixing the $Q_i(F_i)$ and maximizing w.r.t. $q(\mathbf{P}_F)$ and $q(\mathbf{P}_C)$.

Iterating these two steps will keep increasing $\mathcal{L}$. This is an example of a variational *expectation-maximization (EM)* algorithm. We derive the *E*-step so far.

Summary 1: variational approximations

- Computing the parameter posterior ('learning') can be difficult even in simple models.
- Instead of evaluating the posterior directly, restate inference as an **optimization** problem.
- Introduces an approximating posterior instead of the correct one.
- It's called ”variational” because the approximating posterior is varied until optimal.

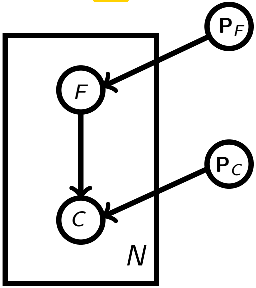
Summary 2: variational approximations

- **Jensen's inequality**: let $f(x)$ be a convex function, and $X \in \{x_1, \ldots, x_N\}$ be a random variable with distribution $P(X)$. Then $f(E(X)) \leq E(f(X))$.
- **Kullback-Leibler divergence** between 2 distributions (or densities): $D(Q||P) = \sum_X Q(x) \log \left(\frac{Q(x)}{P(x)} \right)$
- $D(Q||P) \geq 0$ with equality only if $Q = P$.
- Approximate inference/learning can be done by maximizing $\mathcal{L}(D, Q(H)) = E_{Q(H)} \left(\log \left(\frac{P(D,H)}{Q(H)} \right) \right) \leq \log(P(D))$
- Bound becomes tight when inference is exact.
- Interpretation of variational learning: explain data well while keeping prior beliefs as much as possible.

Learning the loadedness of the coin $\mathbf{P}_C$

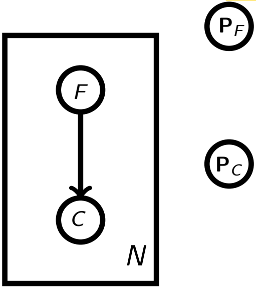
Reminder:
Lower bound $\mathcal{L}(D, Q(H)) = E_{Q(H)}\left(\log\left(\frac{P(D,H)}{Q(H)}\right)\right) \leq \log(P(D))$

$p(\mathbf{C}, \mathbf{F}, \mathbf{P}_F, \mathbf{P}_C)$



$\mathcal{L}(\mathbf{C}, \prod_i Q_i(F_i) q(\mathbf{P}_C) q(\mathbf{P}_F))$

$\prod_i Q_i(F_i) q(\mathbf{P}_F) q(\mathbf{P}_C)$



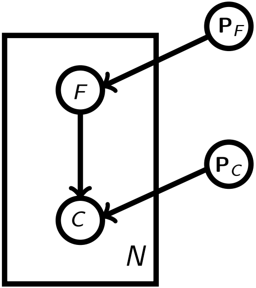
Questions:

1. How do we learn $\mathbf{P}_C$?
2. Can variational approximations avoid infinitely long messages?

Learning the loadedness of the coin $\mathbf{P}_C$

Reminder:
Lower bound $\mathcal{L}(D, Q(H)) = E_{Q(H)}\left(\log\left(\frac{P(D,H)}{Q(H)}\right)\right) \leq \log(P(D))$

$p(\mathbf{C}, \mathbf{F}, \mathbf{P}_F, \mathbf{P}_C)$



$\mathbf{P}_C = (P_h, P_t)$ such that $P_h + P_t = 1$

$$P(C|F = f, \mathbf{P}_C) = \begin{cases} C = h : 0.5 \\ C = t : 0.5 \end{cases}$$

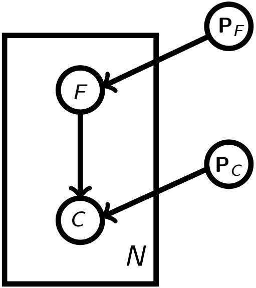
$$P(C|F = l, \mathbf{P}_C) = \begin{cases} C = h : P_h \\ C = t : P_t \end{cases}$$

$\Rightarrow$ difficult part: $p(\mathbf{P}_C)$ is the infintely long message.

Learning the loadedness of the coin $\mathbf{P}_C$

Reminder:
Lower bound $\mathcal{L}(D, Q(H)) = E_{Q(H)}\left(\log\left(\frac{P(D,H)}{Q(H)}\right)\right) \leq \log(P(D))$

$p(\mathbf{C}, \mathbf{F}, \mathbf{P}_F, \mathbf{P}_C)$



$$\mathbf{P}_C = (P_h, P_t) \text{ such that } P_h + P_t = 1$$

$$P(C|F = f, \mathbf{P}_C) = \begin{cases} C = h : 0.5 \\ C = t : 0.5 \end{cases}$$

$$P(C|F = l, \mathbf{P}_C) = \begin{cases} C = h : P_h \\ C = t : P_t \end{cases}$$

$\Rightarrow$ difficult part: $p(\mathbf{P}_C)$ is the infintely long message.

Solution: reparameterization with exponential family distributions/densities.

Exponential family distributions

A distribution/density is said to belong to the exponential family, if it can be written in the form:

$$p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta})\exp(\boldsymbol{\eta}^T\mathbf{u}(\mathbf{x}))$$

- The random variates $\mathbf{x}$ may be discrete or continuous.
- The *sufficient statistics* $\mathbf{u}$ are functions of the $\mathbf{x}$.
- The $\boldsymbol{\eta}$ are the *natural parameters*, one for each sufficient statistic.
- $g(\boldsymbol{\eta})$ is the normalization constant:

$$g(\boldsymbol{\eta})\int d\mathbf{x}\; h(\mathbf{x})\exp(\boldsymbol{\eta}^T\mathbf{u}(\mathbf{x})) = 1$$

Example: coin toss distribution for loaded coin

Reminder:

$\mathbf{P}_C = (P_h, P_t)$ such that $P_h + P_t = 1$, $P(C|F = l, \mathbf{P}_C) = \begin{cases} C = h : P_h \\ C = t : P_t \end{cases}$

$p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta})\exp(\boldsymbol{\eta}^T \mathbf{u}(\mathbf{x}))$

Sufficient statistic: $u(c) = \begin{cases} c = h : 1 \\ c = t : 0 \end{cases}$

Example: coin toss distribution for loaded coin

Reminder:

$\mathbf{P}_C = (P_h, P_t)$ such that $P_h + P_t = 1$, $P(C|F = l, \mathbf{P}_C) = \begin{cases} C = h : P_h \\ C = t : P_t \end{cases}$

$p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta})\exp(\boldsymbol{\eta}^T\mathbf{u}(\mathbf{x}))$

Sufficient statistic: $u(c) = \begin{cases} c = h : 1 \\ c = t : 0 \end{cases}$

Then the distribution can be written as:

$$\begin{aligned} P(C = c|F = l, \mathbf{P}_C) &= P_h^{u(c)}(1 - P_h)^{1-u(c)} \\ &= \exp\left(u(c)\log(P_h) + (1 - u(c))\log(1 - P_h)\right) \\ &= \exp\left(u(c)\log\left(\frac{P_h}{1 - P_h}\right) + \log(1 - P_h)\right) \end{aligned}$$

Example: coin toss distribution for loaded coin

Reminder:

$\mathbf{P}_C = (P_h, P_t)$ such that $P_h + P_t = 1$, $P(C|F = l, \mathbf{P}_C) = \begin{cases} C = h : P_h \\ C = t : P_t \end{cases}$

$p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta})\exp(\boldsymbol{\eta}^T\mathbf{u}(\mathbf{x}))$

Sufficient statistic: $u(c) = \begin{cases} c = h : 1 \\ c = t : 0 \end{cases}$

Then the distribution can be written as:

$$\begin{aligned} P(C = c|F = l, \mathbf{P}_C) &= P_h^{u(c)}(1 - P_h)^{1-u(c)} \\ &= \exp\left(u(c)\log(P_h) + (1 - u(c))\log(1 - P_h)\right) \\ &= \exp\left(u(c)\log\left(\frac{P_h}{1 - P_h}\right) + \log(1 - P_h)\right) \end{aligned}$$

