Markov and Gibbs Random Fields

Bruno Galerne
bruno.galerne@parisdescartes.fr

MAP5, Université Paris Descartes

Master MVA
Cours “Méthodes stochastiques pour l’analyse d’images”
Lundi 6 mars 2017
Outline

The Ising Model

Markov Random Fields and Gibbs Fields

Sampling Gibbs Fields
The main reference for this course is Chapter 4 of the book

*Pattern Theory: The Stochastic Analysis of Real-World Signals*
by D. Mumford and A. Desolneux
[Mumford and Desolneux 2010]
Outline

The Ising Model

Markov Random Fields and Gibbs Fields

Sampling Gibbs Fields
The Ising Model

- The Ising model is the simplest and most famous Gibbs model.
- It has its origin in statistical mechanics but can be presented in the context of image segmentation.

Framework:
- A real-valued image $I \in \mathbb{R}^{M \times N}$ of size $M \times N$ is given (with gray-levels mostly in $[-1, 1]$ after some contrast change).
- One wants to associate a binary image $J \in \{-1, 1\}^{M \times N}$ with values in $\{-1, 1\}$ ($-1$ for black, and $1$ for white) supposed to model the predominantly black/white areas of $I$.
- There is only a finite number of $2^{MN}$ values for $J$.

Input image $I$

A realization $J$ of the associated Ising model (with $T = 1.5$)
The Ising Model

Notation:
- $\Omega_{M,N} = \{0, \ldots, M-1\} \times \{0, \ldots, N-1\}$ is the set of pixel indexes.
- For keeping the notation short we will denote by $\alpha$ or $\beta$ pixel coordinates $(k, l)$, $(m, n)$, so that $I(\alpha) = I(k, l)$ for some $(k, l)$.
- We note $\alpha \sim \beta$ to mean that pixels $\alpha$ and $\beta$ are neighbors for the 4-connectivity (each pixel has 4 neighbors, except at the border).

Energy:
To each couple of images $(I, J)$, $I \in \mathbb{R}^{M \times N}$, $J \in \{-1, 1\}^{M \times N}$, one associates the energy $E(I, J)$ defined by

$$E(I, J) = c \sum_{\alpha \in \Omega_{M,N}} (I(\alpha) - J(\alpha))^2 + \sum_{\alpha \sim \beta} (J(\alpha) - J(\beta))^2$$

where $c > 0$ is a positive constant and the sum $\sum_{\alpha \sim \beta}$ means that each pair of connected pixels is summed once (and not twice).

- The first term of the energy measures the similarity between $I$ and $J$.
- The second term measures the similarity between pixels of $J$ that are neighbors.
The Ising Model

Energy:

\[ E(I, J) = c \sum_{\alpha \in \Omega_{M,N}} (I(\alpha) - J(\alpha))^2 + \sum_{\alpha \sim \beta} (J(\alpha) - J(\beta))^2 \]

The Ising model associated with \( I \): For any fixed image \( I \), this energy enables to define a discrete probability distribution on \( \{-1, 1\}^{M \times N} \) by

\[ p_T(J) = \frac{1}{Z_T} e^{-\frac{1}{T}E(I,J)}, \]

where \( T > 0 \) is a constant called the \textit{temperature} and \( Z_T \) is the normalizing constant

\[ Z_T = \sum_{J \in \{-1,1\}^{M \times N}} e^{-\frac{1}{T}E(I,J)}. \]

The probability distribution \( p_T \) is the Ising model associated with \( I \) (and constant \( c \) and temperature \( T \)).
The Ising Model

Energy:

\[ E(I, J) = c \sum_{\alpha \in \Omega_{M,N}} (I(\alpha) - J(\alpha))^2 + \sum_{\alpha \sim \beta} (J(\alpha) - J(\beta))^2 \]

Probability distribution:

\[ p_T(J) = \frac{1}{Z_T} e^{-\frac{1}{T} E(I, J)} \]

