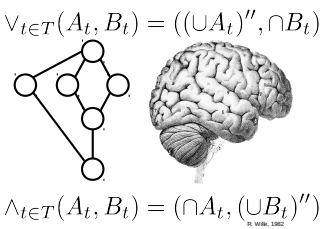
Perceiving, learning and acting with free energy

Bayesian Statistics and Machine Learning

Dominik Endres

April 26, 2018

Philipps Universität Marburg



Outline

Reminder: ingredients of probability

Reminder:
Definition: A σ-algebra over W is a set $\mathcal{F}$ of subsets of W that
- contains W, and
- is closed under union, i.e. if U and V are in $\mathcal{F}$, then so is $U \cup V$, and
- is closed under complementation, i.e. if U is in $\mathcal{F}$, then so is $\bar{U} = \{w : w \in W \wedge w \notin U\}$.

- Probability is built upon the concept of the *possible world* or *elementary outcome* $\in W$.
- Not all possible events need to be assigned a probability, only those that are members of a given σ-algebra $\mathcal{F}$.

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

P1 $P(W) = 1$

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

Reminder: ingredients of probability

Reminder:
Definition: A σ-algebra over W is a set $\mathcal{F}$ of subsets of W that
- contains W, and
- is closed under union, i.e. if U and V are in $\mathcal{F}$, then so is $U \cup V$, and
- is closed under complementation, i.e. if U is in $\mathcal{F}$, then so is $\bar{U} = \{w : w \in W \wedge w \notin U\}$.

- Probability is built upon the concept of the *possible world* or *elementary outcome* $\in W$.
- Not all possible events need to be assigned a probability, only those that are members of a given σ-algebra $\mathcal{F}$.

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

P1 $P(W) = 1$

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

Important consequence: if $\mathcal{F} = 2^W$, then it is sufficent to specify $P(w)$ for all $w \in W$. P2 then allows you to compute $P(U)$ for any $U \subseteq W$.

From here on, we will assume that $\mathcal{F} = 2^W$!

Random variable

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

P1 $P(W) = 1$
P2 If $U, V \in \mathcal{F}$ and $U \cap V = \emptyset$, then $P(U \cup V) = P(U) + P(V)$
We assume that $\mathcal{F} = 2^W$

Definition: a *random variable* X on a set of possible worlds W is a function $X : W \rightarrow Z$ from W to some range Z. If the range is the reals, i.e. $Z \subseteq \mathbb{R}$, then X is also called a *gamble*.

Random variable

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

P1 $P(W) = 1$
P2 If $U, V \in \mathcal{F}$ and $U \cap V = \emptyset$, then $P(U \cup V) = P(U) + P(V)$
We assume that $\mathcal{F} = 2^W$

Definition: a *random variable* X on a set of possible worlds W is a function $X : W \rightarrow Z$ from W to some range Z. If the range is the reals, i.e. $Z \subseteq \mathbb{R}$, then X is also called a *gamble*.

Notes:

- A random variable is neither random, nor is it a variable.
- But its value is unpredictable, if you don't know which $w \in W$ is the 'real world'.
- An instantiation of the value of a random variable (e.g. after you toss a coin) is called a *random variate*.
- *Operationalization* is the process of defining observable random variables
- *Constructs* in Psychology are latent random variables

Probability distribution

Definition: Let Y be a 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 distribution

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

Note:

- Given a probability space $(W, \mathcal{F}, Q)$, and a random variable Y, the corresponding probability distribution over Y can be obtained via $P(Y = y) = \sum_{w: w \in W, Y(w) = y} Q(w)$.
- It is customary to denote the probability distribution over Y by $P(Y)$.
- Instead of writing $P(Y = y)$ for the probability that $Y = y$ under $P(Y)$, it is customary to write $P(y)$.
- Two aspects: structure (this lecture) and content (next lecture)

Use of random variables: computing expectations

Let Y be a *gamble*, i.e. random variable with range $Z \subseteq \mathbb{R}$ and probability distribution $P(Y)$.
The **expected value** or *expectation* of Y w.r.t. $P(Y)$ is defined as

$$\mathrm{E}_{P(Y)}(Y) = \sum_{y \in Z} yP(y)$$

Use of random variables: computing expectations

Let Y be a *gamble*, i.e. random variable with range $Z \subseteq \mathbb{R}$ and probability distribution $P(Y)$.
The **expected value** or *expectation* of Y w.r.t. $P(Y)$ is defined as

$$\mathrm{E}_{P(Y)}(Y) = \sum_{y \in Z} yP(y)$$

Notes:

- $\mathrm{E}_{P(Y)}(Y)$ does not have to be $\in Z$.
- Let Z be the value of a fair die roll. Then $\mathrm{E}_{P(Y)}(Y) = \frac{1}{6}(1 + \ldots + 6) = 3.5$

Joint probability distribution

Reminder:
$P(Y)$ denotes a probability distribution over random variable Y.
$P(y)$ is a shorthand for $P(Y = y)$.

Definition: Let $X_1, \ldots, X_N$ be random variables with ranges $Z_1, \ldots, Z_N$. A **joint probability distribution** $P(X_1, \ldots, X_N)$ is a function $P : \prod_{i=1}^{N} Z_i \rightarrow [0, 1]$ such that

$$\sum_{x_1 \in Z_1} \ldots \sum_{x_N \in Z_N} P(x_1, \ldots, x_N) = 1.$$

Example: joint probability distribution

Reminder:
Assume: the set W of possible worlds can be factorized into two sets, $W = D \times H$, i.e. the elements of W are tuples $j = (d, h)$ where $d \in D$ and $h \in H$.
Let D be the possible elementary outcomes of rolling a die, $D = \{d_1, \ldots, d_6\}$.
Let $H = \{h_1, h_2\}$ be the set of hypotheses h_1 ='the die is fair', and h_2 ='the die will show only the numbers 1,2,3'.

Example: A die is rolled. We don't know whether the die is fair or loaded (i.e. shows only 1,2,3).

Example: joint probability distribution

Reminder:
Assume: the set W of possible worlds can be factorized into two sets, $W = D \times H$, i.e. the elements of W are tuples $j = (d, h)$ where $d \in D$ and $h \in H$.
Let D be the possible elementary outcomes of rolling a die, $D = \{d_1, \ldots, d_6\}$.
Let $H = \{h_1, h_2\}$ be the set of hypotheses h_1 ='the die is fair', and h_2 ='the die will show only the numbers 1,2,3'.

Example: A die is rolled. We don't know whether the die is fair or loaded (i.e. shows only 1,2,3).

Let $X : W \to \{1, 2, 3, 4, 5, 6\}$ be the random variable which assigns the number shown by the die to each possible world.
Let $Y : W \to \{\text{fair}, \text{loaded}\}$ be the random variable which assigns the identity of the die rolled to each possible world.

Example: joint probability distribution

Reminder:
Assume: the set W of possible worlds can be factorized into two sets, $W = D \times H$, i.e. the elements of W are tuples $j = (d, h)$ where $d \in D$ and $h \in H$.
Let D be the possible elementary outcomes of rolling a die, $D = \{d_1, \ldots, d_6\}$.
Let $H = \{h_1, h_2\}$ be the set of hypotheses h_1 ='the die is fair', and h_2 ='the die will show only the numbers 1,2,3'.

Example: A die is rolled. We don't know whether the die is fair or loaded (i.e. shows only 1,2,3).

Let $X : W \to \{1, 2, 3, 4, 5, 6\}$ be the random variable which assigns the number shown by the die to each possible world.
Let $Y : W \to \{\text{fair}, \text{loaded}\}$ be the random variable which assigns the identity of the die rolled to each possible world.

$P(X, Y)$: joint probability distribution over the numbers shown and the fairness of the die.
$P(X = x, Y = y) = P(x, y) =$"the probability that the die showed x and is $y \in \{\text{fair}, \text{unfair}\}$".

Use of random variables: structuring the set W

Reminder:
Assume: the set W of possible worlds can be factorized into two sets, $W = D \times H$, i.e. the elements of W are tuples $j = (d, h)$ where $d \in D$ and $h \in H$.
Let D be the possible elementary outcomes of rolling a die, $D = \{d_1, \ldots, d_6\}$.
Let $H = \{h_1, h_2\}$ be the set of hypotheses h_1 ='the die is fair', and h_2 ='the die will show only the numbers 1,2,3'.

Let $X : W \to \{1, 2, 3, 4, 5, 6\}$ be the number shown.
Let $Y : W \to \{\text{fair}, \text{loaded}\}$ be the fairness of the die.

Both X and Y act on W, but they extract different aspects of the possible worlds/elementary outcomes.
$\Rightarrow$ **Random variables** are useful for structuring and describing sets of possible worlds and/or elementary outcomes.

Marginal probability distribution

Reminder:
$P(Y)$ denotes a probability distribution over random variable Y.
$P(y)$ is a shorthand for $P(Y = y)$.
$P(X_1, \ldots, X_N)$ denotes a joint probability distribution over $X_1, \ldots, X_N$.

Definition: Let $X_1, \ldots, X_N$ be random variables with ranges $Z_1, \ldots, Z_N$, and $P(X_1, \ldots, X_N)$ be their joint probability distribution. Let $I = \{i_1, \ldots, i_K\} \subseteq \{1, \ldots, N\}$ be an index set and $J = \{1, \ldots, N\} \setminus I$ its complement.
The **marginal probability distribution** $P(X_{i_1}, \ldots, X_{i_K})$ is

$$P(x_{i_1}, \ldots, x_{i_K}) = \sum_{x_{j_1} \in Z_{j_1}} \cdots \sum_{x_{j_{N-K}} \in Z_{j_{N-K}}} P(x_1, \ldots, x_N)$$

Marginal probability distribution

Reminder:
$P(Y)$ denotes a probability distribution over random variable Y.
$P(y)$ is a shorthand for $P(Y = y)$.
$P(X_1, \ldots, X_N)$ denotes a joint probability distribution over $X_1, \ldots, X_N$.

Definition: Let $X_1, \ldots, X_N$ be random variables with ranges $Z_1, \ldots, Z_N$, and $P(X_1, \ldots, X_N)$ be their joint probability distribution. Let $I = \{i_1, \ldots, i_K\} \subseteq \{1, \ldots, N\}$ be an index set and $J = \{1, \ldots, N\} \setminus I$ its complement.
The **marginal probability distribution** $P(X_{i_1}, \ldots, X_{i_K})$ is

$$P(x_{i_1}, \ldots, x_{i_K}) = \sum_{x_{j_1} \in Z_{j_1}} \cdots \sum_{x_{j_{N-K}} \in Z_{j_{N-K}}} P(x_1, \ldots, x_N)$$

- The marginal distribution over any subset of random variables is obtained by 'summing out' all other random variables.
- If there are many other random variables/values, 'summing out' can be computationally hard.
- Since the joint distribution $P(X_1, \ldots, X_N)$ is normalized to 1, so are all marginals.

Example: marginal probability distribution

Reminder:
Assume: the set W of possible worlds can be factorized into two sets, $W = D \times H$, i.e. the elements of W are tuples $j = (d, h)$ where $d \in D$ and $h \in H$.
Let D be the possible elementary outcomes of rolling a die, $D = \{d_1, \ldots, d_6\}$.
Let $H = \{h_1, h_2\}$ be the set of hypotheses h_1 ='the die is fair', and h_2 ='the die is loaded, i.e. will show only the numbers 1,2,3'. Let $X : W \to \{1, 2, 3, 4, 5, 6\}$ be the random variable which assigns the number shown by the die to each possible world.
Let $Y : W \to \{\text{fair}, \text{loaded}\}$ be the random variable which assigns the identity of the die rolled to each possible world. $P(X, Y)$ is joint probability distribution over the numbers shown and the fairness of the die.

Example: A die is rolled. We don't know whether the die is fair or loaded (i.e. shows only 1,2,3).

Example: marginal probability distribution

Reminder:
Assume: the set W of possible worlds can be factorized into two sets, $W = D \times H$, i.e. the elements of W are tuples $j = (d, h)$ where $d \in D$ and $h \in H$.
Let D be the possible elementary outcomes of rolling a die, $D = \{d_1, \ldots, d_6\}$.
Let $H = \{h_1, h_2\}$ be the set of hypotheses h_1 ='the die is fair', and h_2 ='the die is loaded, i.e. will show only the numbers 1,2,3'. Let $X : W \to \{1, 2, 3, 4, 5, 6\}$ be the random variable which assigns the number shown by the die to each possible world.
Let $Y : W \to \{\text{fair}, \text{loaded}\}$ be the random variable which assigns the identity of the die rolled to each possible world. $P(X, Y)$ is joint probability distribution over the numbers shown and the fairness of the die.

Example: A die is rolled. We don't know whether the die is fair or loaded (i.e. shows only 1,2,3).

$P(X)$: probability distribution over the numbers shown by the die.

$P(X = x) = P(x) =$"the probability that the die showed x"
$= \sum_y P(x, y)$, where $\sum_y = \sum_{y \in \{\text{fair,loaded}\}}$

Conditional probability distribution

Reminder:
$P(Y)$ denotes a probability distribution over random variable Y.
$P(y)$ is a shorthand for $P(Y = y)$.
$P(X_1, \ldots, X_N)$ denotes a joint probability distribution over $X_1, \ldots, X_N$.

Definition: Let $X_1, \ldots, X_N$ be random variables and $P(X_1, \ldots, X_N)$ be their joint probability distribution. Let $I = \{i_1, \ldots, i_K\}$ and $C = \{c_1, \ldots, c_M\}$ be two index sets such that $I \cup C = \{1, \ldots, N\}$. If $P(X_{c_1}, \ldots, X_{c_M}) > 0$, then the **conditional probability distribution** is

$$P(X_{i_1}, \ldots, X_{i_K} | X_{c_1}, \ldots, X_{c_M}) = \frac{P(X_1, \ldots, X_N)}{P(X_{c_1}, \ldots, X_{c_M})}$$

Note: $P(X_{j_1}, \ldots, X_{j_M}) > 0$ means that this marginal distribution is strictly positive for all values of $X_{j_1}, \ldots, X_{j_M}$.

Example: conditional probability distribution

Reminder:
Assume: the set W of possible worlds can be factorized into two sets, $W = D \times H$, i.e. the elements of W are tuples $j = (d, h)$ where $d \in D$ and $h \in H$.
Let D be the possible elementary outcomes of rolling a die, $D = \{d_1, \ldots, d_6\}$.
Let $H = \{h_1, h_2\}$ be the set of hypotheses h_1 ='the die is fair', and h_2 ='the die is loaded, i.e. will show only the numbers 1,2,3'. Let $X : W \to \{1, 2, 3, 4, 5, 6\}$ be the random variable which assigns the number shown by the die to each possible world.
Let $Y : W \to \{\text{fair}, \text{loaded}\}$ be the random variable which assigns the identity of the die rolled to each possible world.

Example: A die is rolled. We don't know whether the die is fair or loaded (i.e. shows only 1,2,3).

Example: conditional probability distribution

Reminder:
Assume: the set W of possible worlds can be factorized into two sets, $W = D \times H$, i.e. the elements of W are tuples $j = (d, h)$ where $d \in D$ and $h \in H$.
Let D be the possible elementary outcomes of rolling a die, $D = \{d_1, \ldots, d_6\}$.
Let $H = \{h_1, h_2\}$ be the set of hypotheses h_1 ='the die is fair', and h_2 ='the die is loaded, i.e. will show only the numbers 1,2,3'. Let $X : W \to \{1, 2, 3, 4, 5, 6\}$ be the random variable which assigns the number shown by the die to each possible world.
Let $Y : W \to \{\text{fair}, \text{loaded}\}$ be the random variable which assigns the identity of the die rolled to each possible world.

Example: A die is rolled. We don't know whether the die is fair or loaded (i.e. shows only 1,2,3).

$P(X|Y)$: probability distribution over the numbers shown by the die given which die it is $= \frac{P(X,Y)}{P(Y)}$
$P(X = x|Y = y) = P(x|y) =$"the probability that the die showed x given that it was die y"$= \frac{P(x,y)}{P(y)}$.

Example: conditional probability distribution

Reminder:
Assume: the set W of possible worlds can be factorized into two sets, $W = D \times H$, i.e. the elements of W are tuples $j = (d, h)$ where $d \in D$ and $h \in H$.
Let D be the possible elementary outcomes of rolling a die, $D = \{d_1, \ldots, d_6\}$.
Let $H = \{h_1, h_2\}$ be the set of hypotheses h_1 ='the die is fair', and h_2 ='the die is loaded, i.e. will show only the numbers 1,2,3'. Let $X : W \to \{1, 2, 3, 4, 5, 6\}$ be the random variable which assigns the number shown by the die to each possible world.
Let $Y : W \to \{\text{fair}, \text{loaded}\}$ be the random variable which assigns the identity of the die rolled to each possible world.

Example: A die is rolled. We don't know whether the die is fair or loaded (i.e. shows only 1,2,3).

$P(X|Y)$: probability distribution over the numbers shown by the die given which die it is $= \frac{P(X,Y)}{P(Y)}$
$P(X = x|Y = y) = P(x|y) =$"the probability that the die showed x given that it was die y"$= \frac{P(x,y)}{P(y)}$.

Note: writing $P(X|Y) = \frac{P(X,Y)}{P(Y)}$ means that this relationship holds point-wise, i.e. for all possible values of X and Y.