Identify $\eta = \log\left(\frac{P_h}{1-P_h}\right)$ ("logit") and thus $1 - P_h = \frac{1}{1-\exp(\eta)} = \sigma(-\eta)$

$$P(C = c|F = l, \eta) = \underbrace{1}_{h(c)}\underbrace{\sigma(-\eta)}_{g(\eta)}\exp(\eta u(c))$$

Properties of exponential family distributions

Reminder:
$p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta})\exp(\boldsymbol{\eta}^T\mathbf{u}(\mathbf{x}))$
$g(\boldsymbol{\eta})\int d\mathbf{x}\; h(\mathbf{x})\exp(\boldsymbol{\eta}^T\mathbf{u}(\mathbf{x})) = 1$

Expectations can be computed from the normalization constant:

$$(\nabla_{\boldsymbol{\eta}} g(\boldsymbol{\eta}))\underbrace{\int d\mathbf{x}\; h(\mathbf{x})\exp(\boldsymbol{\eta}^T\mathbf{u}(\mathbf{x}))}_{=\frac{1}{g(\boldsymbol{\eta})}} + \underbrace{g(\boldsymbol{\eta})\int d\mathbf{x}\; h(\mathbf{x})\mathbf{u}(\mathbf{x})\exp(\boldsymbol{\eta}^T\mathbf{u}(\mathbf{x}))}_{\langle\mathbf{u}(\mathbf{x})\rangle} = \mathbf{0}$$

and thus the expectation $\langle\mathbf{u}(\mathbf{x})\rangle$ is:

$$\boxed{\langle\mathbf{u}(\mathbf{x})\rangle = -\frac{\nabla_{\boldsymbol{\eta}} g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})} = -\nabla\log(g(\boldsymbol{\eta}))}$$

Likewise, by differentiating again we find:

$$\boxed{\mathrm{Cov}(\mathbf{u}(\mathbf{x})) = -\nabla_{\boldsymbol{\eta}}\nabla_{\boldsymbol{\eta}}\;\log(g(\boldsymbol{\eta}))}$$

where $\nabla_{\boldsymbol{\eta}}\nabla_{\boldsymbol{\eta}}$ computes the Hessian matrix.

Conjugate priors on exp. fam. distributions

Reminder:
$p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta})\exp(\boldsymbol{\eta}^T\mathbf{u}(\mathbf{x}))$
$P(C = c|F = l, \mathbf{P}_C) = \sigma(-\eta)\exp(\eta u(c))$

So far we reparameterized $P(C|F = l, \mathbf{P}_C)$. But the difficult part was $p(\mathbf{P}_C)$.
We'd like to

- Parameterize $p(\mathbf{P}_C)$ with a small number of parameters (short messages), and
- keep that parametric form after observing data.

Solution: a conjugate prior. A prior is conjugate to a likelihood, if the posterior after observing data has the same form as the prior.

Conjugate priors on exp. fam. distributions

Reminder:
$p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta})\exp(\boldsymbol{\eta}^T\mathbf{u}(\mathbf{x}))$

The **conjugate prior** for

$$p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta})\exp(\boldsymbol{\eta}^T\mathbf{u}(\mathbf{x}))$$

is given by:

$$p(\boldsymbol{\eta}|\boldsymbol{\lambda},\nu) = f(\boldsymbol{\lambda},\nu)g(\boldsymbol{\eta})^{\nu}\exp(\nu\boldsymbol{\eta}^T\boldsymbol{\lambda})$$

where

- $\boldsymbol{\lambda}$ are the parameters of the p(oste)rior,
- ν is the concentration parameter,
- $g(\boldsymbol{\eta})$ is the same function as before, and
- $f(\boldsymbol{\lambda},\nu)$ is the normalization constant.

Proof of conjugacy

Reminder:
distribution/density: $p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta})\exp(\boldsymbol{\eta}^T\mathbf{u}(\mathbf{x}))$
conjugate prior: $p(\boldsymbol{\eta}|\boldsymbol{\lambda},\nu) = f(\boldsymbol{\lambda},\nu)g(\boldsymbol{\eta})^\nu\exp(\nu\boldsymbol{\eta}^T\boldsymbol{\lambda})$

Proof: assume we observed N datapoints $\mathbf{x}_{1:N}$.

$$
\begin{aligned}
p(\boldsymbol{\eta}|\boldsymbol{\lambda},\nu,\mathbf{x}_{1:N}) &= \frac{p(\mathbf{x}_{1:N},\boldsymbol{\eta}|\boldsymbol{\lambda},\nu)}{p(\mathbf{x}_{1:N}|\boldsymbol{\lambda},\nu)} \\
&= \frac{\prod_{n=1}^N p(\mathbf{x}_n|\boldsymbol{\eta})\; p(\boldsymbol{\eta}|\boldsymbol{\lambda},\nu)}{\int d\boldsymbol{\eta}\prod_{n=1}^N p(\mathbf{x}_n|\boldsymbol{\eta})\; p(\boldsymbol{\eta}|\boldsymbol{\lambda},\nu)} \\
&= \frac{\prod_{n=1}^N g(\boldsymbol{\eta})h(\mathbf{x}_n)\exp(\boldsymbol{\eta}^T\mathbf{u}(\mathbf{x}_n))\cdot f(\boldsymbol{\lambda},\nu)g(\boldsymbol{\eta})^\nu\exp(\nu\boldsymbol{\eta}^T\boldsymbol{\lambda})}{\int d\boldsymbol{\eta}\prod_{n=1}^N g(\boldsymbol{\eta})h(\mathbf{x}_n)\exp(\boldsymbol{\eta}^T\mathbf{u}(\mathbf{x}_n))\cdot f(\boldsymbol{\lambda},\nu)g(\boldsymbol{\eta})^\nu\exp(\nu\boldsymbol{\eta}^T\boldsymbol{\lambda})} \\
&= \frac{\prod_n h(\mathbf{x}_n)f(\boldsymbol{\lambda},\nu)\; g(\boldsymbol{\eta})^{\nu+N}\exp\left(\boldsymbol{\eta}^T\left(\nu\boldsymbol{\lambda}+\sum_n\mathbf{u}(\mathbf{x}_n)\right)\right)}{\prod_n h(\mathbf{x}_n)f(\boldsymbol{\lambda},\nu)\int d\boldsymbol{\eta}\; g(\boldsymbol{\eta})^{\nu+N}\exp\left(\boldsymbol{\eta}^T\left(\nu\boldsymbol{\lambda}+\sum_n\mathbf{u}(\mathbf{x}_n)\right)\right)} \\
&= \frac{g(\boldsymbol{\eta})^{\nu+N}\exp\left(\boldsymbol{\eta}^T\left(\nu\boldsymbol{\lambda}+\sum_n\mathbf{u}(\mathbf{x}_n)\right)\right)}{\int d\boldsymbol{\eta}\; g(\boldsymbol{\eta})^{\nu+N}\exp\left(\boldsymbol{\eta}^T\left(\nu\boldsymbol{\lambda}+\sum_n\mathbf{u}(\mathbf{x}_n)\right)\right)}
\end{aligned}
$$

Proof of conjugacy, cont.