- Minimizing with respect to \( J \) the energy \( E(I, J) \) is equivalent to find the most probable state of the discrete distribution \( p_T \).
- Note that this most probable state (also called the mode of the distribution) is the same for all temperatures \( T \).
- As \( T \) tends to 0, the distribution \( p_T \) tends to be concentrated at this mode.
- As \( T \) tends to \( +\infty \), \( p_T \) tends to be uniform over all the possible image configurations.
- Hence the temperature parameter \( T \) controls the amount of allowed randomness around the most probable state.
The main questions regarding the Ising model are the following:

1. **How can we sample efficiently from the discrete distribution $p_T$?**

   Although the distribution is a discrete probability on the finite set of binary images $\{-1, 1\}^{M \times N}$, one cannot compute the table of the probability distribution even for $10 \times 10$ images! Indeed, the memory size of the table would be

   $$2^{10 \times 10} \times 4 > 10^{30} \text{ Bytes} = 10^{18} \text{ TeraBytes}$$

2. **How can we compute the common mode(s) of the distributions $p_T$, that is (one of) the most probable state of the distributions $p_T$?**

3. **What is the good order for the temperature value $T$?**
The main questions regarding the Ising model are the following:

1. How can we sample efficiently from the discrete distribution $p_T$? Although the distribution is a discrete probability on the finite set of binary images $\{-1, 1\}^{M \times N}$, one cannot compute the table of the probability distribution even for $10 \times 10$ images! Indeed, the memory size of the table would be

$$2^{10 \times 10} \times 4 > 10^{30} \text{ Bytes} = 10^{18} \text{ TeraBytes}$$

2. How can we compute the common mode(s) of the distributions $p_T$, that is (one of) the most probable state of the distributions $p_T$?

3. What is the good order for the temperature value $T$?
The main questions regarding the Ising model are the following:

1. How can we sample efficiently from the discrete distribution $p_T$? Although the distribution is a discrete probability on the finite set of binary images $\{-1, 1\}^{M \times N}$, one cannot compute the table of the probability distribution even for $10 \times 10$ images! Indeed, the memory size of the table would be

$$2^{10 \times 10} \times 4 > 10^{30} \text{ Bytes} = 10^{18} \text{ TeraBytes}$$

2. How can we compute the common mode(s) of the distributions $p_T$, that is (one of) the most probable state of the distributions $p_T$?

3. What is the good order for the temperature value $T$?
Figure 4.4: Samples from the Ising model: the gray-level image on the top left is scaled to have max/min pixel values 0, +1 and is used as the external magnetic field $I$ in the Ising model. The black/white image in the top-middle is the mode of the probability distribution, the least energy field $J$, when the constant $c = 4$ in the energy. The top-right when the constant $c = 4$ in the energy. The top-right shows values of the energy $E(I, J)$ for $J$ sampled from $P_T(J|I)$ at various temperatures $T$. The bottom line shows random samples $J$ taken from the Ising model conditioned on fixed $I$ (i.e. from $P_T(J|I)$) at temperatures $T = 10, 4, 2.5, 1.5$ and 1. Note how the shapes emerge between the 2nd and 3rd sample, where the mean energy, as seen in the graph, has large derivative with respect to $T$. 

The Ising Model
Outline

The Ising Model

Markov Random Fields and Gibbs Fields

Sampling Gibbs Fields
A General Framework

- The Ising model is a very special case of a large class of stochastic models called **Gibbs fields**.

**General framework:**
- We consider an arbitrary **finite graph** \((V, E)\) with vertices \(V\) and undirected edges \(E \subset V \times V\).
- Each vertex has a random label in a **phase space** \(\Lambda\) (also supposed to be finite, e.g. discrete gray-levels).
- The set \(\Omega = \Lambda^V\) of all label configurations is called the **state space**.
- The elements of \(\Omega = \Lambda^V\) are denoted by \(x, y, \ldots\), that is, \(x = (x_\alpha)_{\alpha \in V}, x_\alpha \in \Lambda\).