Product rule for probability distributions

Reminder:
Random variables $X_1, \ldots, X_N$ with joint prob. dist. $P(X_1, \ldots, X_N)$.
$I = \{i_1, \ldots, i_K\}$ and $C = \{c_1, \ldots, c_M\}$ such that $I \cup C = \{1, \ldots, N\}$.
Conditional prob. dist. :$P(X_{i_1}, \ldots, X_{i_K} | X_{c_1}, \ldots, X_{c_M}) = \frac{P(X_1, \ldots, X_N)}{P(X_{c_1}, \ldots, X_{c_M})}$

A consequence of the definition of the conditional probability distribution is the **product rule for random variables**:

$$P(X_{i_1}, \ldots, X_{i_K} | X_{c_1}, \ldots, X_{c_M}) P(X_{c_1}, \ldots, X_{c_M}) = P(X_1, \ldots, X_N)$$

Note: as before, the equality is point-wise.

Chain rule for probability distributions

Reminder:
Random variables $X_1, \ldots, X_N$ with joint prob. dist. $P(X_1, \ldots, X_N)$.
$I = \{i_1, \ldots, i_K\}$ and $C = \{c_1, \ldots, c_M\}$ such that $I \cup C = \{1, \ldots, N\}$.
Product rule for prob. dist. $P(X_{i_1}, \ldots, X_{i_K} | X_{c_1}, \ldots, X_{c_M}) P(X_{c_1}, \ldots, X_{c_M}) = P(X_1, \ldots, X_N)$

Apply product rule repeatedly:

$$\begin{aligned} P(X_1, \ldots, X_N) &= P(X_1 | X_2, \ldots, X_N) P(X_2, \ldots, X_N) \\ &= P(X_1 | X_2, \ldots, X_N) P(X_2 | X_3, \ldots, X_N) P(X_3, \ldots, X_N) \\ &\vdots \\ &= \prod_{i=1}^{N-1} P(X_i | X_{i+1}, \ldots, X_N) P(X_N) \end{aligned}$$

Holds for any ordering of the X_i !
This is the **chain rule for probability distributions**.

Independence between random variables

Reminder:
$P(X, Y)$ is joint probability distribution of X and Y.
$P(X) = \sum_y P(X, y)$ is the marginal probability distribution of X.
$P(Y) = \sum_x P(x, Y)$ is the marginal probability distribution of Y.

Definition: Two random variables X and Y are independent if and only if

$$P(X, Y) = P(X)P(Y).$$

Note: If X, Y are independent, then $P(X|Y) = \frac{P(X,Y)}{P(Y)} = P(X)$.
Knowing Y does not change knowledge of X.

Motivating example: conditional independence between random variables

Example: A die is rolled *twice*. We don't know whether the die is fair or loaded (i.e. shows only 1,2,3).

Motivating example: conditional independence between random variables

Example: A die is rolled *twice*. We don't know whether the die is fair or loaded (i.e. shows only 1,2,3).

Random variables:

- X_1: value of outcome of 1st roll $\in \{1; \ldots; 6\}$.
- X_2: value of outcome of 2nd roll $\in \{1; \ldots; 6\}$.
- Y: fairness of the die $\in \{\text{fair}, \text{loaded}\}$.

Motivating example: conditional independence between random variables

Example: A die is rolled *twice*. We don't know whether the die is fair or loaded (i.e. shows only 1,2,3).

Random variables:

- X_1: value of outcome of 1st roll $\in \{1; \ldots; 6\}$.
- X_2: value of outcome of 2nd roll $\in \{1; \ldots; 6\}$.
- Y: fairness of the die $\in \{\text{fair}, \text{loaded}\}$.

Question: what is the joint distribution $P(X_1, X_2, Y)$? Knowing it would enable us to compute all marginals and conditionals, e.g. $P(Y|X_1, X_2)$.

Motivating example: conditional independence between random variables

Reminder:
A die is rolled *twice*. We don't know whether the die is fair or loaded (i.e. shows only 1,2,3).
Random variables: X_1, X_2: values of outcomes of 1st and 2nd roll $\in \{1; \ldots; 6\}$.
Y: fairness of the die $\in \{\text{fair}, \text{loaded}\}$.
$P(y) = \frac{1}{2}$, $P(x|Y = \text{fair}) = \frac{1}{6}$
$P(X = 1|Y = \text{loaded}) = P(X = 2|Y = \text{loaded}) = P(X = 3|Y = \text{loaded}) = \frac{1}{3}$
$P(X = 4|Y = \text{loaded}) = P(X = 5|Y = \text{loaded}) = P(X = 6|Y = \text{loaded}) = 0$

Question: what is the joint distribution $P(X_1, X_2, Y)$?
Answer: use chain rule:

$$P(X_1, X_2, Y) = P(X_1|X_2, Y)P(X_2|Y)P(Y)$$

We know $P(Y)$ and $P(X_2|Y)$. What about $P(X_1|X_2, Y)$?

Motivating example: conditional independence between random variables

Reminder:
A die is rolled *twice*. We don't know whether the die is fair or loaded (i.e. shows only 1,2,3).
Random variables: X_1, X_2: values of outcomes of 1st and 2nd roll $\in \{1; \ldots; 6\}$.
Y: fairness of the die $\in \{\text{fair}, \text{loaded}\}$.
$P(y) = \frac{1}{2}$, $P(x|Y = \text{fair}) = \frac{1}{6}$
$P(X = 1|Y = \text{loaded}) = P(X = 2|Y = \text{loaded}) = P(X = 3|Y = \text{loaded}) = \frac{1}{3}$
$P(X = 4|Y = \text{loaded}) = P(X = 5|Y = \text{loaded}) = P(X = 6|Y = \text{loaded}) = 0$

Question: what is the joint distribution $P(X_1, X_2, Y)$?
Answer: use chain rule:

$$P(X_1, X_2, Y) = P(X_1|X_2, Y)P(X_2|Y)P(Y)$$

We know $P(Y)$ and $P(X_2|Y)$. What about $P(X_1|X_2, Y)$?
Once we know the die (i.e. the value of Y), the values of $P(X_i|Y)$ of each die roll should be the same, no matter how often we roll the die.

Motivating example: conditional independence between random variables

Reminder:
A die is rolled *twice*. We don't know whether the die is fair or loaded (i.e. shows only 1,2,3).
Random variables: X_1, X_2: values of outcomes of 1st and 2nd roll $\in \{1; \ldots; 6\}$.
Y: fairness of the die $\in \{\text{fair}, \text{loaded}\}$.
$P(y) = \frac{1}{2}$, $P(x|Y = \text{fair}) = \frac{1}{6}$
$P(X = 1|Y = \text{loaded}) = P(X = 2|Y = \text{loaded}) = P(X = 3|Y = \text{loaded}) = \frac{1}{3}$
$P(X = 4|Y = \text{loaded}) = P(X = 5|Y = \text{loaded}) = P(X = 6|Y = \text{loaded}) = 0$

Question: what is the joint distribution $P(X_1, X_2, Y)$?
Answer: use chain rule:

$$P(X_1, X_2, Y) = P(X_1|X_2, Y)P(X_2|Y)P(Y)$$

We know $P(Y)$ and $P(X_2|Y)$. What about $P(X_1|X_2, Y)$?
Once we know the die (i.e. the value of Y), the values of $P(X_i|Y)$ of each die roll should be the same, no matter how often we roll the die.

$$\Rightarrow P(X_1|X_2, Y) = P(X_1|Y).$$

N.B.: a.k.a 'model assumption'

Motivating example: conditional independence between random variables

Reminder:
A die is rolled *twice*. We don't know whether the die is fair or loaded (i.e. shows only 1,2,3).
Random variables: X_1, X_2: values of outcomes of 1st and 2nd roll $\in \{1; \ldots; 6\}$.
Y: fairness of the die $\in \{\text{fair}, \text{loaded}\}$.
$P(y) = \frac{1}{2}$, $P(x|Y = \text{fair}) = \frac{1}{6}$
$P(X = 1|Y = \text{loaded}) = P(X = 2|Y = \text{loaded}) = P(X = 3|Y = \text{loaded}) = \frac{1}{3}$
$P(X = 4|Y = \text{loaded}) = P(X = 5|Y = \text{loaded}) = P(X = 6|Y = \text{loaded}) = 0$

We believe: $P(X_1|X_2, Y) = P(X_1|Y)$. Thus:

$$\begin{aligned} P(X_1, X_2, Y) &= P(X_1|Y)P(X_2|Y)P(Y) \\ &= P(X_1, X_2|Y)P(Y) \end{aligned}$$

Motivating example: conditional independence between random variables

Reminder:
A die is rolled *twice*. We don't know whether the die is fair or loaded (i.e. shows only 1,2,3).
Random variables: X_1, X_2: values of outcomes of 1st and 2nd roll $\in \{1; \ldots; 6\}$.
Y: fairness of the die $\in \{\text{fair}, \text{loaded}\}$.
$P(y) = \frac{1}{2}$, $P(x|Y = \text{fair}) = \frac{1}{6}$
$P(X = 1|Y = \text{loaded}) = P(X = 2|Y = \text{loaded}) = P(X = 3|Y = \text{loaded}) = \frac{1}{3}$
$P(X = 4|Y = \text{loaded}) = P(X = 5|Y = \text{loaded}) = P(X = 6|Y = \text{loaded}) = 0$

We believe: $P(X_1|X_2, Y) = P(X_1|Y)$. Thus:

$$\begin{aligned} P(X_1, X_2, Y) &= P(X_1|Y)P(X_2|Y)P(Y) \\ &= P(X_1, X_2|Y)P(Y) \end{aligned}$$

$$\Rightarrow P(X_1, X_2|Y) = P(X_1|Y)P(X_2|Y)$$

Like the definition of independence, but everything is conditioned on Y.

Conditional independence between random variables

Definition: Two random variables X_1 and X_2 are *conditionally independent* given a random variable Y if and only if

$$P(X_1, X_2|Y) = P(X_1|Y)P(X_2|Y).$$

Alternatively, X_1 and X_2 are conditionally independent if and only if

- $P(X_1|Y) > 0$
- $P(X_2|Y) > 0$
- $P(X_1|X_2, Y) = P(X_1|Y)$
- $P(X_2|X_1, Y) = P(X_2|Y)$

Conditional independence between random variables

Definition: Two random variables X_1 and X_2 are *conditionally independent* given a random variable Y if and only if

$$P(X_1, X_2|Y) = P(X_1|Y)P(X_2|Y).$$

Conditional independence between random variables

Definition: Two random variables X_1 and X_2 are *conditionally independent* given a random variable Y if and only if

$$P(X_1, X_2|Y) = P(X_1|Y)P(X_2|Y).$$

Notes:

- This definition can be extended to more than 3 random variables by replacing any of X_1, X_2 or Y with a list of random variables.
- Variables that are conditionally independent are usually marginally dependent, and vice versa.

Example: conditional independence vs. marginal dependence

Reminder:
A die is rolled *twice*. We don't know whether the die is fair or loaded (i.e. shows only 1,2,3).
Random variables: X_1, X_2: values of outcomes of 1st and 2nd roll $\in \{1; \ldots; 6\}$.
Y: fairness of the die $\in \{\text{fair}, \text{loaded}\}$.
Conditional independence: $P(X_1|X_2, Y) = P(X_1|Y)$ and $P(X_2|X_1, Y) = P(X_2|Y)$.

Die rolls are conditionally independent given Y.
Marginal probability distribution $P(X_1, X_2)$:

$$\begin{aligned} P(X_1, X_2) &= \sum_y P(X_1, X_2, y) \\ &= \sum_y P(X_1|X_2, y)P(X_2|y)P(y) \\ &= \sum_y P(X_1|y)P(X_2|y)P(y) \end{aligned}$$

Example: conditional independence vs. marginal dependence

Reminder:
A die is rolled *twice*. We don't know whether the die is fair or loaded (i.e. shows only 1,2,3).
Random variables: X_1, X_2: values of outcomes of 1st and 2nd roll $\in \{1; \ldots; 6\}$.
Y: fairness of the die $\in \{\text{fair}, \text{loaded}\}$.
Conditional independence: $P(X_1|X_2, Y) = P(X_1|Y)$ and $P(X_2|X_1, Y) = P(X_2|Y)$.
Marginal prob. dist. $P(X_1, X_2) = \sum_y P(X_1|y)P(X_2|y)P(y)$

On the other hand:

$$
\begin{aligned}
P(X_1) &= \sum_y P(X_1, y) \\
&= \sum_y P(X_1|y)P(y) \\
P(X_2) &= \sum_y P(X_2|y)P(y)
\end{aligned}
$$

Example: conditional independence vs. marginal dependence

Reminder:
A die is rolled *twice*. We don't know whether the die is fair or loaded (i.e. shows only 1,2,3).
Random variables: X_1, X_2: values of outcomes of 1st and 2nd roll $\in \{1; \ldots; 6\}$.
Y: fairness of the die $\in \{\text{fair}, \text{loaded}\}$.
Conditional independence: $P(X_1|X_2, Y) = P(X_1|Y)$ and $P(X_2|X_1, Y) = P(X_2|Y)$.
Marginal prob. dist. $P(X_1, X_2) = \sum_y P(X_1|y)P(X_2|y)P(y)$

Therefore

$$\begin{aligned} P(X_1)P(X_2) &= \sum_y P(X_1|y)P(y) \sum_y P(X_2|y)P(y) \\ &\neq \sum_y P(X_1|y)P(X_2|y)P(y) \\ &= P(X_1, X_2) \end{aligned}$$

$\Rightarrow X_1$ and X_2 are marginally dependent.
$\Rightarrow$ One die roll contains information about the other if we *do not* know Y.
$\Rightarrow$ The marginal dependence goes in *both* directions.

Summary: random variables

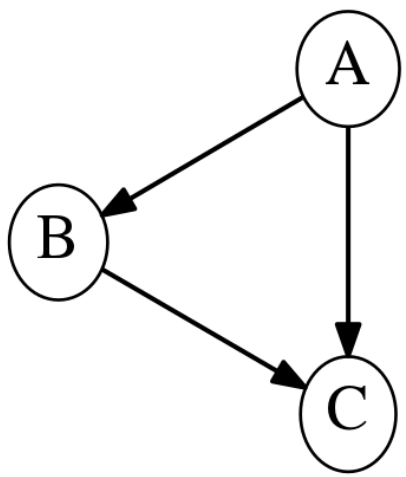
- A *random variable* X on a set of possible worlds W is a function $X : W \to Z$ from W to some range Z.
- A *probability distribution* is a function $P : Z \to [0, 1]$ such that $\sum_{x \in Z} P(X = x) = 1$.
- Chain rule: $P(X_1, \ldots, X_N) = \prod_{i=1}^{N} P(X_i | X_{i+1}, \ldots, X_N)$
- Conditional independence $P(X_1, X_2 | Y) = P(X_1 | Y) P(X_2 | Y)$.
- Conditional independence $P(X_1 | X_2, Y) = P(X_1 | Y)$.
  - Expressed by omitting all variables that X_1 does not depend on after the conditioning line (here: X_2 omitted).
- Marginal probability distribution $P(X_1) = \sum_y P(X_1, y)$. Can be hard to compute!

Bayesian networks

A type of probabilistic graphical model which expresses conditional (in)dependence relationships.

Random variables	**Bayesian networks**
Random variables A, B, C	Nodes of a graph
Conditional (in)dependence	Directed edges
Chain rule decomposition	directed acyclic graph (DAG)

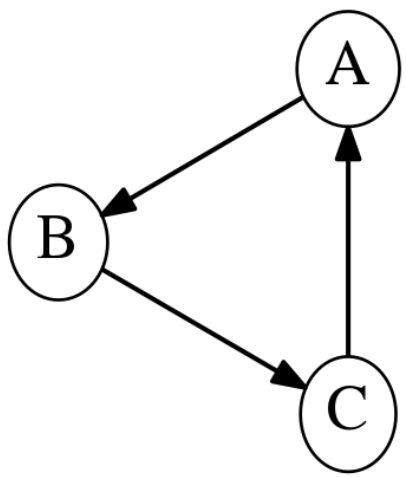
The graph represents a set of *constraints* on the joint probability distribution of the random variables.



A Bayesian network with 3 random variables A,B,C.

Example: closed loop

NOT a directed acyclic graph (DAG), thus not a Bayesian network.
Closed directed loop $A \to B \to C$.

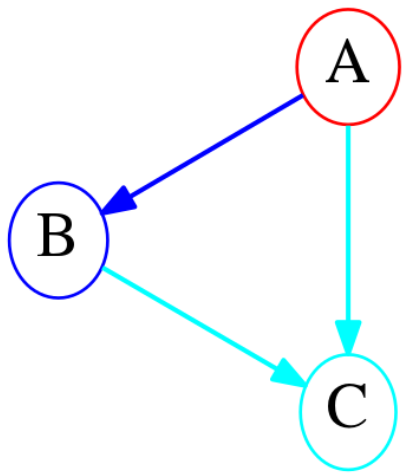


Representing constraints on joint distributions

Reminder:
Chain rule for prob. dist. $P(X_1, \ldots, X_N) = \prod_{i=1}^{N} P(X_i | X_{i+1}, \ldots, X_N)$

Example: 3 random variables A, B, C. Joint distribution

$$P(A, B, C) = P(A)P(B|A)P(C|A, B)$$

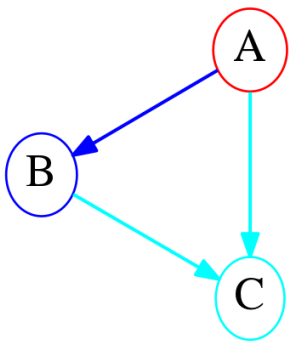


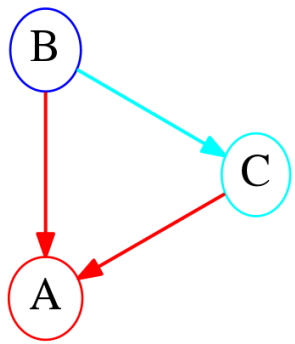
- Each node represents a random variable
- If and only if there is an edge from A to B, then A appears in the conditional distribution of B given B's predecessors in the factorization chain: $P(B|A, \ldots)$.