Reminder:
distribution/density: $p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta})\exp(\boldsymbol{\eta}^T\mathbf{u}(\mathbf{x}))$
conjugate prior: $p(\boldsymbol{\eta}|\boldsymbol{\lambda},\nu) = f(\boldsymbol{\lambda},\nu)g(\boldsymbol{\eta})^{\nu}\exp(\nu\boldsymbol{\eta}^T\boldsymbol{\lambda})$

Proof: assume we observed N datapoints $\mathbf{x}_{1:N}$.

$$p(\boldsymbol{\eta}|\boldsymbol{\lambda},\nu,\mathbf{x}_{1:N}) = \frac{g(\boldsymbol{\eta})^{\nu+N}\exp\left(\boldsymbol{\eta}^T\left(\nu\boldsymbol{\lambda}+\sum_n \mathbf{u}(\mathbf{x}_n)\right)\right)}{\int d\boldsymbol{\eta}\; g(\boldsymbol{\eta})^{\nu+N}\exp\left(\boldsymbol{\eta}^T\left(\nu\boldsymbol{\lambda}+\sum_n \mathbf{u}(\mathbf{x}_n)\right)\right)}$$

Define the **posterior parameters** as

$$\begin{aligned} \tilde{\nu} &= \nu + N \\ \tilde{\boldsymbol{\lambda}} &= \frac{\nu\boldsymbol{\lambda}+\sum_n \mathbf{u}(\mathbf{x}_n)}{\tilde{\nu}} \end{aligned}$$

and identify $f(\tilde{\boldsymbol{\lambda}},\tilde{\nu}) = \left(\int d\boldsymbol{\eta}\; g(\boldsymbol{\eta})^{\nu+N}\exp\left(\boldsymbol{\eta}^T\left(\nu\boldsymbol{\lambda}+\sum_n \mathbf{u}(\mathbf{x}_n)\right)\right)\right)^{-1}$

$$\Rightarrow\; p(\boldsymbol{\eta}|\boldsymbol{\lambda},\nu,\mathbf{x}_{1:N}) = f(\tilde{\boldsymbol{\lambda}},\tilde{\nu})g(\boldsymbol{\eta})^{\tilde{\nu}}\exp\left(\tilde{\nu}\boldsymbol{\eta}^T\tilde{\boldsymbol{\lambda}}\right) = p(\boldsymbol{\eta}|\tilde{\boldsymbol{\lambda}},\tilde{\nu})$$

Example: density of $\mathbf{P}_C$

Reminder:
distribution/density: $p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta})\exp(\boldsymbol{\eta}^T\mathbf{u}(\mathbf{x}))$
conjugate prior: $p(\boldsymbol{\eta}|\boldsymbol{\lambda},\nu) = f(\boldsymbol{\lambda},\nu)g(\boldsymbol{\eta})^\nu\exp(\nu\boldsymbol{\eta}^T\boldsymbol{\lambda})$
natural parameter: $\eta = \log\left(\frac{P_h}{1-P_h}\right)$, $\sigma(-\eta) = 1 - P_h$

We found for the coin toss distribution:

$$P(C = c|F = l, \eta) = \underbrace{1}_{h(c)}\underbrace{\sigma(-\eta)}_{g(\eta)}\exp(\eta u(c))$$

Thus, the exponential family conjugate prior is:

$$p(\eta|\lambda,\nu) = f(\lambda,\nu)\sigma(-\eta)^\nu\exp(\nu\eta\lambda)$$

Question: how to compute $f(\lambda,\nu)$?

Example: density of $\mathbf{P}_C$, contd.

Reminder:
distribution/density: $p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta})\exp(\boldsymbol{\eta}^T\mathbf{u}(\mathbf{x}))$
conjugate prior: $p(\boldsymbol{\eta}|\boldsymbol{\lambda},\nu) = f(\boldsymbol{\lambda},\nu)g(\boldsymbol{\eta})^\nu \exp(\nu\boldsymbol{\eta}^T\boldsymbol{\lambda})$
natural parameter: $\eta = \log\left(\frac{P_h}{1-P_h}\right)$, $\sigma(-\eta) = 1 - P_h$

To transform this into a "textbook form" and compute the normalization constant, note that

$$
\begin{aligned}
p(\eta|\lambda,\nu) &= f(\lambda,\nu)(1-P_h)^\nu \exp\left(\nu\lambda\log\left(\frac{P_h}{1-P_h}\right)\right) \\
&= f(\lambda,\nu)\exp\left(\nu\lambda\log(P_h) + \nu(1-\lambda)\log(1-P_h)\right) \\
&= f(\lambda,\nu)P_h^{\nu\lambda}(1-P_h)^{\nu(1-\lambda)}
\end{aligned}
$$

Substitute $\alpha = \nu\lambda$, $\beta = \nu(1-\lambda)$, $\frac{d\eta}{dP_h} = \frac{1}{P_h(1-P_h)}$, $f(\lambda,\nu) = B(\alpha,\beta)^{-1}$:

$$\boxed{p(P_h|\alpha,\beta) = B(\alpha,\beta)^{-1}P_h^{\alpha-1}(1-P_h)^{\beta-1}}$$

i.e. the conjugate prior is a Beta-distribution!

Properties of exponential family conjugate priors

Reminder:
distribution/density: $p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta})\exp(\boldsymbol{\eta}^T\mathbf{u}(\mathbf{x}))$
conjugate prior: $p(\boldsymbol{\eta}|\boldsymbol{\lambda},\nu) = f(\boldsymbol{\lambda},\nu)g(\boldsymbol{\eta})^{\nu}\exp(\nu\boldsymbol{\eta}^T\boldsymbol{\lambda})$

Similar to exponential family distribution ($\langle\boldsymbol{\eta}\rangle$ w.r.t $p(\boldsymbol{\eta}|\boldsymbol{\lambda},\nu)$):

$$\boxed{\langle\boldsymbol{\eta}\rangle = -\frac{\nabla_{\boldsymbol{\lambda}}\log(f(\boldsymbol{\lambda},\nu))}{\nu}}$$

$$\boxed{\langle\log(g(\boldsymbol{\eta}))\rangle + \boldsymbol{\lambda}^T\langle\boldsymbol{\eta}\rangle = -\frac{\partial\log(f(\boldsymbol{\lambda},\nu))}{\partial\nu}}$$

Important for variational learning: KL-divergence between p(osteri)ors with different parameters

$$\boxed{D(p(\boldsymbol{\eta}|\tilde{\boldsymbol{\lambda}},\tilde{\nu})||p(\boldsymbol{\eta}|\boldsymbol{\lambda},\nu)) = \log\left(\frac{f(\tilde{\boldsymbol{\lambda}},\tilde{\nu})}{f(\boldsymbol{\lambda},\nu)}\right) - (\tilde{\nu}-\nu)\frac{\partial\log(f(\tilde{\boldsymbol{\lambda}},\tilde{\nu}))}{\partial\tilde{\nu}} + (\tilde{\boldsymbol{\lambda}}^T - \boldsymbol{\lambda}^T)\nu\langle\boldsymbol{\eta}\rangle}$$

(where $\langle\boldsymbol{\eta}\rangle = \langle\boldsymbol{\eta}\rangle_{p(\boldsymbol{\eta}|\tilde{\boldsymbol{\lambda}},\tilde{\nu})}$)

Learning $\mathbf{P}_C$ and $\mathbf{P}_F$

Reminder:
$\mathbf{P}_C$: probability that coin shows 'heads' when loaded. $\mathbf{P}_F$: probability that coin is fair $\in [0, 1]$.
Approximating density $\left[\prod_{i=1}^N Q_i(F_i)\right] q(\mathbf{P}_F)q(\mathbf{P}_C)$

We wish to maximize (E_q is expectation w.r.t approximating density):

$$
\begin{aligned}
\mathcal{L} &= \sum_{i=1}^{N} E_q\left[\log\left(\frac{P(C_i|F_i,\mathbf{P}_C)P(F_i|\mathbf{P}_F)}{Q_i(F_i)}\right)\right] \\
&\quad - E_q\left[\log\left(\frac{q(\mathbf{P}_F)}{p(\mathbf{P}_F)}\right)\right] - E_q\left[\log\left(\frac{q(\mathbf{P}_C)}{p(\mathbf{P}_C)}\right)\right]
\end{aligned}
$$

Learning $\mathbf{P}_C$ and $\mathbf{P}_F$

Reminder:
$\mathbf{P}_C$: probability that coin shows 'heads' when loaded. $\mathbf{P}_F$: probability that coin is fair $\in [0, 1]$.
Approximating density $\left[\prod_{i=1}^{N} Q_i(F_i)\right] q(\mathbf{P}_F)q(\mathbf{P}_C)$

We wish to maximize (E_q is expectation w.r.t approximating density):

$$\begin{aligned}\mathcal{L} &= \sum_{i=1}^{N} E_q\left[\log\left(\frac{P(C_i|F_i,\mathbf{P}_C)P(F_i|\mathbf{P}_F)}{Q_i(F_i)}\right)\right] \\ &\quad - E_q\left[\log\left(\frac{q(\mathbf{P}_F)}{p(\mathbf{P}_F)}\right)\right] - E_q\left[\log\left(\frac{q(\mathbf{P}_C)}{p(\mathbf{P}_C)}\right)\right]\end{aligned}$$

- We saw how to maximize $\mathcal{L}$ w.r.t. $Q_i(F_i)$
- We will now maximize w.r.t. $\mathbf{P}_F$ (and $\mathbf{P}_C$ as an exercise)
- for this, we only consider parts of $\mathcal{L}$ depending on $q(\mathbf{P}_F)$.