**Example of the Ising model:**
- The vertices \(V\) are the pixels of \(\Omega_{M,N}\).
- The edges are the ones of the 4-connectivity.
- The phase space is \(\Lambda = \{-1, 1\}\).
- The state space is the set of binary images \(\Omega = \{-1, 1\}^{M \times N}\).
Markov Random Fields

Definition (Markov Random Field)

A random variable $\mathcal{X} = (\mathcal{X}_\alpha)_{\alpha \in V}$ with values in the state space $\Omega = \Lambda^V$ is a Markov random field (MRF) if for any partition $V = V_1 \cup V_2 \cup V_3$ of the set of vertices $V$ such that there is no edge between vertices of $V_1$ and $V_3$, and denoting by $\mathcal{X}_i = \mathcal{X}|_{V_i}$ the restriction of $\mathcal{X}$ to $V_i$, $i = 1, 2, 3$, one has

$$P(\mathcal{X}_1 = x_1 | \mathcal{X}_2 = x_2, \mathcal{X}_3 = x_3) = P(\mathcal{X}_1 = x_1 | \mathcal{X}_2 = x_2)$$

whenever both conditional probabilities are well defined, that is whenever $P(\mathcal{X}_2 = x_2, \mathcal{X}_3 = x_3) > 0$.

- The above property is called the **Markov property**.
- It is equivalent to saying that $\mathcal{X}_1$ and $\mathcal{X}_3$ are conditionally independent given $\mathcal{X}_2$, that is,

$$P(\mathcal{X}_1 = x_1, \mathcal{X}_3 = x_3 | \mathcal{X}_2 = x_2) = P(\mathcal{X}_1 = x_1 | \mathcal{X}_2 = x_2)P(\mathcal{X}_3 = x_3 | \mathcal{X}_2 = x_2)$$

whenever the conditional probabilities are well defined, that is whenever $P(\mathcal{X}_2 = x_2) > 0$...
Proof: Conditional independence $\Rightarrow$ Markov property

Assuming $\mathbb{P}(\mathcal{X}_2 = x_2, \mathcal{X}_3 = x_3) > 0$ (since otherwise there is nothing to show):

$$
\mathbb{P}(\mathcal{X}_1 = x_1 | \mathcal{X}_2 = x_2, \mathcal{X}_3 = x_3) = \frac{\mathbb{P}(\mathcal{X}_1 = x_1, \mathcal{X}_2 = x_2, \mathcal{X}_3 = x_3)}{\mathbb{P}(\mathcal{X}_2 = x_2, \mathcal{X}_3 = x_3)} \\
= \frac{\mathbb{P}(\mathcal{X}_1 = x_1, \mathcal{X}_3 = x_3 | \mathcal{X}_2 = x_2)}{\mathbb{P}(\mathcal{X}_3 = x_3 | \mathcal{X}_2 = x_2)} \\
= \frac{\mathbb{P}(\mathcal{X}_1 = x_1 | \mathcal{X}_2 = x_2)}{\mathbb{P}(\mathcal{X}_3 = x_3 | \mathcal{X}_2 = x_2)} \mathbb{P}(\mathcal{X}_3 = x_3 | \mathcal{X}_2 = x_2) \\
= \mathbb{P}(\mathcal{X}_1 = x_1 | \mathcal{X}_2 = x_2).$

Proof: Markov property $\Rightarrow$ Conditional independence

Assuming $\mathbb{P}(X_2 = x_2, X_3 = x_3) > 0$ (since otherwise there is nothing to show):

$$\mathbb{P}(X_1 = x_1 | X_2 = x_2) \mathbb{P}(X_3 = x_3 | X_2 = x_2)$$

$$= \mathbb{P}(X_1 = x_1 | X_2 = x_2, X_3 = x_3) \mathbb{P}(X_3 = x_3 | X_2 = x_2)$$

$$= \frac{\mathbb{P}(X_1 = x_1, X_2 = x_2, X_3 = x_3)}{\mathbb{P}(X_2 = x_2, X_3 = x_3)} \frac{\mathbb{P}(X_2 = x_2, X_3 = x_3)}{\mathbb{P}(X_2 = x_2)}$$

$$= \mathbb{P}(X_1 = x_1, X_3 = x_3 | X_2 = x_2).$$
Definition (Clique)
Given a graph \((V, E)\), a subset \(C \subset V\) is a clique if any distinct vertices of \(C\) are linked with an edge of \(E\):

\[
\forall \alpha, \beta \in C, \quad \alpha \neq \beta \Rightarrow (\alpha, \beta) \in E.
\]

The set of cliques of a graph is denoted by \(C\).