Alternative ordering of variables

Reminder:
Chain rule for prob. dist. $P(X_1, \ldots, X_N) = \prod_{i=1}^{N} P(X_i | X_{i+1}, \ldots, X_N)$
Independece of random variables $P(X, Y) = P(X)P(Y)$.

Order of factorization of joint distribution can be exchanged:





$P(A, B, C) = P(A)P(B|A)P(C|A, B)$ $P(A, B, C) = P(B)P(C|B)P(A|B, C)$

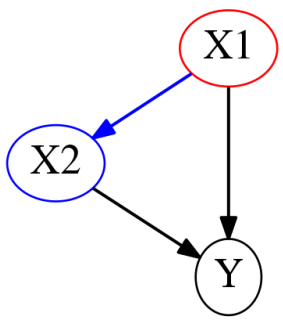
- Both graphs describe possible factorizations of $P(A, B, C)$.
- Here, both factorizations are equivalent w.r.t. the dependency structure: a given variable is conditionally dependent on all others.
  - A consequence of probabilistic (in)dependence being a mutual property.
  - Both graphs are *fully connected*.

Example: rolling a die twice. Bad ordering

Reminder:
Each node represents a random variable.
If there is an edge from A to B, then A appears in the conditional distribution of B given B's predecessors in the factorization chain: $P(B|A, \ldots)$.

Random variables: X_1, X_2: value of 1st and 2nd roll, Y: fairness.
Factorization of joint probability distribution:

$$P(X_1, X_2, Y) = P(X_1)P(X_2|X_1)P(Y|X_1, X_2)$$



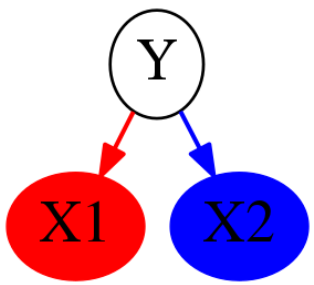
$\Rightarrow$ the factorization order $X_1 \to X_2 \to Y$ is not a good choice, because all variables are dependent on each other.

Example: rolling a die twice. Good ordering

Reminder:
Each node represents a random variable.
If there is an edge from A to B, then A appears in the conditional distribution of B given B's predecessors in the factorization chain: $P(B|A, \ldots)$.

Random variables: X_1, X_2: value of 1st and 2nd roll, Y: fairness.
Factorization of joint probability distribution:

$$P(X_1, X_2, Y) = P(Y)P(X_1|Y)\underbrace{P(X_2|X_1, Y)}_{P(X_2|Y)}$$



$\Rightarrow$ the factorization order $Y \to X_1 \to X_2$ is a better choice of ordering, because conditional independence relationships are represented in the graph!

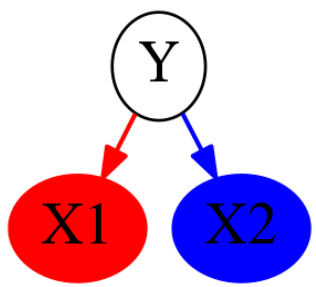
Example: rolling a die twice. Good ordering

Reminder:
Each node represents a random variable.
If there is an edge from A to B, then A appears in the conditional distribution of B given B's predecessors in the factorization chain: $P(B|A, \ldots)$.

Random variables: X_1, X_2: value of 1st and 2nd roll, Y: fairness.
Factorization of joint probability distribution:

$$P(X_1, X_2, Y) = P(Y)P(X_1|Y)\underbrace{P(X_2|X_1, Y)}_{P(X_2|Y)}$$

Filled nodes:
observed variables



$\Rightarrow$ the factorization order $Y \to X_1 \to X_2$ is a better choice of ordering, because conditional independence relationships are represented in the graph!

Good vs. bad random variable ordering

Question: in what sense is the factorization

$$P(X_1, X_2, Y) = P(Y)P(X_1|Y)P(X_2|Y)$$

better than

$$P(X_1, X_2, Y) = P(X_1)P(X_2|X_1)P(Y|X_1, X_2)$$

Good vs. bad random variable ordering

Question: in what sense is the factorization

$$P(X_1, X_2, Y) = P(Y)P(X_1|Y)P(X_2|Y)$$

better than

$$P(X_1, X_2, Y) = P(X_1)P(X_2|X_1)P(Y|X_1, X_2)$$

Answer 1: consider the number of probabilities which you have to assign: if a random variable can take on N different values, then you have to guess/estimate $N - 1$ probabilities to determine its probability distribution.

- Good ordering: $1 + (5 \times 2) + (5 \times 2) = 21$. Because of i.i.d. property, actually only 11.
- Bad ordering: $5 + (5 \times 6) + 1 \times (6 \times 6) = 71$.

$\Rightarrow$ far less probabilities for the good ordering.

Good vs. bad random variable ordering

Question: in what sense is the factorization

$$P(X_1, X_2, Y) = P(Y)P(X_1|Y)P(X_2|Y)$$

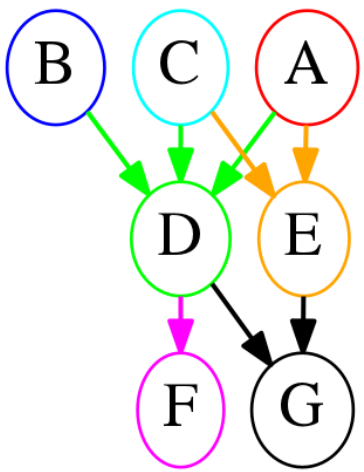
better than

$$P(X_1, X_2, Y) = P(X_1)P(X_2|X_1)P(Y|X_1, X_2)$$

Answer 2: The good ordering represents our information about the structure of the problem: die fairness determines probabilities of outcomes, not the other way round. We might say that the good ordering represents the 'causal structure' of the problem. *Caveat*: for a Bayesian network to represent causal structure, additional conditions must hold (see e.g. Pearl(2000):Causality).

Bayesian network terminology

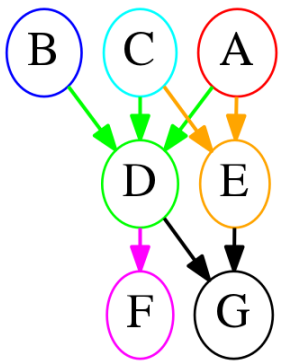
Reminder:
Each node represents a random variable.
If there is an edge from A to B, then A appears in the conditional distribution of B given B's predecessors in the factorization chain: $P(B|A, \ldots)$.



A,B,C are the *parents* of D. $\text{pa}_D = \{A, B, C\}$.
D,E are the *children* of A.
A,B,C,D,E are the *ancestors* of G.
D,F,G are the *descendants* of B.
A,B,C are the *roots* (no parents).
F,G are the *leaves* (no children).

Conditional independence given parents

Reminder:
Each node represents a random variable.
If and only if there is an edge from A to B, then A appears in the conditional distribution of B given B's predecessors in the factorization chain: $P(B|A, \ldots)$.
Set of parents of node A is pa_A.



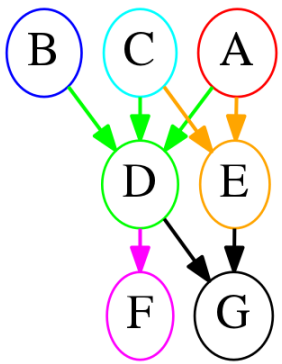
$$\text{pa}_A = \text{pa}_B = \text{pa}_C = \emptyset$$
$$\text{pa}_D = \{A, B, C\}$$
$$\text{pa}_E = \{A, C\}$$
$$\text{pa}_F = \{D\}$$
$$\text{pa}_G = \{D, E\}$$

Factorization of joint distribution: choose an ordering such that pa_X always precede X in the factorization chain. Always possible because graph is a DAG. Let $P(X|\emptyset) = P(X)$.

$$\begin{aligned} P(A, B, C, D, E, F, G) &= P(A)\, P(B)\, P(C) \\ &\times P(D|A, B, C)\, P(E|A, C) \\ &\times P(G|D, E)\, P(F|D) \end{aligned}$$

Conditional independence given parents

Reminder:
Each node represents a random variable.
If and only if there is an edge from A to B, then A appears in the conditional distribution of B given B's predecessors in the factorization chain: $P(B|A, \ldots)$.
Set of parents of node A is pa_A.



$$\text{pa}_A = \text{pa}_B = \text{pa}_C = \emptyset$$
$$\text{pa}_D = \{A, B, C\}$$
$$\text{pa}_E = \{A, C\}$$
$$\text{pa}_F = \{D\}$$
$$\text{pa}_G = \{D, E\}$$

Alternatively, we can write this as

$$P(A, B, C, D, E, F, G) =$$

$$\begin{aligned} &= P(A)\,P(B)\,P(C) && = P(A|\text{pa}_A)\,P(B|\text{pa}_B)\,P(C|\text{pa}_C) \\ &\times P(D|A, B, C)\,P(E|A, C) && \times P(D|\text{pa}_D)\,P(E|\text{pa}_E) \\ &\times P(F|D)\,P(G|D, E) && \times P(F|\text{pa}_F)\,P(G|\text{pa}_G) \end{aligned}$$

Translating a graph structure into a factorization

The expression for the joint distribution

$$\begin{aligned} P(A,B,C,D,E,F,G) &= P(A|\text{pa}_A)\, P(B|\text{pa}_B)\, P(C|\text{pa}_C) \\ &\times P(D|\text{pa}_D)\, P(E|\text{pa}_E) \\ &\times P(F|\text{pa}_F)\, P(G|\text{pa}_G) \end{aligned}$$

no longer depends on the chosen factorization order, only on the parent-child relationships expressed in the graph!
(because multiplication is commutative).

Translating a graph structure into a factorization

The expression for the joint distribution

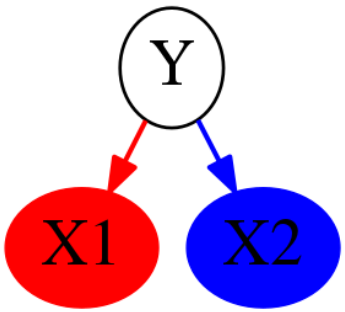
$$\begin{aligned} P(A,B,C,D,E,F,G) &= P(A|\text{pa}_A)\, P(B|\text{pa}_B)\, P(C|\text{pa}_C) \\ &\times P(D|\text{pa}_D)\, P(E|\text{pa}_E) \\ &\times P(F|\text{pa}_F)\, P(G|\text{pa}_G) \end{aligned}$$

no longer depends on the chosen factorization order, only on the parent-child relationships expressed in the graph!
(because multiplication is commutative).
Algorithm for translating a Bayesian network into a factorization of a joint distribution:

- Given: random variables $X_1, \ldots, X_N$ and a DAG G with nodes labeled $X_1, \ldots, X_N$.
- For all X_i, identify pa_{X_i} from G.
- Output $P(X_1, \ldots, X_N) = \prod_{i=1}^{N} P(X_i|\text{pa}_{X_i})$

Example: from graph to factorization

Random variables: X_1, X_2: value of 1st and 2nd roll, Y: fairness.



- $\text{pa}_Y = \emptyset$
- $\text{pa}_{X_1} = \{Y\}$
- $\text{pa}_{X_2} = \{Y\}$

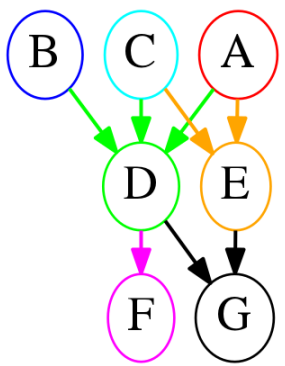
$\Rightarrow P(X_1, X_2, Y) = P(Y)P(X_1|Y)P(X_2|Y)$

Given a factorization:

$$\begin{aligned} P(A,B,C,D,E,F,G) &= P(A)\,P(B)\,P(C) \\ &\times P(D|A,B,C)\,P(E|A,C) \\ &\times P(G|D,E)\,P(F|D) \end{aligned}$$

building the graph is straightforward:

1. Identify and draw the roots: A, B, C
2. Find all children of the roots: D, E
3. Draw arrows for each cond. dependence
4. Iterate 2. and 3. until leaves are reached



Summary: Bayesian networks

A type of probabilistic graphical model which expresses conditional (in)dependence relationships.

Random variables	**Bayesian networks**
Random variables A, B, C	Nodes of a graph
Conditional (in)dependence	Directed edges
Chain rule decomposition	directed acyclic graph (DAG)

- Good decompositions keep the number of probabilities to estimate small.
- Good decompositions represent our knowledge/assumptions about probabilistic (in)dependence relationships between the random variables involved.
- A given chain-rule factorization can translated into a DAG.
- A given DAG can be translated into a chain-rule factorization.

Factor graphs

Let $\mathbf{X} = \{X_1, \ldots, X_N\}$ be the arguments of a function $q(X_1, \ldots, X_N)$. Let the sets $\mathbf{s}_1, \ldots, \mathbf{s}_K$ be subsets of $\mathbf{X}$, and the functions $f_1, \ldots, f_K$ depend only on the arguments in the corresponding $\mathbf{s}_i$. The f_i are a factorization of q if

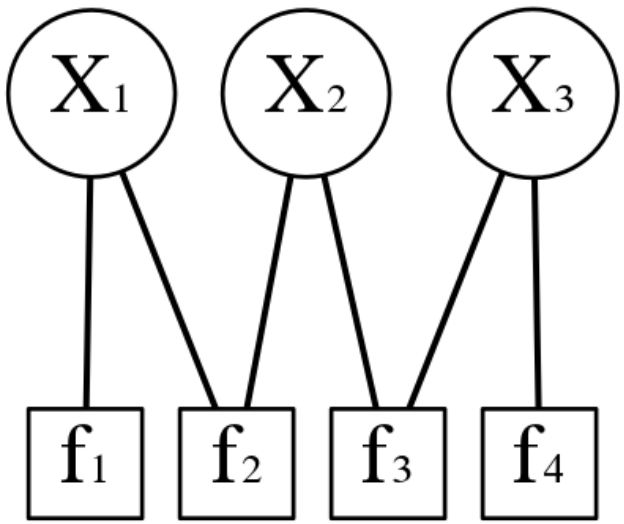
$$q(\mathbf{X}) = \prod_{i=1}^{K} f_i(\mathbf{s}_i)$$

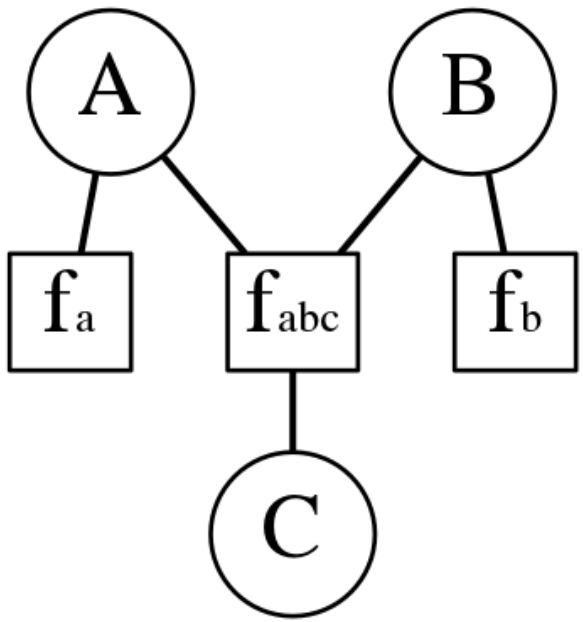
To construct a factor graph from a factorization,

- draw one variable node (circle) for each variable X_j,
- draw one factor node (square) for each factor f_i,
- if $X_j \in \mathbf{s}_i$, connect the variable node of X_j with the factor node of f_i.

Example: factor graphs

$f_1(X_1)\, f_2(X_1, X_2)\, f_3(X_2, X_3)\, f_4(X_3)$ $\quad$ $f_a(A)\, f_b(B)\, f_{abc}(A, B, C)$





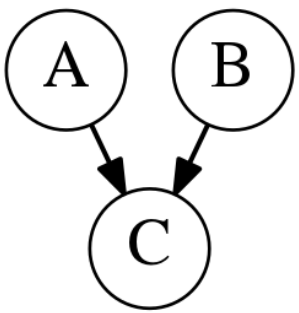
Factor graphs are *bipartite*: factors are only connected to variables, and vice versa!