Maximizing $\mathcal{L}$

Reminder:
$\mathbf{P}_C$: probability that coin shows 'heads' when loaded. $\mathbf{P}_F$: probability that coin is fair $\in [0, 1]$.
Approximating density $\left[\prod_{i=1}^{N} Q_i(F_i)\right] q(\mathbf{P}_F)q(\mathbf{P}_C)$

$$\begin{aligned}
\mathcal{L} &= \sum_{i=1}^{N} E_q\left[\log\left(P(F_i|\mathbf{P}_F)\right)\right] - E_q\left[\log\left(\frac{q(\mathbf{P}_F)}{p(\mathbf{P}_F)}\right)\right] + C \\
&= \sum_{i=1}^{N} E_q\left[\log\left(P(F_i|\mathbf{P}_F)\right)\right] - D\left(q(\mathbf{P}_F)||p(\mathbf{P}_F)\right) + C \\
&= \mathcal{L}_{\mathcal{F}} + C
\end{aligned}$$

- Maximize $\mathcal{L}_{\mathcal{F}}$ w.r.t. $\mathbf{P}_F \Rightarrow$ maximize $\mathcal{L}$ w.r.t. $\mathbf{P}_F$.
- C contains all parts of $\mathcal{L}$ not depending on $\mathbf{P}_F$
- $\mathbf{P}_F$ and $\mathbf{P}_C$ do not interact when $Q_i(F_i)$ is fixed $\Rightarrow$ can optimize independently!

Maximizing $\mathcal{L}$, contd.

Reminder:
$\mathbf{P}_C$: probability that coin shows 'heads' when loaded. $\mathbf{P}_F$: probability that coin is fair $\in [0, 1]$.
Approximating density $\left[\prod_{i=1}^{N} Q_i(F_i)\right] q(\mathbf{P}_F)q(\mathbf{P}_C)$
$\mathcal{L}_{\mathcal{F}} = \sum_{i=1}^{N} E_q \left[\log\left(P(F_i|\mathbf{P}_F)\right)\right] - D\left(q(\mathbf{P}_F)||p(\mathbf{P}_F)\right)$

To make this maximization tractable (and more general), assume that likelihoods and p(oste)riors are in the exponential family:

$$\begin{aligned}
\boldsymbol{\eta} &= \eta(\mathbf{P}_F) \\
P(F_i|\mathbf{P}_F) &= p(F_i|\boldsymbol{\eta}) = h(F_i)g(\boldsymbol{\eta})\exp(\boldsymbol{\eta}^T\mathbf{u}(F_i)) \\
p(\mathbf{P}_F) &= p(\boldsymbol{\eta}|\boldsymbol{\lambda}, \tilde{\nu}) = f(\boldsymbol{\lambda}, \nu)g(\boldsymbol{\eta})^{\nu}\exp(\nu\boldsymbol{\eta}^T\boldsymbol{\lambda}) \\
q(\mathbf{P}_F) &= p(\boldsymbol{\eta}|\tilde{\boldsymbol{\lambda}}, \tilde{\nu}) = f(\tilde{\boldsymbol{\lambda}}, \tilde{\nu})g(\boldsymbol{\eta})^{\tilde{\nu}}\exp(\tilde{\nu}\boldsymbol{\eta}^T\tilde{\boldsymbol{\lambda}})
\end{aligned}$$

Maximizing $\mathcal{L}$, contd.

Reminder:

$\mathbf{P}_C$: probability that coin shows 'heads' when loaded. $\mathbf{P}_F$: probability that coin is fair $\in [0, 1]$.

Approximating density $\left[\prod_{i=1}^{N} Q_i(F_i)\right] q(\mathbf{P}_F)q(\mathbf{P}_C)$

$\mathcal{L}_{\mathcal{F}} = \sum_{i=1}^{N} E_q \left[\log\left(P(F_i|\mathbf{P}_F)\right)\right] - D\left(q(\mathbf{P}_F)||p(\mathbf{P}_F)\right)$

$P(F_i|\mathbf{P}_F) = p(F_i|\boldsymbol{\eta}) = h(F_i)g(\boldsymbol{\eta})\exp(\boldsymbol{\eta}^T\mathbf{u}(F_i))$

$p(\mathbf{P}_F) = p(\boldsymbol{\eta}|\boldsymbol{\lambda}, \tilde{\nu}) = f(\boldsymbol{\lambda}, \nu)g(\boldsymbol{\eta})^{\nu}\exp(\nu\boldsymbol{\eta}^T\boldsymbol{\lambda})$

$q(\mathbf{P}_F) = p(\boldsymbol{\eta}|\tilde{\boldsymbol{\lambda}}, \tilde{\nu}) = f(\tilde{\boldsymbol{\lambda}}, \tilde{\nu})g(\boldsymbol{\eta})^{\tilde{\nu}}\exp(\tilde{\nu}\boldsymbol{\eta}^T\tilde{\boldsymbol{\lambda}})$

Computing the terms of $\mathcal{L}_{\mathcal{F}}$:

$$\begin{aligned}
E_q\left[\log\left(P(F_i|\mathbf{P}_F)\right)\right] &= \langle\log(h(F_i))\rangle_{Q_i(F_i)} + \langle\log(g(\boldsymbol{\eta}))\rangle_{q(\mathbf{P}_F)} \\
&= \langle\boldsymbol{\eta}\rangle_{q(\mathbf{P}_F)}\,\langle\mathbf{u}(F_i)\rangle_{Q_i(F_i)} \\
D\left(q(\mathbf{P}_F)||p(\mathbf{P}_F)\right) &= \log\left(\frac{f(\tilde{\boldsymbol{\lambda}}, \tilde{\nu})}{f(\boldsymbol{\lambda}, \nu)}\right) - (\tilde{\nu} - \nu)\frac{\partial\log(f(\tilde{\boldsymbol{\lambda}}, \tilde{\nu}))}{\partial\tilde{\nu}} \\
&\quad + (\tilde{\boldsymbol{\lambda}}^T - \boldsymbol{\lambda}^T)\nu\langle\boldsymbol{\eta}\rangle_{q(\mathbf{P}_F)}
\end{aligned}$$

Maximizing $\mathcal{L}$, contd.

Reminder:
$\mathcal{L}_{\mathcal{F}} = \sum_{i=1}^{N} E_q \left[\log\left(P(F_i|\mathbf{P}_F)\right)\right] - D\left(q(\mathbf{P}_F)||p(\mathbf{P}_F)\right)$
$E_q\left[\log\left(P(F_i|\mathbf{P}_F)\right)\right] = \langle \log(h(F_i))\rangle_{Q_i(F_i)} + \langle \log(g(\boldsymbol{\eta}))\rangle_{q(\mathbf{P}_F)} + \langle \boldsymbol{\eta}\rangle_{q(\mathbf{P}_F)} \langle \mathbf{u}(F_i)\rangle_{Q_i(F_i)}$
$D\left(q(\mathbf{P}_F)||p(\mathbf{P}_F)\right) = \log\left(\frac{f(\tilde{\boldsymbol{\lambda}},\tilde{\nu})}{f(\boldsymbol{\lambda},\nu)}\right) - (\tilde{\nu} - \nu)\frac{\partial \log(f(\tilde{\boldsymbol{\lambda}},\tilde{\nu}))}{\partial \tilde{\nu}} + (\tilde{\boldsymbol{\lambda}}^T - \boldsymbol{\lambda}^T)\nu\langle \boldsymbol{\eta}\rangle_{q(\mathbf{P}_F)}$