Example

1. The singletons \(\{\alpha\}\) of \(V\) are always cliques.
2. For the graph of the 4-connectivity, the cliques are
   - the singletons,
   - the pairs of horizontal adjacent pixels,
   - the pairs of vertical adjacent pixels.
Gibbs Fields

Definition (Families of potentials)
A family of function $U_C : \Omega \rightarrow \mathbb{R}$, $C \in \mathcal{C}$, is said to be a family of potentials if each function $U_C$ only depends of the restriction on the cliques $C$:

$$\forall x, y \in \Omega, \quad x|_C = y|_C \Rightarrow U_C(x) = U_C(y).$$

Definition (Gibbs distribution and Gibbs field)
A Gibbs distribution on the state space $\Omega = \Lambda^V$ is a probability distribution $P$ that comes from an energy function deriving from a family of potentials on cliques $(U_C)_{C \in \mathcal{C}}$:

$$P(x) = \frac{1}{Z} e^{-E(x)}, \quad \text{where} \quad E(x) = \sum_{C \in \mathcal{C}} U_C(x),$$

and $Z$ is the normalizing constant $Z = \sum_{x \in \Omega} e^{-E(x)}$. A Gibbs field is a random variable $\mathcal{X} = (\mathcal{X}_\alpha)_{\alpha \in \mathcal{V}} \in \Omega$ such that its law is a Gibbs distribution.
Gibbs Fields

Gibbs distribution:

\[ P(x) = \frac{1}{Z} e^{-E(x)}, \quad \text{where} \quad E(x) = \sum_{C \in C} U_C(x). \]

Example

The distribution of the Ising model is given by

\[ p_T(J) = \frac{1}{Z_T} e^{-\frac{1}{T} E(I,J)}, \quad \text{where} \]

\[ E(I,J) = c \sum_{\alpha \in \Omega_{M,N}} (I(\alpha) - J(\alpha))^2 + \sum_{\alpha \sim \beta} (J(\alpha) - J(\beta))^2. \]

Hence it is a Gibbs distribution that derives from the family of potentials

\[ U_{\{\alpha\}} = \frac{c}{T} (I(\alpha) - J(\alpha))^2, \quad \alpha \in V, \]

and

\[ U_{\{\alpha,\beta\}} = \frac{1}{T} (J(\alpha) - J(\beta))^2, \quad (\alpha, \beta) \in E. \]
Gibbs Fields

Gibbs distribution:

\[ P(x) = \frac{1}{Z} e^{-E(x)}, \quad \text{where} \quad E(x) = \sum_{C \in C} U_C(x). \]

**Remark:** One can always introduce a temperature parameter \( T \): For all \( T > 0 \),

\[ P_T(x) = \frac{1}{Z_T} e^{-\frac{1}{T} E(x)} \]

is also a Gibbs distribution.
The Hammersley-Clifford Theorem

**Theorem (Hammersley-Clifford)**

Gibbs Fields and MRF are equivalent in the following sense:

1. If $\mathcal{X}$ is a Gibbs field then it satisfies the Markov property.
2. If $P$ is the distribution of a MRF such that $P(x) > 0$ for all $x \in \Omega$, then there exists an energy function $E$ deriving from a family of potentials $(U_C)_{C \in \mathcal{C}}$ such that

$$P(x) = \frac{1}{Z} e^{-E(x)}.$$

**Take-away message:**

Gibbs fields and Markov random fields are essentially the same thing.
Proof of Gibbs fields are MRF

- Let $\mathcal{X}$ be a Gibbs field with distribution $P(x) = \frac{1}{Z} e^{-E(x)}$, where $E(x) = \sum_{C \in \mathcal{C}} U_C(x)$.