Translating Bayesian Nets into factor graphs

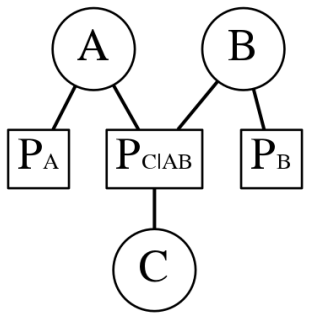
Bayesian network	factor graph
Node	Variable node
(Un)conditional distribution	Factor node
Origin node of directed edge	Edge from variable to factor
Target node of directed edge	Edge from factor to variable

Example: $P(A, B, C) = P(A)P(B)P(C|A, B)$

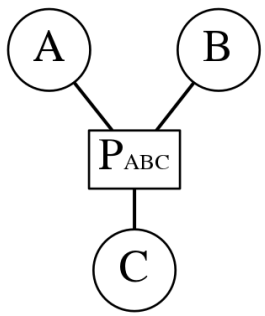
Bay. net



A factor graph

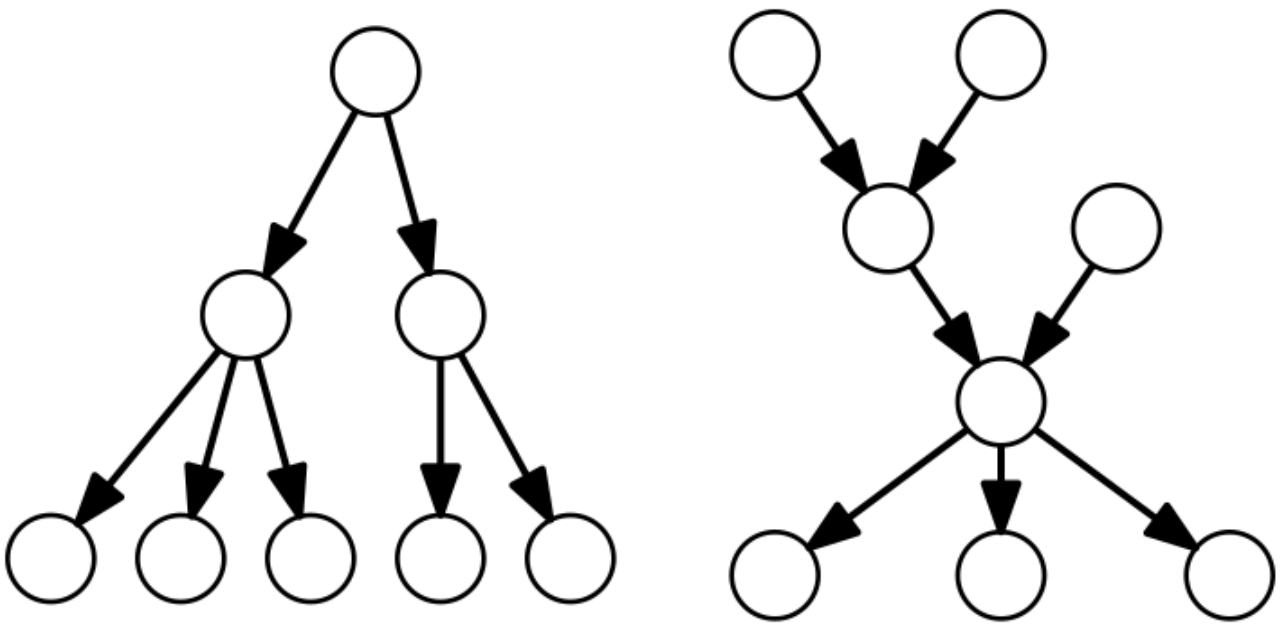


Another factor graph



Trees and polytrees

Singly connected Bayesian networks can be marginalized efficiently after conversion to a factor graph.
A graph is singly connected, if there is exactly one path between any pair of nodes (ignoring the arrows). Trees and polytrees.



Summary: sum-product algorithm

- Message from factor t to variable Y:

$$\mu_{t\to Y}(Y) = \sum_{Z_1} \dots \sum_{Z_M} t(Y, Z_1, \dots, Z_M) \prod_{Z \in \mathrm{ne}_t \setminus Y} \mu_{Z\to t}(Z)$$

- Message from variable Y to factor t:

$$\mu_{Y\to t}(Y) = \prod_{I \in \mathrm{ne}_Y \setminus t} \mu_{I\to Y}(Y)$$

- Message from a factor leaf node f to variable Y: $\mu_{f\to Y}(Y) = f(Y)$
- Message from a variable leaf node Z to factor t: $\mu_{Z\to t}(Z) = 1$
- Marginal of Y:

$$f(Y) = \prod_{t \in \mathrm{ne}_Y} \mu_{t\to Y}(Y)$$

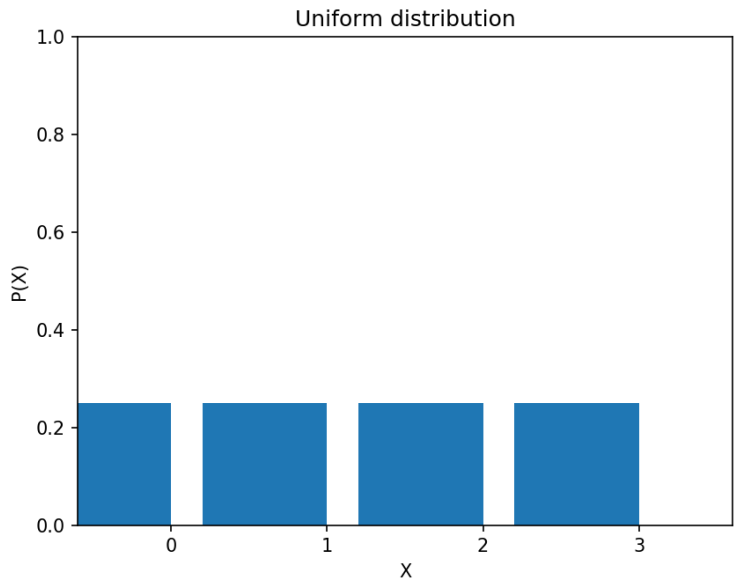
.

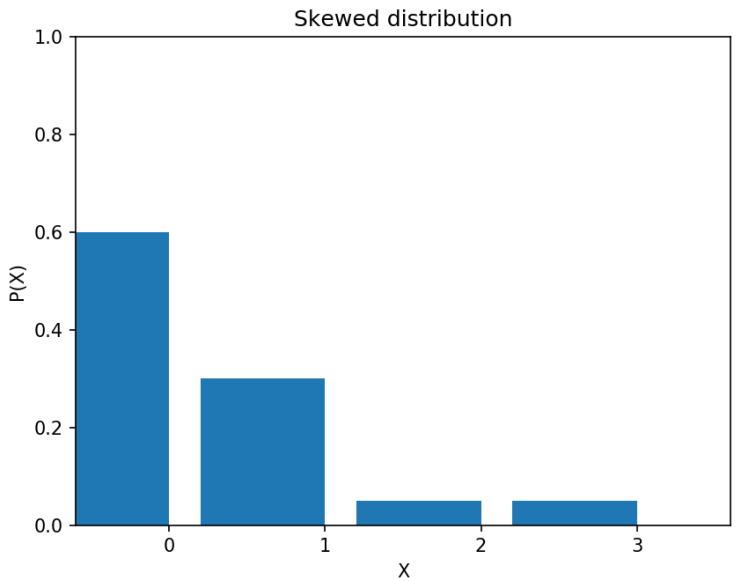
Entropy

Questions:

- How much do we expect to learn by taking a measurement?
- How can we quantify this 'amount' of learning?
- Information theory [Shannon 1948] formalizes answers in the context of communication

Simpler question: Which $P(X)$ would lead you to expect more information about x from a measurement?



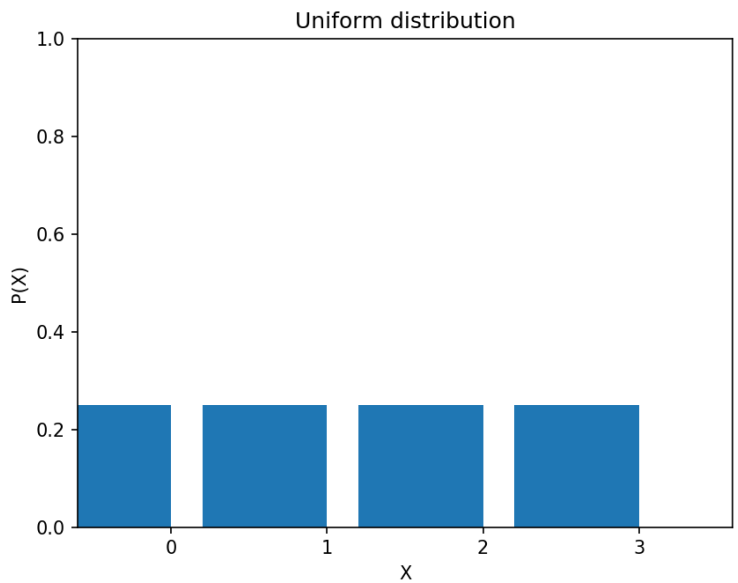


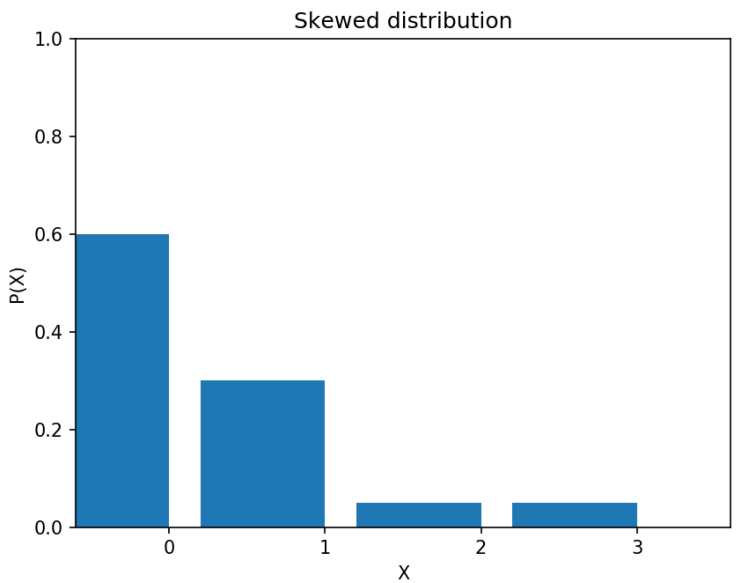
Entropy

Questions:

- How much do we expect to learn by taking a measurement?
- How can we quantify this 'amount' of learning?
- Information theory [Shannon 1948] formalizes answers in the context of communication

Simpler question: Which $P(X)$ would lead you to expect more information about x from a measurement?

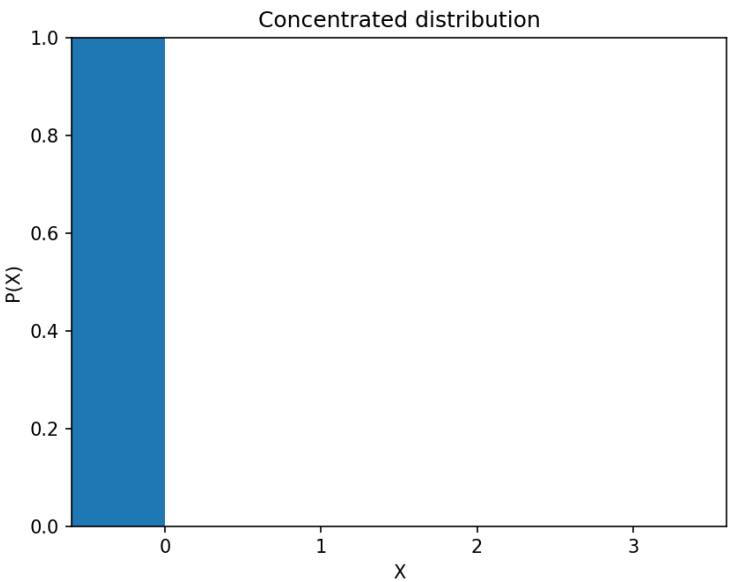


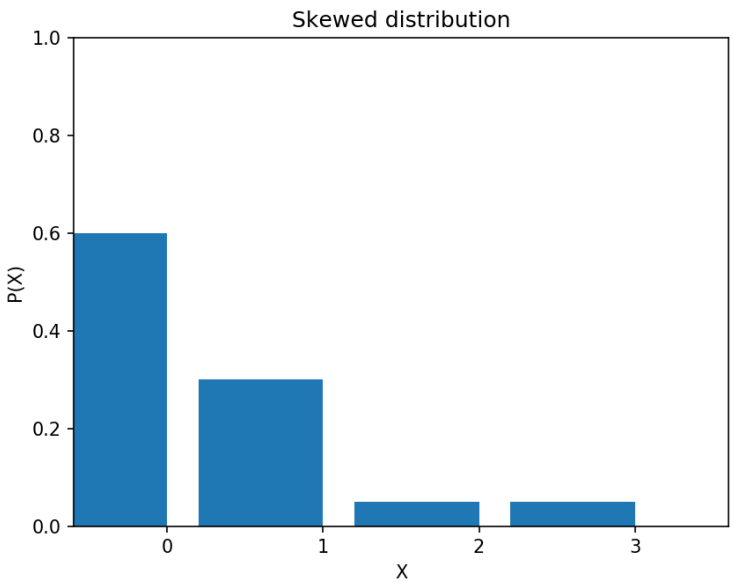


Left $P(X)$, because we are less certain before measurement.

Entropy

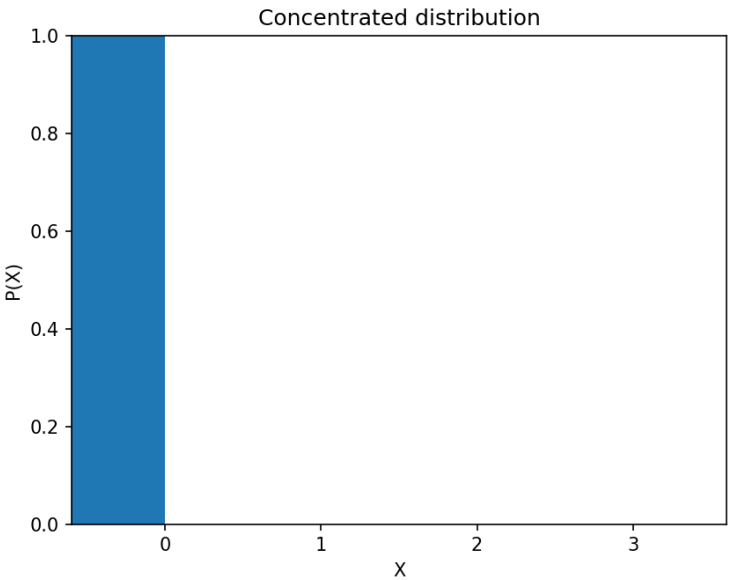
Simpler question: Which $P(X)$ would lead you to expect more information about x from a measurement?

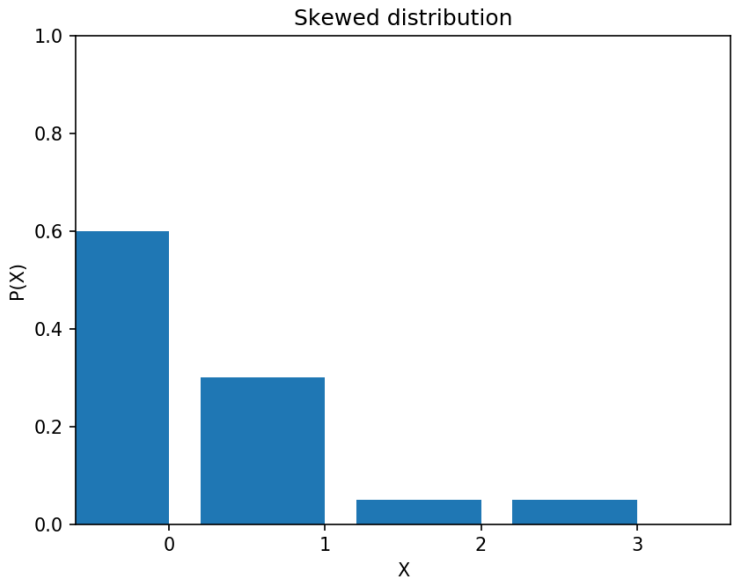




Entropy

Simpler question: Which $P(X)$ would lead you to expect more information about x from a measurement?

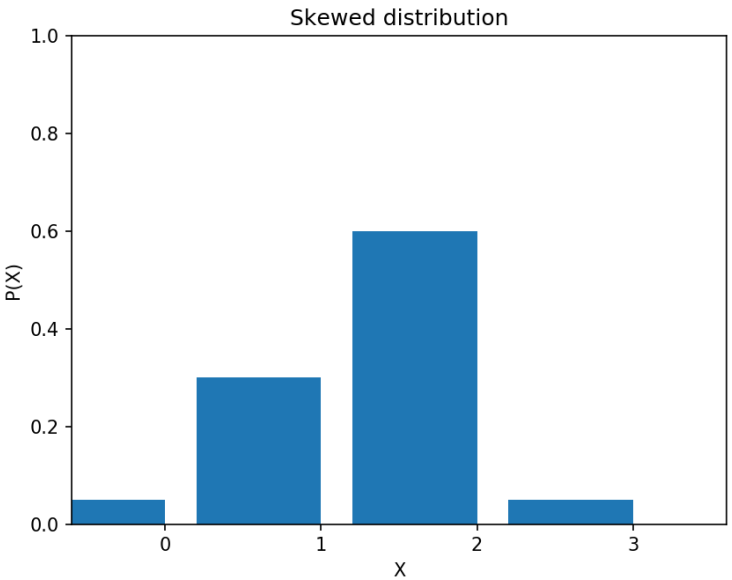


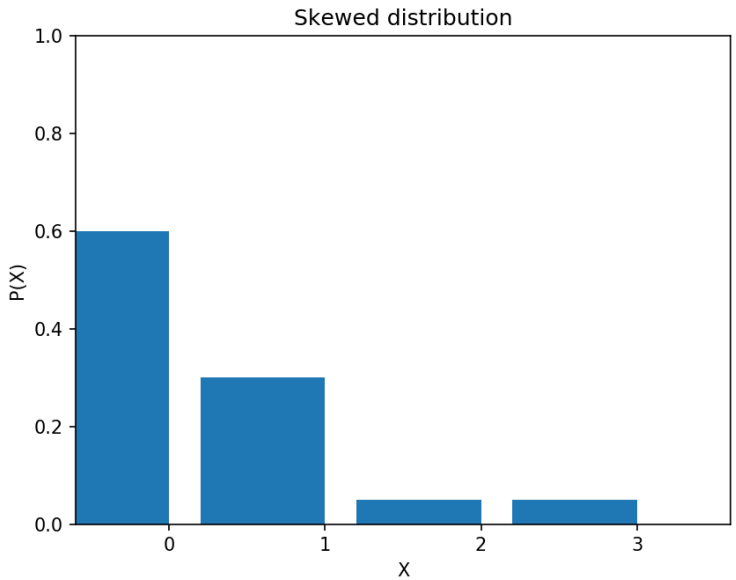


Right $P(X)$, because we are less certain before measurement.
$\Rightarrow$ the more concentrated or 'peaky' the distribution, the less we expect to learn from measuring X.