Extremum of $\mathcal{L}_{\mathcal{F}}$ can be found with (convex) optimizer or by setting derivatives to zero:

$$\begin{aligned}
\nabla_{\tilde{\boldsymbol{\lambda}}}\mathcal{L}_{\mathcal{F}} &= \sum_{i=1}^{N} \nabla_{\tilde{\boldsymbol{\lambda}}} E_q\left[\log\left(P(F_i|\mathbf{P}_F)\right)\right] - \nabla_{\tilde{\boldsymbol{\lambda}}} D\left(q(\mathbf{P}_F)||p(\mathbf{P}_F)\right) = \mathbf{0} \\
\frac{\partial \mathcal{L}_{\mathcal{F}}}{\partial \tilde{\nu}} &= \sum_{i=1}^{N} \frac{\partial E_q\left[\log\left(P(F_i|\mathbf{P}_F)\right)\right]}{\partial \tilde{\nu}} - \frac{\partial D\left(q(\mathbf{P}_F)||p(\mathbf{P}_F)\right)}{\partial \tilde{\nu}} = 0
\end{aligned}$$

The derivatives of $\mathcal{L}_\mathcal{F}$

Reminder:
$\nabla_{\tilde{\boldsymbol{\lambda}}} \mathcal{L}_\mathcal{F} = \sum_{i=1}^N \nabla_{\tilde{\boldsymbol{\lambda}}} E_q \left[\log \left(P(F_i|\mathbf{P}_F)\right)\right] - \nabla_{\tilde{\boldsymbol{\lambda}}} D\left(q(\mathbf{P}_F)||p(\mathbf{P}_F)\right)$
$\frac{\partial \mathcal{L}_\mathcal{F}}{\partial \tilde{\nu}} = \sum_{i=1}^N \frac{\partial E_q[\log(P(F_i|\mathbf{P}_F))]}{\partial \tilde{\nu}} - \frac{\partial D(q(\mathbf{P}_F)||p(\mathbf{P}_F))}{\partial \tilde{\nu}}$
$E_q \left[\log \left(P(F_i|\mathbf{P}_F)\right)\right] = \langle \log(h(F_i))\rangle_{Q_i(F_i)} + \langle \log(g(\boldsymbol{\eta}))\rangle_{q(\mathbf{P}_F)} + \langle \boldsymbol{\eta}\rangle_{q(\mathbf{P}_F)} \; \langle \mathbf{u}(F_i)\rangle_{Q_i(F_i)}$
$D\left(q(\mathbf{P}_F)||p(\mathbf{P}_F)\right) = \log\left(\frac{f(\tilde{\boldsymbol{\lambda}},\tilde{\nu})}{f(\boldsymbol{\lambda},\nu)}\right) - (\tilde{\nu} - \nu)\frac{\partial \log(f(\tilde{\boldsymbol{\lambda}},\tilde{\nu}))}{\partial \tilde{\nu}} + (\tilde{\boldsymbol{\lambda}}^T - \boldsymbol{\lambda}^T)\nu\langle \boldsymbol{\eta}\rangle_{q(\mathbf{P}_F)}$
Exponential family expectations: $\langle \boldsymbol{\eta}\rangle = -\frac{\nabla_{\boldsymbol{\lambda}} \log(f(\boldsymbol{\lambda},\nu))}{\nu}$, $\langle \log(g(\boldsymbol{\eta}))\rangle + \boldsymbol{\lambda}^T\langle \boldsymbol{\eta}\rangle = -\frac{\partial \log(f(\boldsymbol{\lambda},\nu))}{\partial \nu}$

Using the properties of exponential family distributions:

$$
\begin{aligned}
\nabla_{\tilde{\boldsymbol{\lambda}}} E_q \left[\log \left(P(F_i|\mathbf{P}_F)\right)\right] &= -\frac{\partial^2 \log(f(\tilde{\boldsymbol{\lambda}},\tilde{\nu}))}{\partial \tilde{\boldsymbol{\lambda}} \partial \tilde{\nu}} - \langle \boldsymbol{\eta}\rangle_{q(\mathbf{P}_F)} - \tilde{\boldsymbol{\lambda}}^T \nabla_{\tilde{\boldsymbol{\lambda}}}\langle \boldsymbol{\eta}\rangle_{q(\mathbf{P}_F)} \\
&\quad + \nabla_{\tilde{\boldsymbol{\lambda}}}\langle \boldsymbol{\eta}\rangle_{q(\mathbf{P}_F)} \; \langle \mathbf{u}(F_i)\rangle_{Q_i(F_i)} \\
\nabla_{\tilde{\boldsymbol{\lambda}}} D\left(q(\mathbf{P}_F)||p(\mathbf{P}_F)\right) &= -\tilde{\nu}\langle \boldsymbol{\eta}\rangle_{q(\mathbf{P}_F)} - (\tilde{\nu} - \nu)\frac{\partial^2 \log(f(\tilde{\boldsymbol{\lambda}},\tilde{\nu}))}{\partial \tilde{\boldsymbol{\lambda}} \partial \tilde{\nu}} \\
&\quad + \nu\langle \boldsymbol{\eta}\rangle_{q(\mathbf{P}_F)} + (\tilde{\boldsymbol{\lambda}}^T - \boldsymbol{\lambda}^T)\nu\nabla_{\tilde{\boldsymbol{\lambda}}}\langle \boldsymbol{\eta}\rangle_{q(\mathbf{P}_F)}
\end{aligned}
$$

The derivatives of $\mathcal{L}_{\mathcal{F}}$

Reminder:

$\nabla_{\tilde{\boldsymbol{\lambda}}} \mathcal{L}_{\mathcal{F}} = \sum_{i=1}^{N} \nabla_{\tilde{\boldsymbol{\lambda}}} E_q \left[\log \left(P(F_i|\mathbf{P}_F)\right)\right] - \nabla_{\tilde{\boldsymbol{\lambda}}} D\left(q(\mathbf{P}_F)||p(\mathbf{P}_F)\right)$

$\frac{\partial \mathcal{L}_{\mathcal{F}}}{\partial \tilde{\nu}} = \sum_{i=1}^{N} \frac{\partial E_q[\log(P(F_i|\mathbf{P}_F))]}{\partial \tilde{\nu}} - \frac{\partial D(q(\mathbf{P}_F)||p(\mathbf{P}_F))}{\partial \tilde{\nu}}$

$E_q \left[\log \left(P(F_i|\mathbf{P}_F)\right)\right] = \langle \log(h(F_i))\rangle_{Q_i(F_i)} + \langle \log(g(\boldsymbol{\eta}))\rangle_{q(\mathbf{P}_F)} + \langle \boldsymbol{\eta}\rangle_{q(\mathbf{P}_F)} \; \langle \mathbf{u}(F_i)\rangle_{Q_i(F_i)}$

$D\left(q(\mathbf{P}_F)||p(\mathbf{P}_F)\right) = \log\left(\frac{f(\tilde{\boldsymbol{\lambda}},\tilde{\nu})}{f(\boldsymbol{\lambda},\nu)}\right) - (\tilde{\nu}-\nu)\frac{\partial \log(f(\tilde{\boldsymbol{\lambda}},\tilde{\nu}))}{\partial \tilde{\nu}} + (\tilde{\boldsymbol{\lambda}}^T - \boldsymbol{\lambda}^T)\nu\langle \boldsymbol{\eta}\rangle_{q(\mathbf{P}_F)}$

Exponential family expectations: $\langle \boldsymbol{\eta}\rangle = -\frac{\nabla_{\boldsymbol{\lambda}} \log(f(\boldsymbol{\lambda},\nu))}{\nu}$, $\langle \log(g(\boldsymbol{\eta}))\rangle + \boldsymbol{\lambda}^T\langle \boldsymbol{\eta}\rangle = -\frac{\partial \log(f(\boldsymbol{\lambda},\nu))}{\partial \nu}$