Entropy

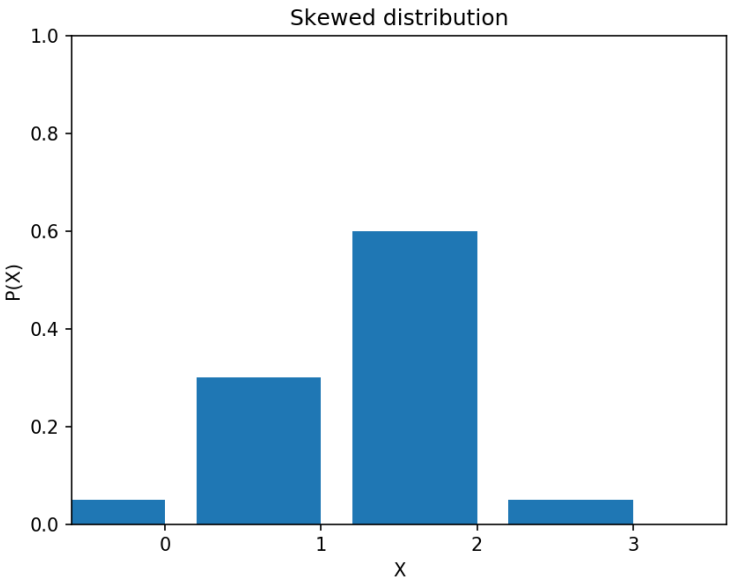
Simpler question: Which $P(X)$ would lead you to expect more information about x from a measurement?

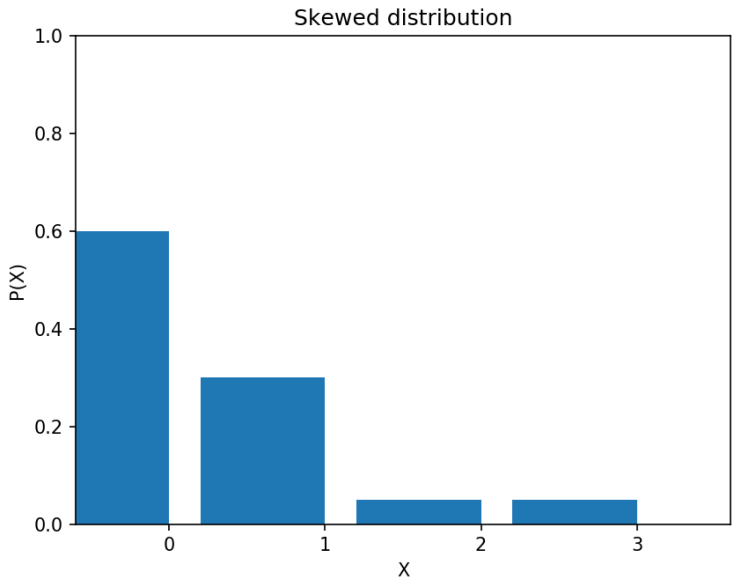




Entropy

Simpler question: Which $P(X)$ would lead you to expect more information about x from a measurement?

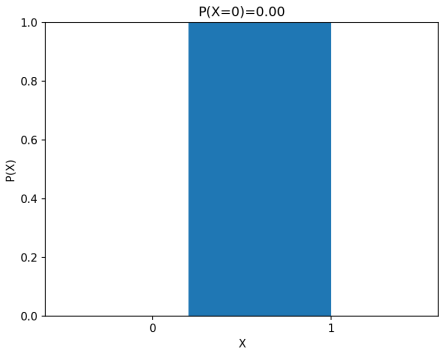


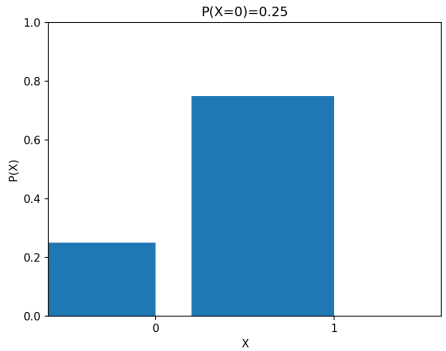


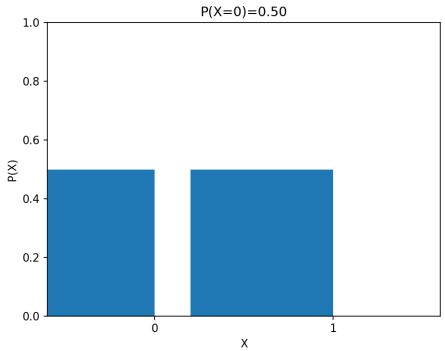
No preference, if we just swap labels on x-axis. $\Rightarrow$ expected amount of learning does not depend on X directly, only on $P(X)$

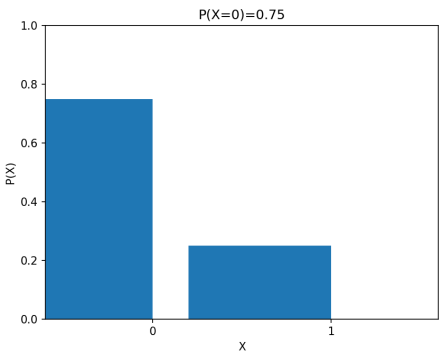
Entropy, simplest scenario: binary X

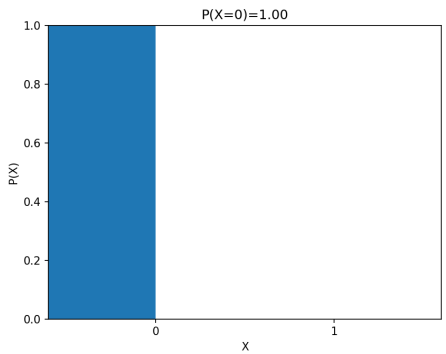
Question: how much do we expect to learn from a measurement of X?





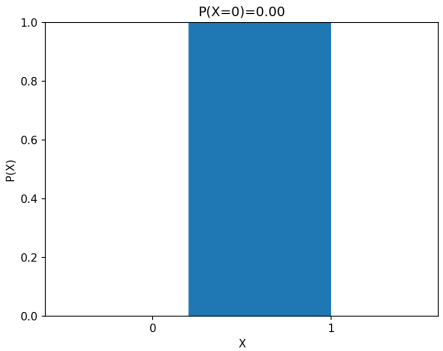




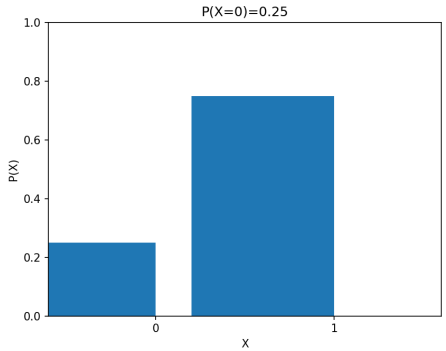


Entropy, simplest scenario: binary X

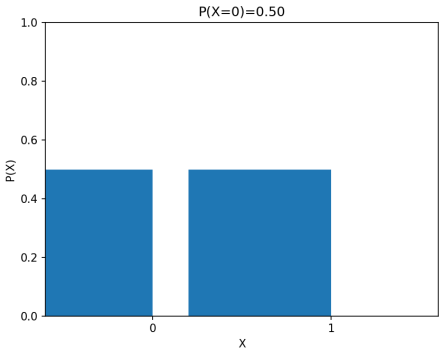
Question: how much do we expect to learn from a measurement of X?



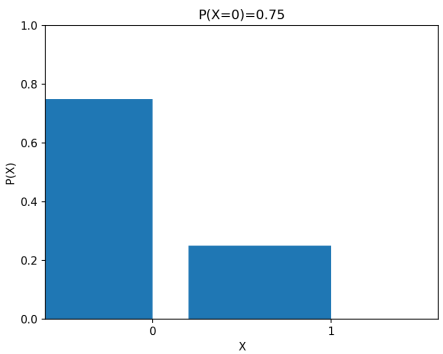
Nothing



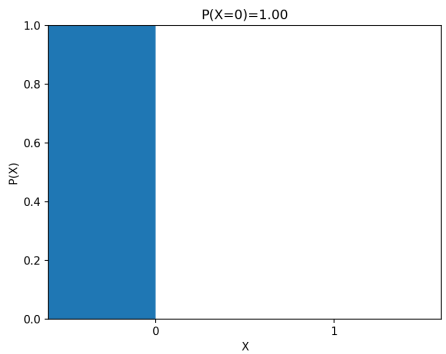
Some



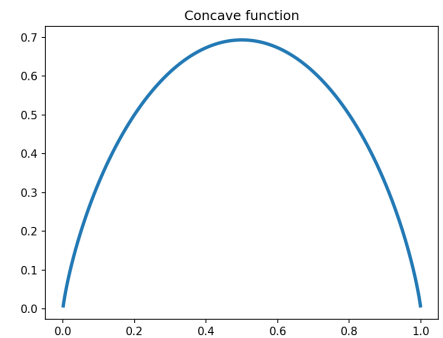
the most?



Some



Nothing



$\Rightarrow$ expected amount of learning should be a concave function of $P(X)$

Entropy: two independent random variables

Assume X and Y are independent random variables, i.e.

$$P(X, Y) = P(X)P(Y)$$

and let $H(X), H(Y)$ be the expected amount of information from measuring X and Y. Then

$$H(X, Y) = H(X) + H(Y)$$

because we would expect to learn nothing about Y from knowing X.

Entropy axioms

Let $H(X)$ be a measure of expected information with the properties

- $H(X) \geq 0$ (we always learn something or nothing)
- $H(X) = \sum_X P(X) f\left(P(X)\right)$ (it is an expectation depending on $P(X)$ only)
- $P(X) f\left(P(X)\right)$ is concave and smooth
- For $P(X, Y) = P(X)P(Y)$, $H(X, Y) = H(X) + H(Y)$ (independent infos add up)

Entropy axioms

Let $H(X)$ be a measure of expected information with the properties

- $H(X) \geq 0$ (we always learn something or nothing)
- $H(X) = \sum_X P(X) f\left(P(X)\right)$ (it is an expectation depending on $P(X)$ only)
- $P(X) f\left(P(X)\right)$ is concave and smooth
- For $P(X, Y) = P(X)P(Y)$, $H(X, Y) = H(X) + H(Y)$ (independent infos add up)

It is then possible to show that [Chakrabarti & Chakrabarty, 2005] that $f(P(X)) = -\log(P(X))$ is the only possible choice. The resulting

$$H(X) = -\sum_X P(X) \log\left(P(X)\right)$$

is called **information entropy** or **Shannon entropy**. It measures expected information gain after a measurement of X.

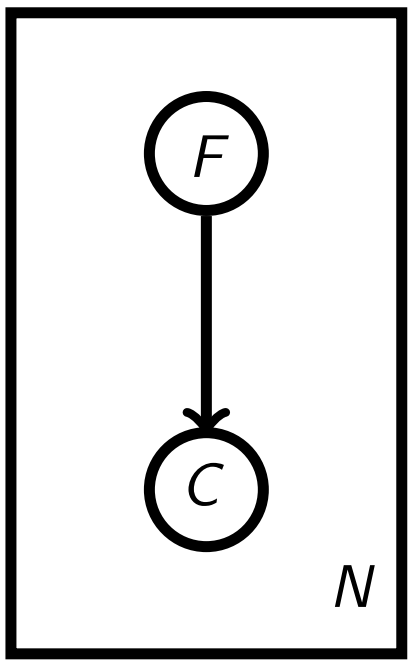
Inference & learning example: how loaded is my coin, and how often ?

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

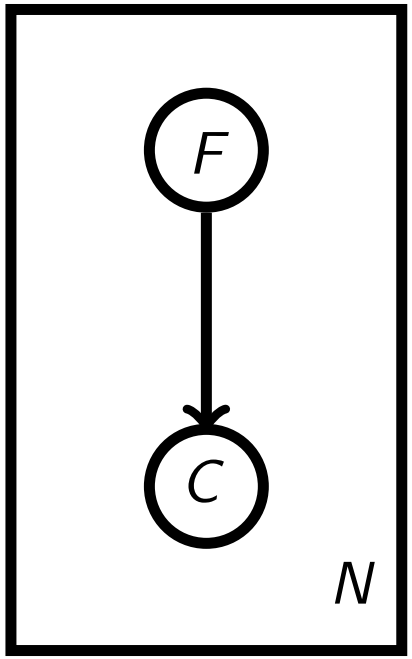
- **Perception**: is the coin loaded at toss n ? $P(F_n|C_n)$.
- **Learning**: How often is it loaded ?
- **Learning**: How loaded is it ?
- **Action**: Change C to what you like



Inference & learning example: how caffeinated is my coffee, and how much is there ?

Coffee percolator in departmental kitchen:

- Possible outcomes $C \in \{lots, little\}$.
- Possible coffee types: $F \in \{caf, decaf\}$
- You toss N times.
- You observe sequence $S = (lots, little, \ldots, lots)$.



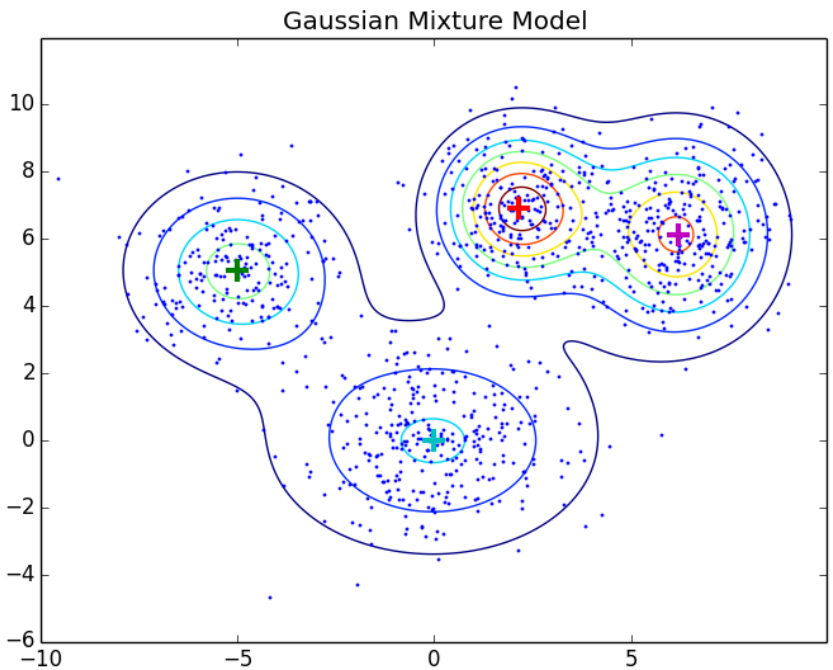
Questions

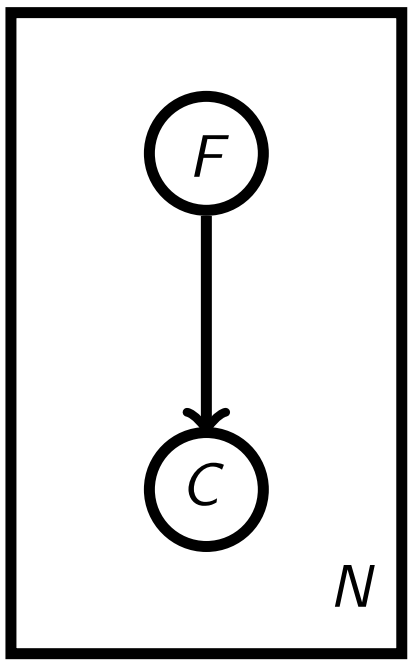
- **Perception**: is the coffee caffeinated at day n ? $P(F_n|C_n)$.
- **Learning**: How often is it caffeinated ?
- **Learning**: How much coffee should I expect ?
- **Action**: make own coffee

Inference & learning example: clustering

Gaussian clustering

- Possible outcomes $C \sim \mathcal{N}(\mu_f, \Sigma_f)$.
- Possible clusters: $F \in \{1, \ldots, K\}$



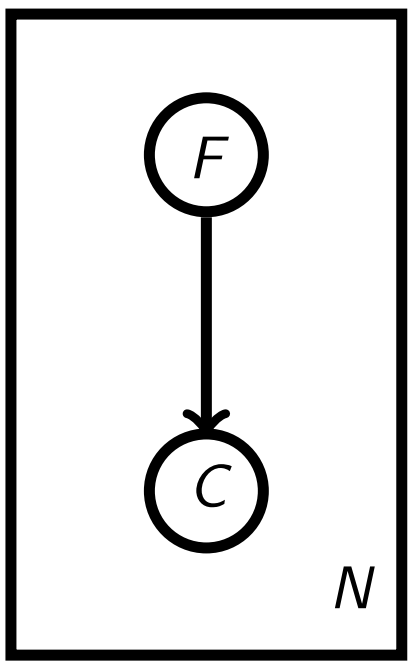


Learning example: how loaded is my coin, and how often ?

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?



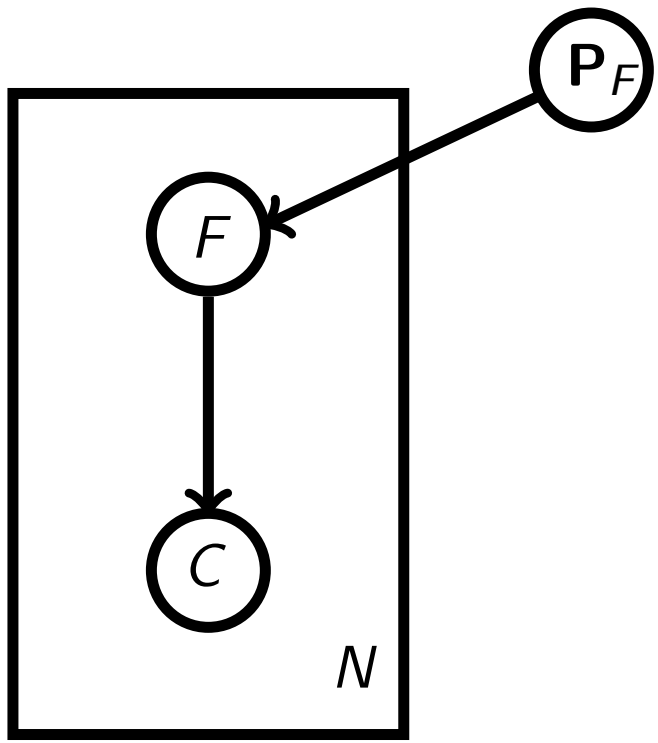
Learning example: how loaded is my coin, and how often ?

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 ?

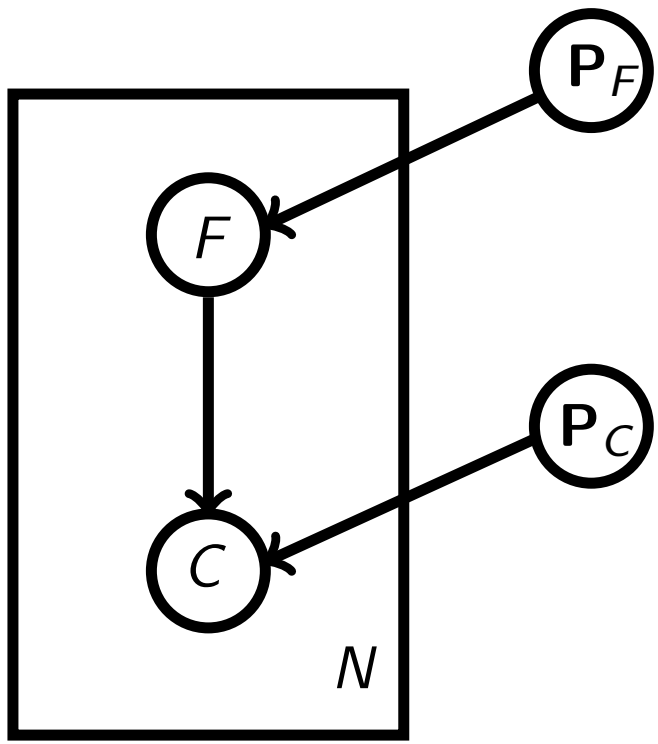
Reminder:
You are given a (new?) coin for each toss. You toss N times, observing $S = (t, h, \ldots, t)$
Possible outcomes $C_n \in h, t$. Possible coins types: $F_n \in f, l$.

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 ?

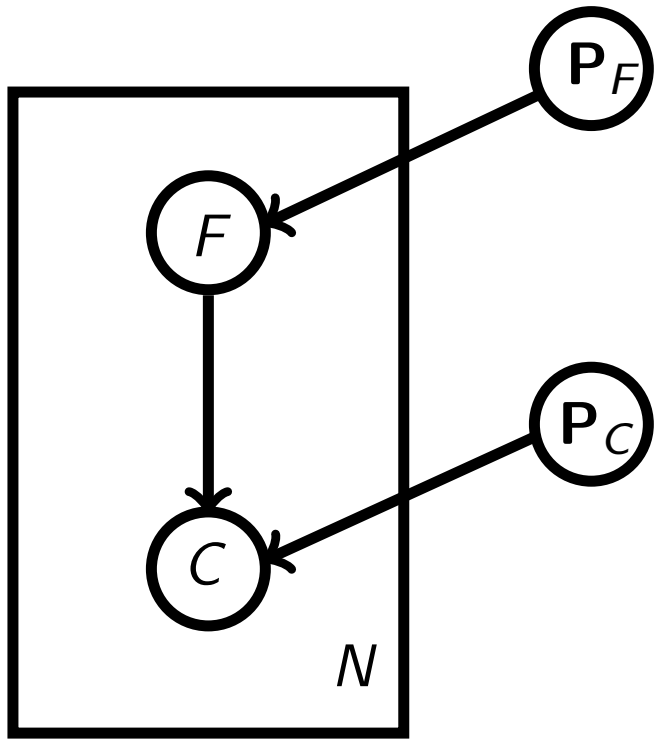
Reminder:
You are given a (new?) coin for each toss. You toss N times, observing $S = (t, h, \ldots, t)$
Possible outcomes $C_n \in h, t$. Possible coins types: $F_n \in f, l$.

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

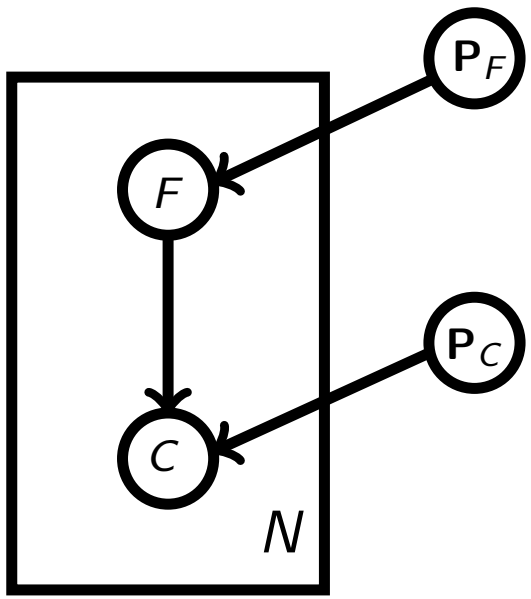


Reminder: C_n: data
F_n: hidden/latent vars.
$\mathbf{P}_F$,$\mathbf{P}_C$: parameters.

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

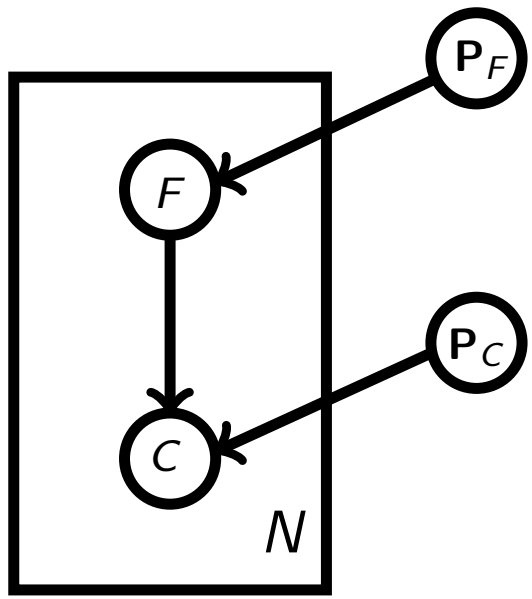
Reminder:
You are given a (new?) coin for each toss. You toss N times, observing $S = (t, h, \ldots, t)$
Possible outcomes $C_n \in h, t$. Possible coins types: $F_n \in f, l$.
$\mathbf{P}_F$ and $\mathbf{P}_C$ parameterize the distributions of F and $P(C|F = l)$.

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

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

Can we use conditioning:

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



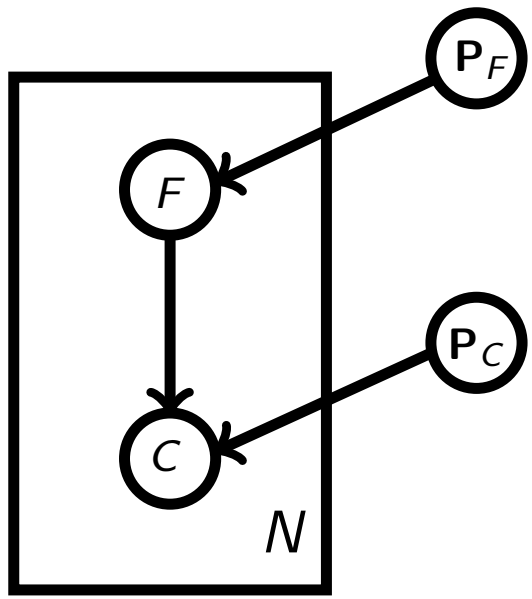
Evaluating the parameter posterior

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. $\int_Z dx\, p(x) = 1$

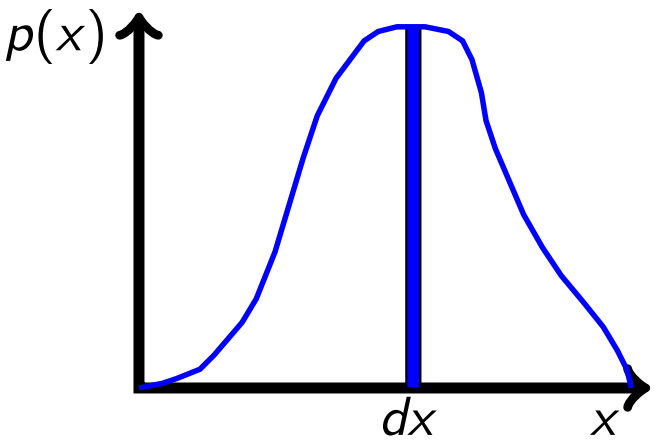
Notes:

- $p(X)$ 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: Radon-Nikodym theorem.

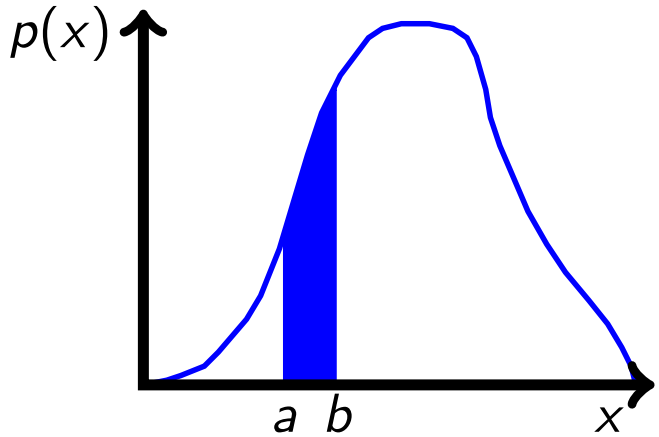
Probability density: graphical interpretation

Probability density: graphical interpretation

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



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



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: $p(X|Y) = \frac{p(X,Y)}{p(Y)}$

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: $p(X|Y) = \frac{p(X,Y)}{p(Y)}$
- 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: $p(X|Y) = \frac{p(X,Y)}{p(Y)}$
- 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: $p(X, Y, Z) = p(X|Y, Z)p(Y|Z)p(Z)$.

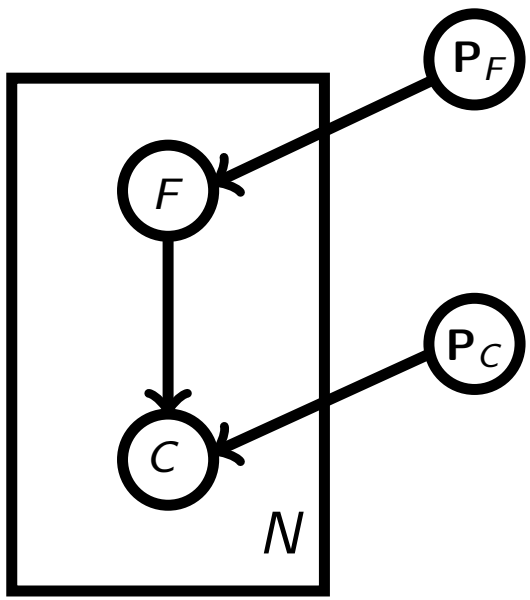
Note: densities and probability distribution can appear together in one expression.

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?

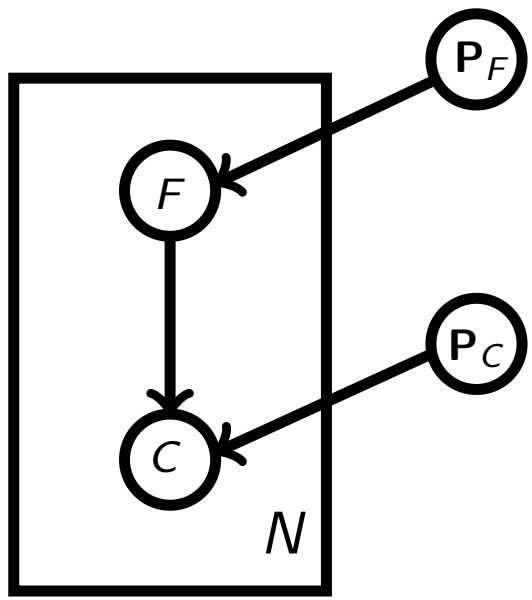
Ques.: Density $p(\mathbf{P}_F|S)$?

We can use conditioning:

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

Likewise, for $\mathbf{P}_C$

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

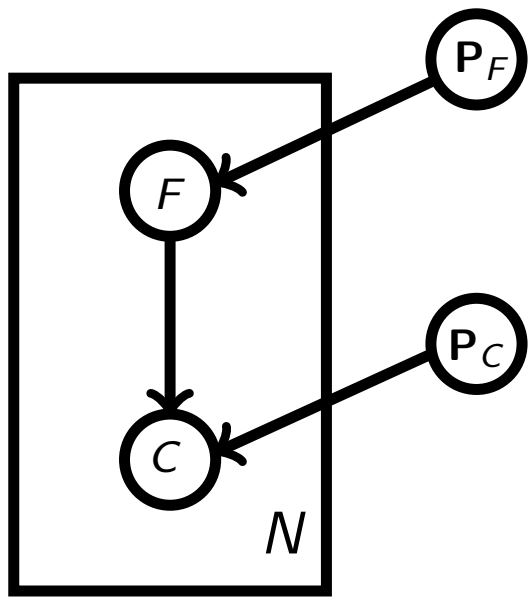


Evaluating the parameter posterior, contd.

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 $\mathbf{P}_C$

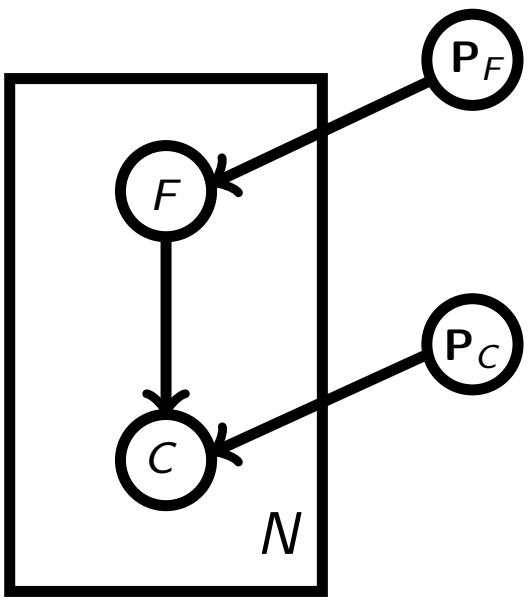
$$p(\mathbf{P}_C|S) = \frac{P(S|\mathbf{P}_C)p(\mathbf{P}_C)}{P(S)}$$

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

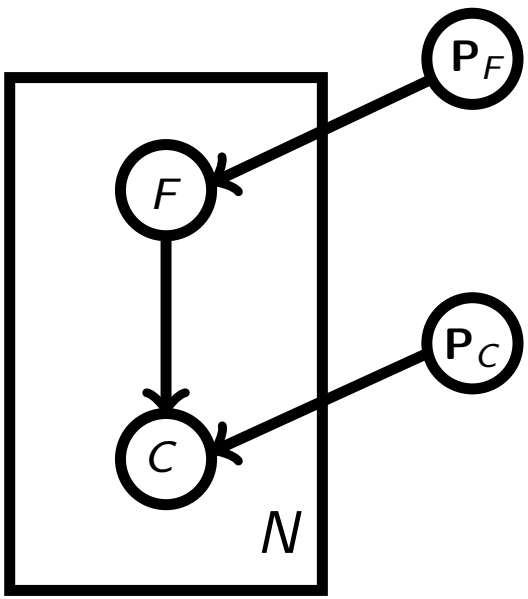
Bayesian network



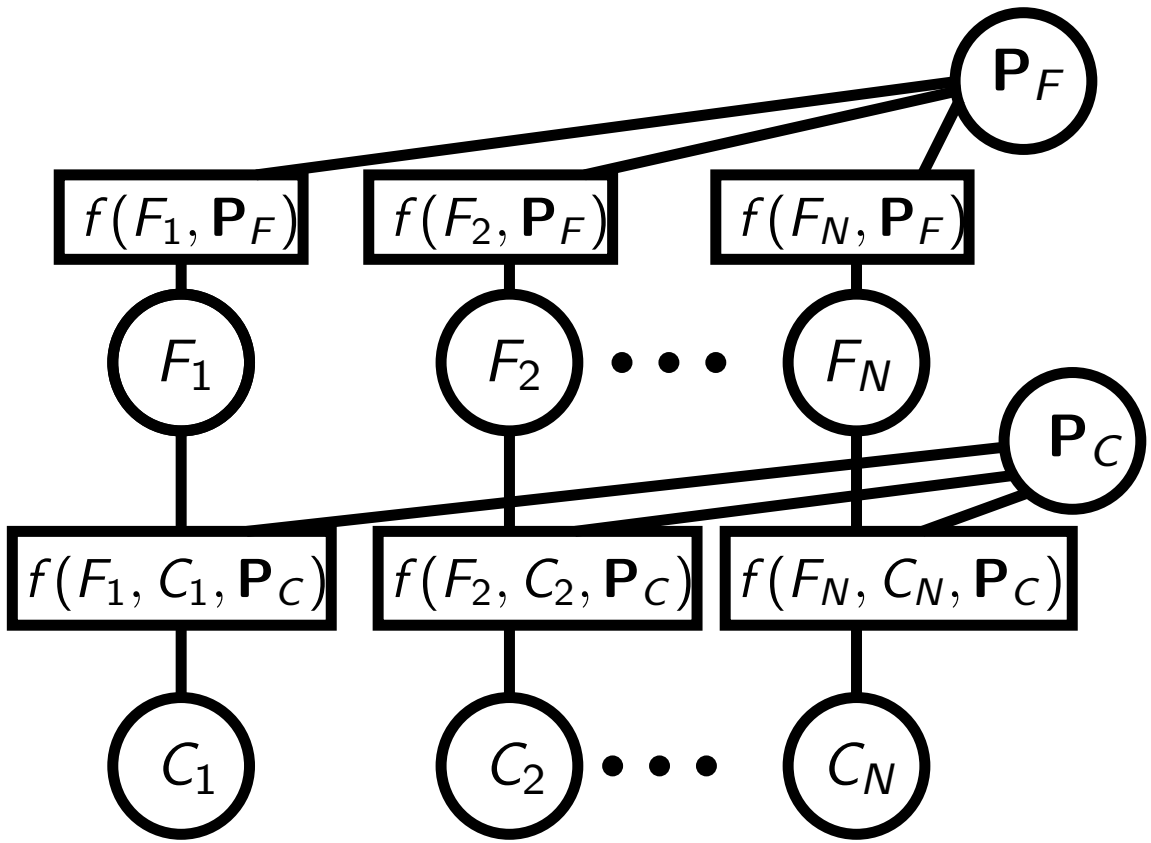
Evaluating the parameter posterior, contd.