Using the properties of exponential family distributions:

$$\begin{aligned}
\nabla_{\tilde{\boldsymbol{\lambda}}} E_q \left[\log \left(P(F_i|\mathbf{P}_F)\right)\right] &= -\frac{\partial^2 \log(f(\tilde{\boldsymbol{\lambda}},\tilde{\nu}))}{\partial \tilde{\boldsymbol{\lambda}}\partial \tilde{\nu}} - \langle \boldsymbol{\eta}\rangle_{q(\mathbf{P}_F)} - \tilde{\boldsymbol{\lambda}}^T \nabla_{\tilde{\boldsymbol{\lambda}}}\langle \boldsymbol{\eta}\rangle_{q(\mathbf{P}_F)} \\
&\quad + \nabla_{\tilde{\boldsymbol{\lambda}}}\langle \boldsymbol{\eta}\rangle_{q(\mathbf{P}_F)} \; \langle \mathbf{u}(F_i)\rangle_{Q_i(F_i)} \\
\nabla_{\tilde{\boldsymbol{\lambda}}} D\left(q(\mathbf{P}_F)||p(\mathbf{P}_F)\right) &= -\tilde{\nu}\langle \boldsymbol{\eta}\rangle_{q(\mathbf{P}_F)} - (\tilde{\nu}-\nu)\frac{\partial^2 \log(f(\tilde{\boldsymbol{\lambda}},\tilde{\nu}))}{\partial \tilde{\boldsymbol{\lambda}}\partial \tilde{\nu}} \\
&\quad + \nu\langle \boldsymbol{\eta}\rangle_{q(\mathbf{P}_F)} + (\tilde{\boldsymbol{\lambda}}^T - \boldsymbol{\lambda}^T)\nu\nabla_{\tilde{\boldsymbol{\lambda}}}\langle \boldsymbol{\eta}\rangle_{q(\mathbf{P}_F)}
\end{aligned}$$

Note: for N observations/datapoints, there are N many $\nabla_{\tilde{\boldsymbol{\lambda}}} E_q[]$ terms.

When is $\nabla_{\tilde{\boldsymbol{\lambda}}} \mathcal{L}_{\mathcal{F}} = 0$?

Reminder:

$\nabla_{\tilde{\boldsymbol{\lambda}}} \mathcal{L}_{\mathcal{F}} = \sum_{i=1}^{N} \nabla_{\tilde{\boldsymbol{\lambda}}} E_q \left[\log \left(P(F_i|\mathbf{P}_F)\right)\right] - \nabla_{\tilde{\boldsymbol{\lambda}}} D\left(q(\mathbf{P}_F)||p(\mathbf{P}_F)\right)$

$\frac{\partial \mathcal{L}_{\mathcal{F}}}{\partial \tilde{\nu}} = \sum_{i=1}^{N} \frac{\partial E_q[\log(P(F_i|\mathbf{P}_F))]}{\partial \tilde{\nu}} - \frac{\partial D(q(\mathbf{P}_F)||p(\mathbf{P}_F))}{\partial \tilde{\nu}}$

$E_q \left[\log \left(P(F_i|\mathbf{P}_F)\right)\right] = \langle \log(h(F_i)) \rangle_{Q_i(F_i)} + \langle \log(g(\boldsymbol{\eta})) \rangle_{q(\mathbf{P}_F)} + \langle \boldsymbol{\eta} \rangle_{q(\mathbf{P}_F)} \langle \mathbf{u}(F_i) \rangle_{Q_i(F_i)}$

$D\left(q(\mathbf{P}_F)||p(\mathbf{P}_F)\right) = \log\left(\frac{f(\tilde{\boldsymbol{\lambda}},\tilde{\nu})}{f(\boldsymbol{\lambda},\nu)}\right) - (\tilde{\nu} - \nu)\frac{\partial \log(f(\tilde{\boldsymbol{\lambda}},\tilde{\nu}))}{\partial \tilde{\nu}} + (\tilde{\boldsymbol{\lambda}}^T - \boldsymbol{\lambda}^T)\nu \langle \boldsymbol{\eta} \rangle_{q(\mathbf{P}_F)}$

Exponential family expectations: $\langle \boldsymbol{\eta} \rangle = -\frac{\nabla_{\boldsymbol{\lambda}} \log(f(\boldsymbol{\lambda},\nu))}{\nu}$, $\langle \log(g(\boldsymbol{\eta})) \rangle + \boldsymbol{\lambda}^T \langle \boldsymbol{\eta} \rangle = -\frac{\partial \log(f(\boldsymbol{\lambda},\nu))}{\partial \nu}$

In total, after collecting terms:

$$
\begin{aligned}
\nabla_{\tilde{\boldsymbol{\lambda}}} \mathcal{L}_{\mathcal{F}} \quad = \quad & \frac{\partial^2 \log(f(\tilde{\boldsymbol{\lambda}},\tilde{\nu}))}{\partial \tilde{\boldsymbol{\lambda}} \partial \tilde{\nu}} (\tilde{\nu} - (\nu + N)) \\
& + \langle \boldsymbol{\eta} \rangle_{q(\mathbf{P}_F)} (\tilde{\nu} - (\nu + N)) \\
& + \left(-\tilde{\boldsymbol{\lambda}}^T (N + \nu) + \nu \boldsymbol{\lambda}^T + \sum_i \langle \mathbf{u}(F_i) \rangle_{Q(F_i)} \right) \nabla_{\tilde{\boldsymbol{\lambda}}} \langle \boldsymbol{\eta} \rangle_{q(\mathbf{P}_F)}
\end{aligned}
$$

When is $\nabla_{\tilde{\boldsymbol{\lambda}}} \mathcal{L}_{\mathcal{F}} = 0$?

Reminder:

$\nabla_{\tilde{\boldsymbol{\lambda}}} \mathcal{L}_{\mathcal{F}} = \sum_{i=1}^{N} \nabla_{\tilde{\boldsymbol{\lambda}}} E_q \left[\log\left(P(F_i|\mathbf{P}_F)\right)\right] - \nabla_{\tilde{\boldsymbol{\lambda}}} D\left(q(\mathbf{P}_F)||p(\mathbf{P}_F)\right)$

$\frac{\partial \mathcal{L}_{\mathcal{F}}}{\partial \tilde{\nu}} = \sum_{i=1}^{N} \frac{\partial E_q[\log(P(F_i|\mathbf{P}_F))]}{\partial \tilde{\nu}} - \frac{\partial D(q(\mathbf{P}_F)||p(\mathbf{P}_F))}{\partial \tilde{\nu}}$

$E_q \left[\log\left(P(F_i|\mathbf{P}_F)\right)\right] = \langle \log(h(F_i))\rangle_{Q_i(F_i)} + \langle \log(g(\boldsymbol{\eta}))\rangle_{q(\mathbf{P}_F)} + \langle \boldsymbol{\eta}\rangle_{q(\mathbf{P}_F)} \langle \mathbf{u}(F_i)\rangle_{Q_i(F_i)}$

$D\left(q(\mathbf{P}_F)||p(\mathbf{P}_F)\right) = \log\left(\frac{f(\tilde{\boldsymbol{\lambda}},\tilde{\nu})}{f(\boldsymbol{\lambda},\nu)}\right) - (\tilde{\nu} - \nu)\frac{\partial \log(f(\tilde{\boldsymbol{\lambda}},\tilde{\nu}))}{\partial \tilde{\nu}} + (\tilde{\boldsymbol{\lambda}}^T - \boldsymbol{\lambda}^T)\nu\langle \boldsymbol{\eta}\rangle_{q(\mathbf{P}_F)}$

Exponential family expectations: $\langle \boldsymbol{\eta}\rangle = -\frac{\nabla_{\boldsymbol{\lambda}} \log(f(\boldsymbol{\lambda},\nu))}{\nu}$, $\langle \log(g(\boldsymbol{\eta}))\rangle + \boldsymbol{\lambda}^T\langle \boldsymbol{\eta}\rangle = -\frac{\partial \log(f(\boldsymbol{\lambda},\nu))}{\partial \nu}$