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



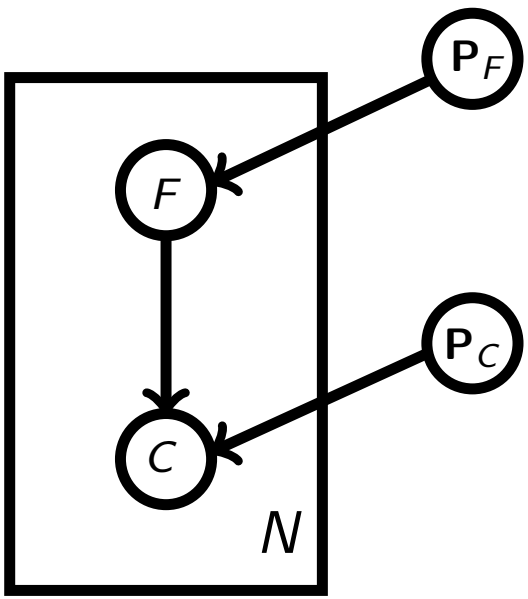
Factor graph



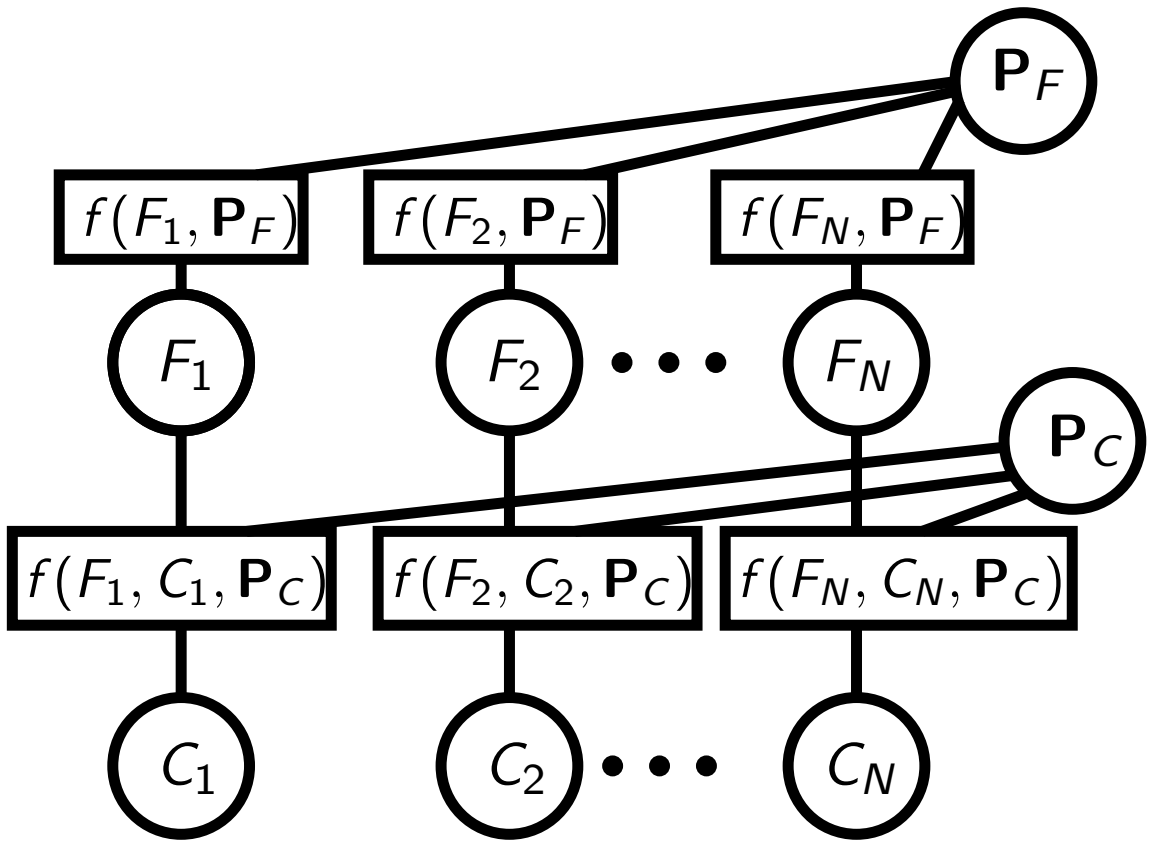
Evaluating the parameter posterior, contd.

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



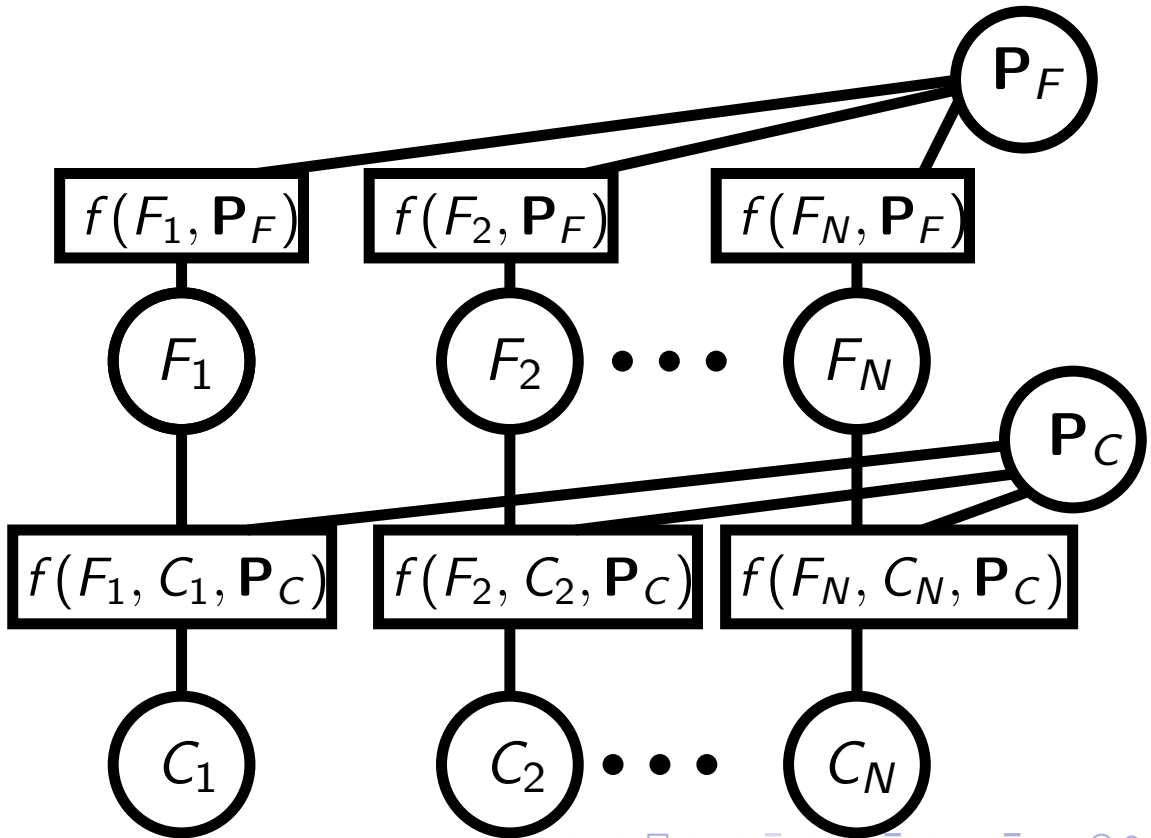
$\Rightarrow$ lots of loops!

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 $S = (t, h, \ldots, t)$
$P(F) = \mathbf{P}_F$ and $P(C|F = l) = \mathbf{P}_C$. We want: posterior densitites $p(\mathbf{P}_F|S)$ and $p(\mathbf{P}_C|S)$.

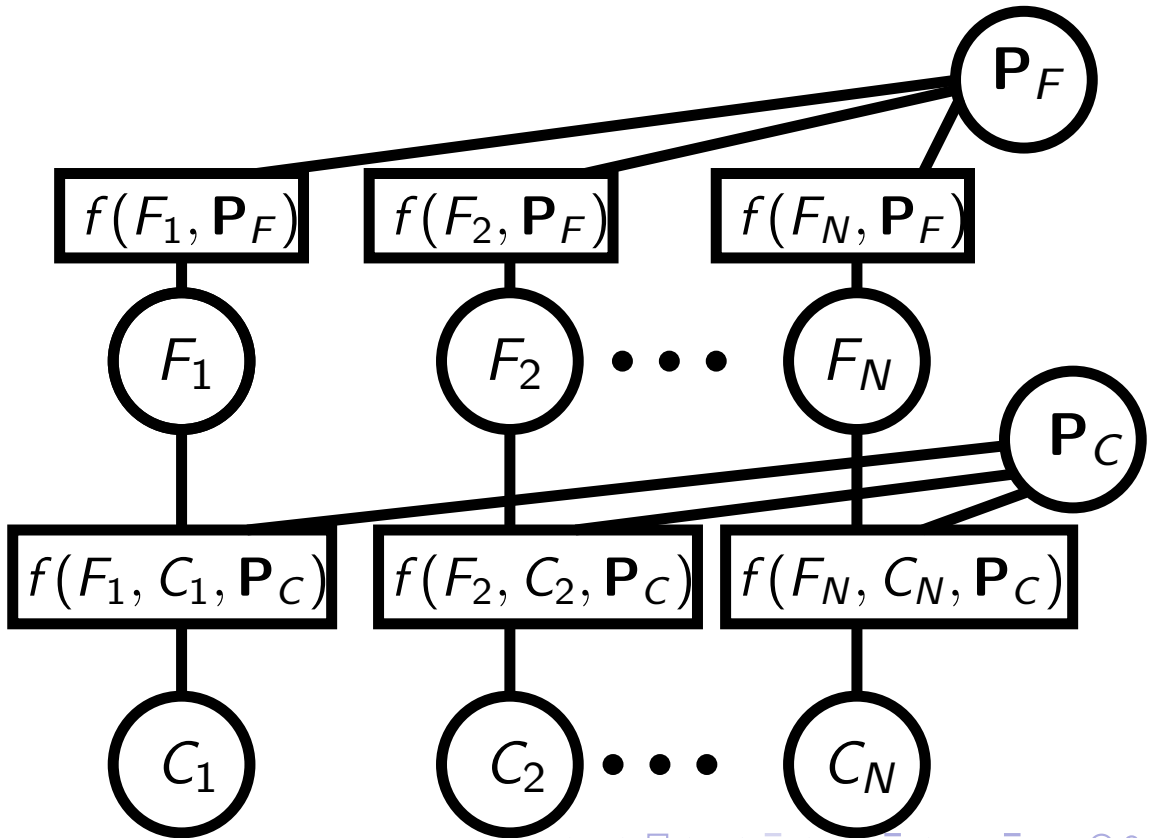
Prob. 1: loopy graph.
Prob. 2: Messages from factors to $\mathbf{P}_C$:

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

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

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

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 $\mathbf{P}_C$:

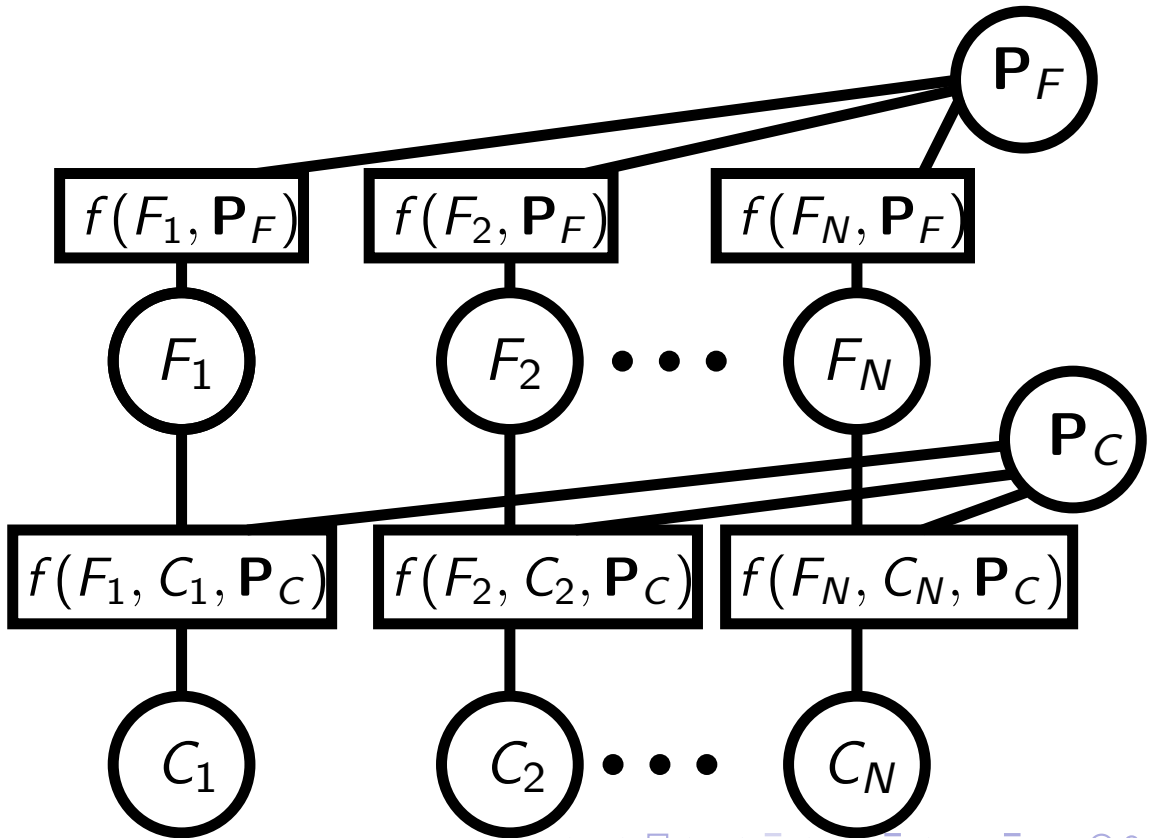
$$\mu_{f(F_1,\mathbf{P}_F)\to\mathbf{P}_F}(\mathbf{P}_F)$$

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 (make simpler) the problem 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 $S = (t, h, \ldots, t)$
$P(F) = \mathbf{P}_F$ 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 (make simpler) the problem 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 future data?
Answer: marginal probability $P(S')$ of future data S' is high.
Problem: S' unknown $\Rightarrow$ use S instead, assume stationarity.

Restating the problem: inference as optimization

General optimization problem statement (D data, H hypotheses/models):

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

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 (D data, H hypotheses/models):

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

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 (D data, H hypotheses/models):

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

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.

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

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:

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

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 $\forall 0 \leq \lambda \leq 1$ 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

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

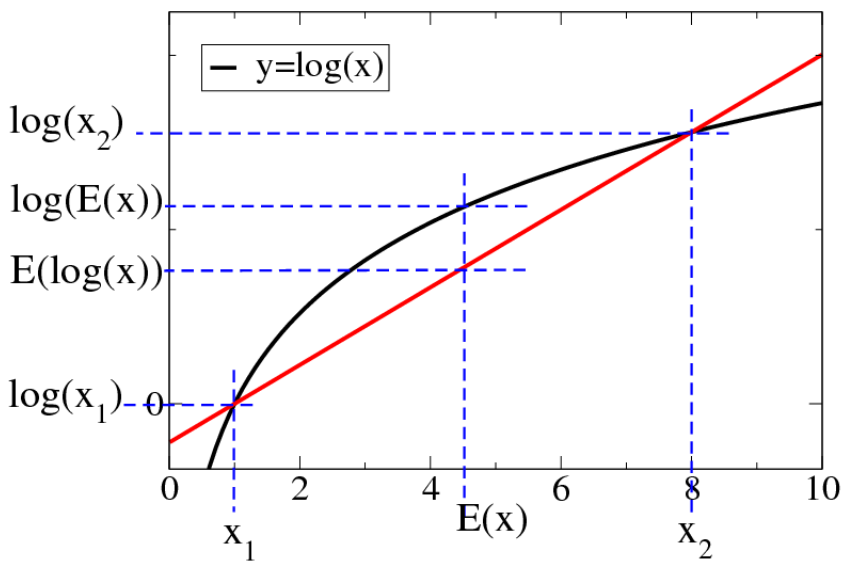
Example: $\log(x)$ is concave.

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

Then

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

.



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 $X \in \{x_1, \ldots, X_N\}$, then $P(X' = x_i) = p'_i = \frac{p_i}{1-p_N}$ is a distribution over $X \in \{x_1, \ldots, X_{N-1}\}$

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

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:

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

$\Rightarrow$ because $D(Q(H)||P(H|D)) \geq 0$, 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/percepts 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 $Q(H)$.

Interpretation 1: We found

$$\mathcal{L}(D, Q(H)) = \log(P(D)) - D(Q(H)||P(H|D))$$

In [Friston, 2006, Hohwy, 2013] terms:

- $-\log(P(D))$: surprise. Hard to compute.
- $D(Q(H)||P(H|D))$: perceptual divergence. Hard.
- $-\mathcal{L}(D, Q(H))$: free energy/predicition error

prediction error= surprise + perceptual divergence

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

Reminder:
Given: observable data D and hidden variables/parameters/percepts 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 $Q(H)$.

Interpretation 1: We found

$$\mathcal{L}(D, Q(H)) = \log(P(D)) - D(Q(H)||P(H|D))$$

In [Friston, 2006, Hohwy, 2013] terms:

- $-\log(P(D))$: surprise. Hard to compute.
- $D(Q(H)||P(H|D))$: perceptual divergence. Hard.
- $-\mathcal{L}(D, Q(H))$: free energy/predicition error

prediction error= surprise + perceptual divergence

Perception: minimize free energy $\Leftrightarrow$ minimize $D(Q(H)||P(H|D))$ by changing $Q(H)$.

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 $Q(H)$.

Question: can $\mathcal{L}(D, Q(H))$ be interpreted?
Answer 2: Note that $P(D, H) = P(D|H)P(H)$. Thus

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

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:

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

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 $\mathcal{L}(D, Q(H))$, 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.

Action: change D to make $P(D|H)$ large, i.e. make the world follow your expectations.
Active inference: choose those data that make $P(D|H)$ large.

Yet 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 3:

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

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

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

Yet 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
- $\Rightarrow$ find most probable expected joint states of data and internal model, while keeping internal model as uninformative as possible (Occam's razor)

Yet 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
- $\Rightarrow$ find most probable expected joint states of data and internal model, while keeping internal model as uninformative as possible (Occam's razor)

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.

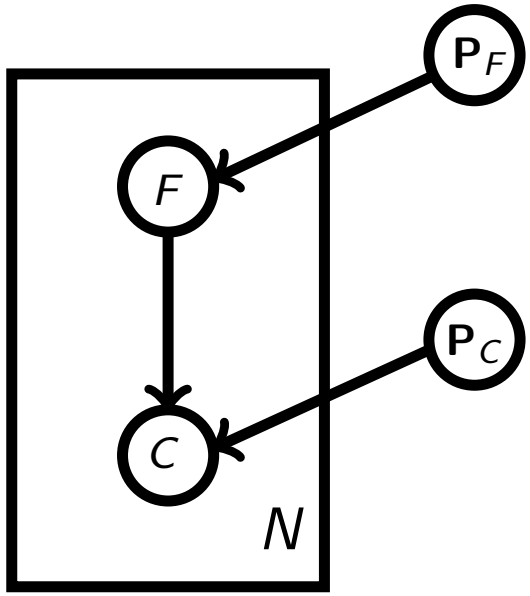
Back to the example: how loaded is my coin, and how often?

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: $q(C_i, F_i, \mathbf{P}_F, \mathbf{P}_C)$

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

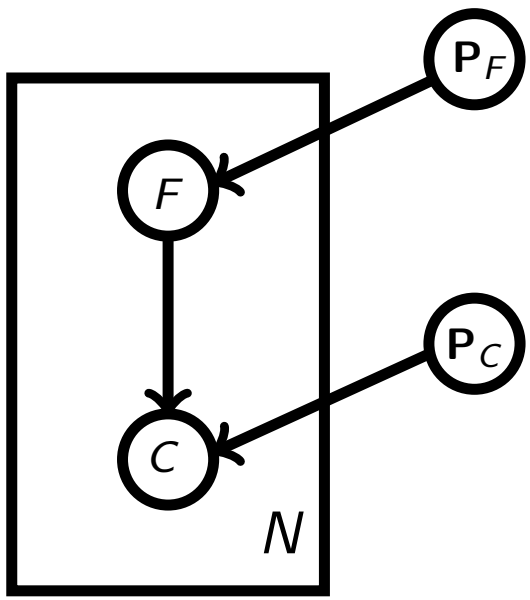


Breaking the loops

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.

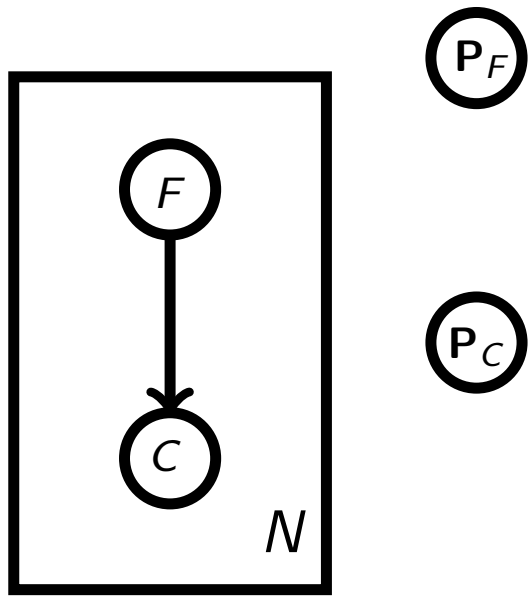


Breaking the loops

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

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

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



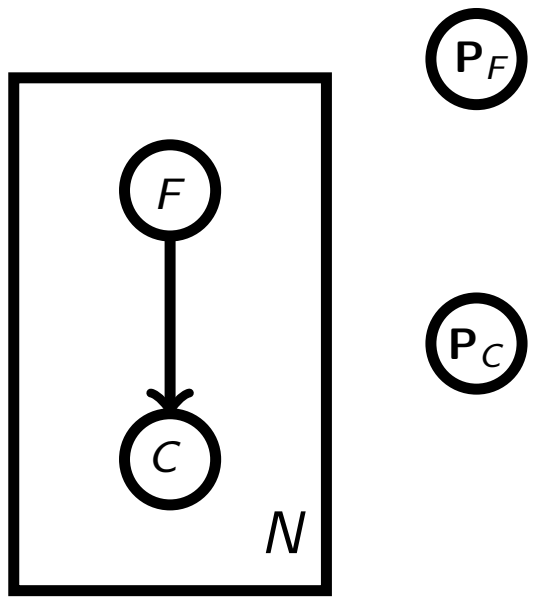
Breaking the loops

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:

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

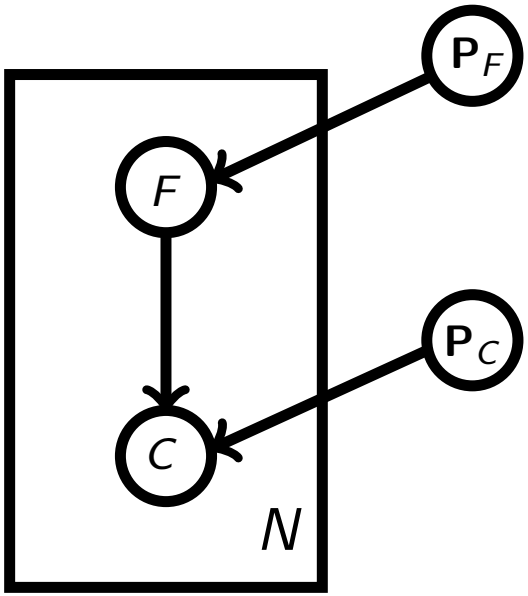
But this looks like the parameters $\mathbf{P}_F$, $\mathbf{P}_C$ are disconnected from the data C_i!



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

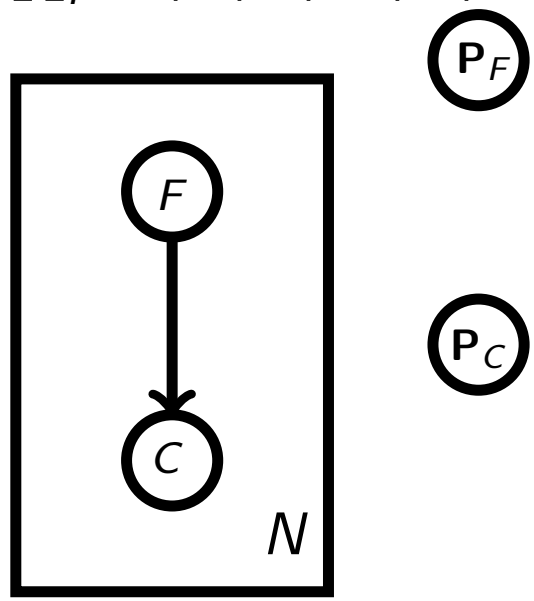
Reminder:
Lower bound $\mathcal{L}(D, Q(H)) = E_{Q(H)}\left(\log\left(\frac{P(D,H)}{Q(H)}\right)\right) \leq \log(P(D))$

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



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

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



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

Computing the bound

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

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

Putting it all together

Thus, we find for the bound

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

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.

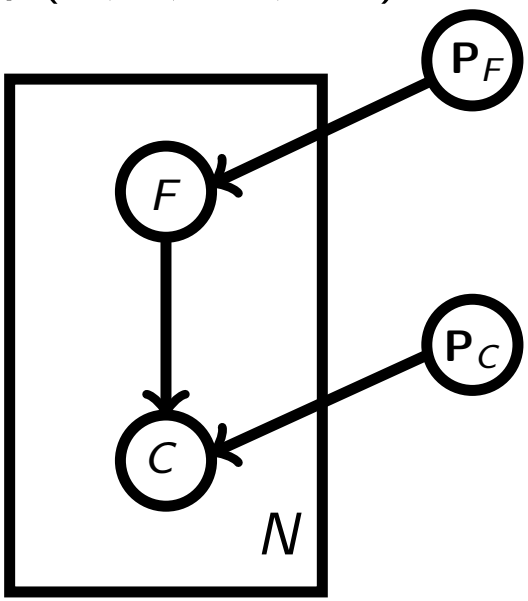
Summary 2: variational approximations

- **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 $\mathbf{P}_C$

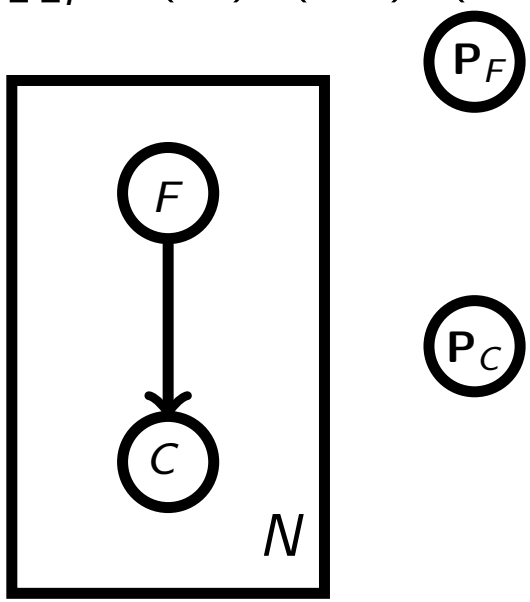
Reminder:
Lower bound $\mathcal{L}(D, Q(H)) = E_{Q(H)}\left(\log\left(\frac{P(D,H)}{Q(H)}\right)\right) \leq \log(P(D))$

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



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

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



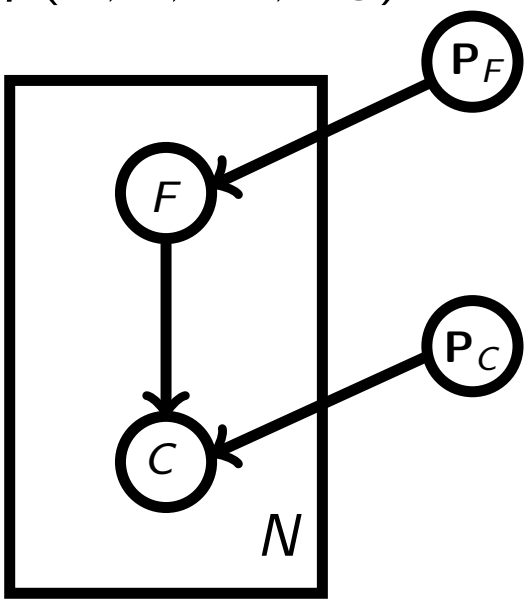
Questions:

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

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



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

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

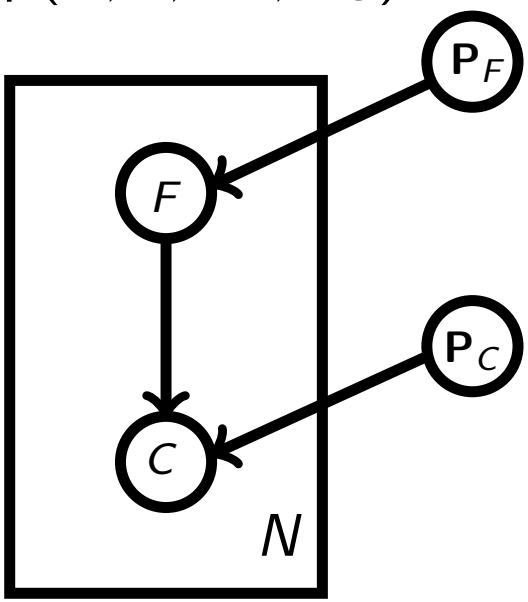
$$P(C|F = l, \mathbf{P}_C) = \begin{cases} C = h : P_h \\ C = t : P_t \end{cases}$$

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

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:

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

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:

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

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 $\mathcal{L}_{\mathcal{F}}$:

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

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:

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

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

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:

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

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:

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

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

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

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:

$$\frac{\partial E_q[\log(P(F_i|\mathbf{P}_F))]}{\partial \tilde{\nu}} = -\frac{\partial^2 \log(f(\tilde{\boldsymbol{\lambda}},\tilde{\nu}))}{\partial \tilde{\nu}^2} - \tilde{\boldsymbol{\lambda}}^T \frac{\partial \langle \boldsymbol{\eta} \rangle_{q(\mathbf{P}_F)}}{\partial \tilde{\nu}} + \frac{\partial \langle \boldsymbol{\eta} \rangle_{q(\mathbf{P}_F)}}{\partial \tilde{\nu}} \langle \mathbf{u}(F_i) \rangle_{Q_i(F_i)}$$

$$\frac{\partial D(q(\mathbf{P}_F)||p(\mathbf{P}_F))}{\partial \tilde{\nu}} = \frac{\partial \log(f(\tilde{\boldsymbol{\lambda}},\tilde{\nu}))}{\partial \tilde{\nu}} - \frac{\partial \log(f(\tilde{\boldsymbol{\lambda}},\tilde{\nu}))}{\partial \tilde{\nu}} - (\tilde{\nu}-\nu)\frac{\partial^2 \log(f(\tilde{\boldsymbol{\lambda}},\tilde{\nu}))}{\partial \tilde{\nu}^2} + (\tilde{\boldsymbol{\lambda}}^T - \boldsymbol{\lambda}^T)\nu\frac{\partial \langle \boldsymbol{\eta} \rangle_{q(\mathbf{P}_F)}}{\partial \tilde{\nu}}$$

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

Reminder:

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

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

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

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:

$$\frac{\partial E_q[\log(P(F_i|\mathbf{P}_F))]}{\partial \tilde{\nu}} = -\frac{\partial^2 \log(f(\tilde{\boldsymbol{\lambda}},\tilde{\nu}))}{\partial \tilde{\nu}^2} - \tilde{\boldsymbol{\lambda}}^T \frac{\partial \langle \boldsymbol{\eta} \rangle_{q(\mathbf{P}_F)}}{\partial \tilde{\nu}} + \frac{\partial \langle \boldsymbol{\eta} \rangle_{q(\mathbf{P}_F)}}{\partial \tilde{\nu}} \langle \mathbf{u}(F_i) \rangle_{Q_i(F_i)}$$

$$\frac{\partial D(q(\mathbf{P}_F)||p(\mathbf{P}_F))}{\partial \tilde{\nu}} = \frac{\partial \log(f(\tilde{\boldsymbol{\lambda}},\tilde{\nu}))}{\partial \tilde{\nu}} - \frac{\partial \log(f(\tilde{\boldsymbol{\lambda}},\tilde{\nu}))}{\partial \tilde{\nu}} - (\tilde{\nu}-\nu)\frac{\partial^2 \log(f(\tilde{\boldsymbol{\lambda}},\tilde{\nu}))}{\partial \tilde{\nu}^2} + (\tilde{\boldsymbol{\lambda}}^T - \boldsymbol{\lambda}^T)\nu \frac{\partial \langle \boldsymbol{\eta} \rangle_{q(\mathbf{P}_F)}}{\partial \tilde{\nu}}$$

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:

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

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

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

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:

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

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

Reminder:

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

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

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

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:

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

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 $\mathcal{L}$.