In total, after collecting terms:

$$
\begin{aligned}
\nabla_{\tilde{\boldsymbol{\lambda}}} \mathcal{L}_{\mathcal{F}} \quad = \quad & \frac{\partial^2 \log(f(\tilde{\boldsymbol{\lambda}},\tilde{\nu}))}{\partial \tilde{\boldsymbol{\lambda}} \partial \tilde{\nu}} (\tilde{\nu} - (\nu + N)) \\
& + \langle \boldsymbol{\eta}\rangle_{q(\mathbf{P}_F)} (\tilde{\nu} - (\nu + N)) \\
& + \left(-\tilde{\boldsymbol{\lambda}}^T (N+\nu) + \nu \boldsymbol{\lambda}^T + \sum_i \langle \mathbf{u}(F_i)\rangle_{Q(F_i)}\right) \nabla_{\tilde{\boldsymbol{\lambda}}} \langle \boldsymbol{\eta}\rangle_{q(\mathbf{P}_F)}
\end{aligned}
$$

Hence for $\nabla_{\tilde{\boldsymbol{\lambda}}} \mathcal{L}_{\mathcal{F}} = 0$, it is sufficient (and generally necessary) that

$$
\begin{aligned}
\tilde{\nu} &= \nu + N \\
\tilde{\boldsymbol{\lambda}} &= \frac{\nu \boldsymbol{\lambda}^T + \sum_i \langle \mathbf{u}(F_i)\rangle_{Q(F_i)}}{\nu + N}
\end{aligned}
$$

When is $\frac{\partial \mathcal{L}_{\mathcal{F}}}{\partial \tilde{\nu}} = 0$

Reminder:

$\frac{\partial \mathcal{L}_{\mathcal{F}}}{\partial \tilde{\nu}} = \sum_{i=1}^{N} \frac{\partial E_q[\log(P(F_i|\mathbf{P}_F))]}{\partial \tilde{\nu}} - \frac{\partial D(q(\mathbf{P}_F)||p(\mathbf{P}_F))}{\partial \tilde{\nu}}$

$E_q\left[\log\left(P(F_i|\mathbf{P}_F)\right)\right] = \langle \log(h(F_i))\rangle_{Q_i(F_i)} + \langle \log(g(\boldsymbol{\eta}))\rangle_{q(\mathbf{P}_F)} + \langle \boldsymbol{\eta}\rangle_{q(\mathbf{P}_F)} \ \langle \mathbf{u}(F_i)\rangle_{Q_i(F_i)}$

$D\left(q(\mathbf{P}_F)||p(\mathbf{P}_F)\right) = \log\left(\frac{f(\tilde{\boldsymbol{\lambda}},\tilde{\nu})}{f(\boldsymbol{\lambda},\nu)}\right) - (\tilde{\nu} - \nu)\frac{\partial \log(f(\tilde{\boldsymbol{\lambda}},\tilde{\nu}))}{\partial \tilde{\nu}} + (\tilde{\boldsymbol{\lambda}}^T - \boldsymbol{\lambda}^T)\nu\langle \boldsymbol{\eta}\rangle_{q(\mathbf{P}_F)}$

Exponential family expectations: $\langle \boldsymbol{\eta}\rangle = -\frac{\nabla_{\boldsymbol{\lambda}} \log(f(\boldsymbol{\lambda},\nu))}{\nu}$, $\langle \log(g(\boldsymbol{\eta}))\rangle + \boldsymbol{\lambda}^T\langle \boldsymbol{\eta}\rangle = -\frac{\partial \log(f(\boldsymbol{\lambda},\nu))}{\partial \nu}$

Using the properties of exponential family distributions:

$$\begin{aligned}
\frac{\partial E_q\left[\log\left(P(F_i|\mathbf{P}_F)\right)\right]}{\partial \tilde{\nu}} &= -\frac{\partial^2 \log(f(\tilde{\boldsymbol{\lambda}},\tilde{\nu}))}{\partial \tilde{\nu}^2} - \tilde{\boldsymbol{\lambda}}^T \frac{\partial \langle \boldsymbol{\eta}\rangle_{q(\mathbf{P}_F)}}{\partial \tilde{\nu}} \\
&\quad + \frac{\langle \boldsymbol{\eta}\rangle_{q(\mathbf{P}_F)}}{\partial \tilde{\nu}} \ \langle \mathbf{u}(F_i)\rangle_{Q_i(F_i)} \\
\frac{\partial D\left(q(\mathbf{P}_F)||p(\mathbf{P}_F)\right)}{\partial \tilde{\nu}} &= \frac{\partial \log(f(\tilde{\boldsymbol{\lambda}},\tilde{\nu}))}{\partial \tilde{\nu}} - \frac{\partial \log(f(\tilde{\boldsymbol{\lambda}},\tilde{\nu}))}{\partial \tilde{\nu}} - (\tilde{\nu} - \nu)\frac{\partial^2 \log(f(\tilde{\boldsymbol{\lambda}},\tilde{\nu}))}{\partial \tilde{\nu}^2} \\
&\quad + (\tilde{\boldsymbol{\lambda}}^T - \boldsymbol{\lambda}^T)\nu\frac{\partial \langle \boldsymbol{\eta}\rangle_{q(\mathbf{P}_F)}}{\partial \tilde{\nu}}
\end{aligned}$$

When is $\frac{\partial \mathcal{L}_{\mathcal{F}}}{\partial \tilde{\nu}} = 0$

Reminder:

$\frac{\partial \mathcal{L}_{\mathcal{F}}}{\partial \tilde{\nu}} = \sum_{i=1}^{N} \frac{\partial E_q[\log(P(F_i|\mathbf{P}_F))]}{\partial \tilde{\nu}} - \frac{\partial D(q(\mathbf{P}_F)||p(\mathbf{P}_F))}{\partial \tilde{\nu}}$

$E_q\left[\log\left(P(F_i|\mathbf{P}_F)\right)\right] = \langle \log(h(F_i))\rangle_{Q_i(F_i)} + \langle \log(g(\boldsymbol{\eta}))\rangle_{q(\mathbf{P}_F)} + \langle \boldsymbol{\eta}\rangle_{q(\mathbf{P}_F)} \langle \mathbf{u}(F_i)\rangle_{Q_i(F_i)}$

$D\left(q(\mathbf{P}_F)||p(\mathbf{P}_F)\right) = \log\left(\frac{f(\tilde{\boldsymbol{\lambda}},\tilde{\nu})}{f(\boldsymbol{\lambda},\nu)}\right) - (\tilde{\nu}-\nu)\frac{\partial \log(f(\tilde{\boldsymbol{\lambda}},\tilde{\nu}))}{\partial \tilde{\nu}} + (\tilde{\boldsymbol{\lambda}}^T - \boldsymbol{\lambda}^T)\nu\langle\boldsymbol{\eta}\rangle_{q(\mathbf{P}_F)}$

Exponential family expectations: $\langle\boldsymbol{\eta}\rangle = -\frac{\nabla_{\boldsymbol{\lambda}} \log(f(\boldsymbol{\lambda},\nu))}{\nu}$, $\langle \log(g(\boldsymbol{\eta}))\rangle + \boldsymbol{\lambda}^T\langle\boldsymbol{\eta}\rangle = -\frac{\partial \log(f(\boldsymbol{\lambda},\nu))}{\partial \nu}$

Using the properties of exponential family distributions:

$$
\begin{aligned}
\frac{\partial E_q\left[\log\left(P(F_i|\mathbf{P}_F)\right)\right]}{\partial \tilde{\nu}} &= -\frac{\partial^2 \log(f(\tilde{\boldsymbol{\lambda}},\tilde{\nu}))}{\partial \tilde{\nu}^2} - \tilde{\boldsymbol{\lambda}}^T\frac{\partial\langle\boldsymbol{\eta}\rangle_{q(\mathbf{P}_F)}}{\partial\tilde{\nu}} \\
&\quad + \frac{\langle\boldsymbol{\eta}\rangle_{q(\mathbf{P}_F)}}{\partial\tilde{\nu}} \langle\mathbf{u}(F_i)\rangle_{Q_i(F_i)} \\
\frac{\partial D\left(q(\mathbf{P}_F)||p(\mathbf{P}_F)\right)}{\partial\tilde{\nu}} &= \frac{\partial \log(f(\tilde{\boldsymbol{\lambda}},\tilde{\nu}))}{\partial\tilde{\nu}} - \frac{\partial \log(f(\tilde{\boldsymbol{\lambda}},\tilde{\nu}))}{\partial\tilde{\nu}} - (\tilde{\nu}-\nu)\frac{\partial^2 \log(f(\tilde{\boldsymbol{\lambda}},\tilde{\nu}))}{\partial\tilde{\nu}^2} \\
&\quad + (\tilde{\boldsymbol{\lambda}}^T - \boldsymbol{\lambda}^T)\nu\frac{\partial\langle\boldsymbol{\eta}\rangle_{q(\mathbf{P}_F)}}{\partial\tilde{\nu}}
\end{aligned}
$$

Note: for N observations/datapoints, there are N many $\frac{\partial E_q[]}{\partial\tilde{\nu}}$ terms.

When is $\frac{\partial \mathcal{L}_{\mathcal{F}}}{\partial \tilde{\nu}} = 0$?

Reminder:

$\frac{\partial \mathcal{L}_{\mathcal{F}}}{\partial \tilde{\nu}} = \sum_{i=1}^{N} \frac{\partial E_q[\log(P(F_i|\mathbf{P}_F))]}{\partial \tilde{\nu}} - \frac{\partial D(q(\mathbf{P}_F)||p(\mathbf{P}_F))}{\partial \tilde{\nu}}$

$E_q\left[\log\left(P(F_i|\mathbf{P}_F)\right)\right] = \langle \log(h(F_i))\rangle_{Q_i(F_i)} + \langle \log(g(\boldsymbol{\eta}))\rangle_{q(\mathbf{P}_F)} + \langle \boldsymbol{\eta}\rangle_{q(\mathbf{P}_F)} \langle \mathbf{u}(F_i)\rangle_{Q_i(F_i)}$

$D\left(q(\mathbf{P}_F)||p(\mathbf{P}_F)\right) = \log\left(\frac{f(\tilde{\boldsymbol{\lambda}},\tilde{\nu})}{f(\boldsymbol{\lambda},\nu)}\right) - (\tilde{\nu} - \nu)\frac{\partial \log(f(\tilde{\boldsymbol{\lambda}},\tilde{\nu}))}{\partial \tilde{\nu}} + (\tilde{\boldsymbol{\lambda}}^T - \boldsymbol{\lambda}^T)\nu\langle \boldsymbol{\eta}\rangle_{q(\mathbf{P}_F)}$

Exponential family expectations: $\langle \boldsymbol{\eta}\rangle = -\frac{\nabla_{\boldsymbol{\lambda}} \log(f(\boldsymbol{\lambda},\nu))}{\nu}$, $\langle \log(g(\boldsymbol{\eta}))\rangle + \boldsymbol{\lambda}^T\langle \boldsymbol{\eta}\rangle = -\frac{\partial \log(f(\boldsymbol{\lambda},\nu))}{\partial \nu}$

In total, after collecting terms:

$$\begin{aligned}\frac{\partial \mathcal{L}_{\mathcal{F}}}{\partial \tilde{\nu}} &= \frac{\partial^2 \log(f(\tilde{\boldsymbol{\lambda}},\tilde{\nu}))}{\partial \tilde{\nu}^2}(\tilde{\nu} - (\nu + N)) \\ &\quad + \left(\tilde{\boldsymbol{\lambda}}^T(-N - \nu) + \sum_i \langle \mathbf{u}(F_i)\rangle_{Q_i(F_i)} + \nu\tilde{\boldsymbol{\lambda}}\right)\end{aligned}$$

When is $\frac{\partial \mathcal{L}_{\mathcal{F}}}{\partial \tilde{\nu}} = 0$?

Reminder:
$\frac{\partial \mathcal{L}_{\mathcal{F}}}{\partial \tilde{\nu}} = \sum_{i=1}^{N} \frac{\partial E_q[\log(P(F_i|\mathbf{P}_F))]}{\partial \tilde{\nu}} - \frac{\partial D(q(\mathbf{P}_F)||p(\mathbf{P}_F))}{\partial \tilde{\nu}}$
$E_q\left[\log\left(P(F_i|\mathbf{P}_F)\right)\right] = \langle \log(h(F_i)) \rangle_{Q_i(F_i)} + \langle \log(g(\boldsymbol{\eta})) \rangle_{q(\mathbf{P}_F)} + \langle \boldsymbol{\eta} \rangle_{q(\mathbf{P}_F)} \langle \mathbf{u}(F_i) \rangle_{Q_i(F_i)}$
$D\left(q(\mathbf{P}_F)||p(\mathbf{P}_F)\right) = \log\left(\frac{f(\tilde{\boldsymbol{\lambda}},\tilde{\nu})}{f(\boldsymbol{\lambda},\nu)}\right) - (\tilde{\nu} - \nu)\frac{\partial \log(f(\tilde{\boldsymbol{\lambda}},\tilde{\nu}))}{\partial \tilde{\nu}} + (\tilde{\boldsymbol{\lambda}}^T - \boldsymbol{\lambda}^T)\nu\langle \boldsymbol{\eta} \rangle_{q(\mathbf{P}_F)}$
Exponential family expectations: $\langle \boldsymbol{\eta} \rangle = -\frac{\nabla_{\boldsymbol{\lambda}} \log(f(\boldsymbol{\lambda},\nu))}{\nu}$, $\langle \log(g(\boldsymbol{\eta})) \rangle + \boldsymbol{\lambda}^T \langle \boldsymbol{\eta} \rangle = -\frac{\partial \log(f(\boldsymbol{\lambda},\nu))}{\partial \nu}$

In total, after collecting terms:

$$\begin{aligned} \frac{\partial \mathcal{L}_{\mathcal{F}}}{\partial \tilde{\nu}} &= \frac{\partial^2 \log(f(\tilde{\boldsymbol{\lambda}},\tilde{\nu}))}{\partial \tilde{\nu}^2}(\tilde{\nu} - (\nu + N)) \\ &\quad + \left(\tilde{\boldsymbol{\lambda}}^T(-N - \nu) + \sum_i \langle \mathbf{u}(F_i) \rangle_{Q_i(F_i)} + \nu\tilde{\boldsymbol{\lambda}}\right) \end{aligned}$$

Hence, as before, for $\frac{\partial \mathcal{L}_{\mathcal{F}}}{\partial \tilde{\nu}} = 0$, it is sufficient (and generally necessary) that

$$\begin{aligned} \tilde{\nu} &= \nu + N \\ \tilde{\boldsymbol{\lambda}} &= \frac{\nu\boldsymbol{\lambda} + \sum_i \langle \mathbf{u}(F_i) \rangle_{Q(F_i)}}{\nu + N} \end{aligned}$$

Summary: maximizing $\mathcal{L}$ w.r.t. $q(\mathbf{P}_F)$

Both gradient conditions required that

$$\begin{aligned} \tilde{\nu} &= \nu + N \\ \tilde{\boldsymbol{\lambda}} &= \frac{\nu\boldsymbol{\lambda} + \sum_i \langle \mathbf{u}(F_i) \rangle_{Q(F_i)}}{\nu + N} \end{aligned}$$

- $\tilde{\boldsymbol{\lambda}}, \tilde{\nu}$: parameters of approximating posterior
- $\boldsymbol{\lambda}, \nu$: parameters of prior
- This is the *M* (maximize) step of an EM-algorithm
- works in this form for any conjugate p(oste)rior pairs in the exponential family.
- compare to exact exponential family updates!

Summary: the expectation-maximization algorithm

- Variational inference/learning maximizes a lower bound on the marginal $P(D)$.
- Maximization procedure can be decomposed into groups of variables:
  - **Latent variables**: (approximating) distributions of coin loadedness
  - **'Parameters'**: distribution over probability of drawing a fair coin
- Maximization is done for each group separately:
  - **Latent variables**: effectively compute expectations, hence 'E-step'
  - **'Parameters'**: maxmimize bound, hence 'M-step'.
- In Bayesian treatment, both steps similar: compute expectation and maximize
- Can be generalized to more groups of variables (deep models etc.)
- If EM is not possible (or one is too lazy to derive it..): just run optimizer on $\mathcal{L}$.