- Let $V = V_1 \cup V_2 \cup V_3$ be any partition of the set of vertices $V$ such that there is no edge between vertices of $V_1$ and $V_3$.

- For each $i = 1, 2, 3$, one denotes by $x_i = x_{\mid V_i}$ the restriction of a state $x$ to $V_i$.

- Remark that a clique $C \in \mathcal{C}$ cannot intersect both $V_1$ and $V_3$, that is, either $C \subset V_1 \cup V_2$ or $C \subset V_2 \cup V_3$.

- Since $U_C(x)$ only depends on the restriction of $x$ to $C$, $U_C(x)$ either depends on $(x_1, x_2)$ or $(x_2, x_3)$.

One can thus write

$$E(x) = E_1(x_1, x_2) + E_2(x_2, x_3) \quad \text{and thus} \quad P(x) = F_1(x_1, x_2)F_2(x_2, x_3).$$

Hence

$$P(x_1 \mid x_2, x_3) = \frac{P(x_1, x_2, x_3)}{P(x_2, x_3)} = \frac{P(x_1, x_2, x_3)}{\sum_{y \in \Lambda V_1} P(y, x_2, x_3)} = \frac{F_1(x_1, x_2)F_2(x_2, x_3)}{\sum_{y \in \Lambda V_1} F_1(y, x_2)F_2(x_2, x_3)}$$

$$= \frac{F_1(x_1, x_2)}{\sum_{y \in \Lambda V_1} F_1(y, x_2)}$$

$$= \frac{P(x_1, x_2)}{P(x_2)} = P(x_1 \mid x_2).$$
The Hammersley-Clifford Theorem

- Hence a Gibbs field satisfies the Markov property, it is a MRF.
- The difficult part of the Hammersley-Clifford theorem is the converse implication!
- It solely relies on the Möbius inversion formula.

Proposition (Möbius inversion formula)

Let \( V \) be a finite subset and \( f \) and \( g \) be two functions defined on the set \( \mathcal{P}(V) \) of subsets of \( V \). Then,

\[
\forall A \subset V, \quad f(A) = \sum_{B \subset A} (-1)^{|A \setminus B|} g(B) \iff \forall A \subset V, \quad g(A) = \sum_{B \subset A} f(B).
\]
Proof of Möbius inversion formula

Simple fact: Let $n$ be an integer, then

$$\sum_{k=0}^{n} \binom{n}{k} (-1)^k = \begin{cases} 1 & \text{if } n = 0, \\ 0 & \text{if } n > 0, \end{cases}$$

(since it is equal to $(1 + (-1))^n$ !)

Proof of $\forall A \subset V, \; f(A) = \sum_{B \subset A} (-1)^{|A \setminus B|} g(B) \Rightarrow \forall A \subset V, \; g(A) = \sum_{B \subset A} f(B)$:

Let $A \subset V$. Then,

$$\sum_{B \subset A} f(B) = \sum_{B \subset A} \sum_{D \subset B} (-1)^{|B \setminus D|} g(D)$$

$$= \sum_{D \subset A} \sum_{B \supset D} (-1)^{|B \setminus D|} g(D)$$

$$= \sum_{D \subset A} \sum_{E \subset A \setminus D} (-1)^{|E|} g(D)$$

$$= \sum_{D \subset A} g(D) \sum_{E \subset A \setminus D} (-1)^{|E|}$$

$$= \sum_{D \subset A} g(D) \sum_{k=0}^{\left|\frac{A \setminus D}{k}\right|} \binom{\left|\frac{A \setminus D}{k}\right|}{k} (-1)^k = g(A).$$

= 0 if $|A \setminus D| > 0$ and 1 if $D = A$
Proof of Möbius inversion formula

Proof of $\forall A \subset V, \ g(A) = \sum_{B \subset A} f(B) \Rightarrow \forall A \subset V, \ f(A) = \sum_{B \subset A} (-1)^{|A \setminus B|} g(B)$:

Let $A \subset V$. Then,

$$\sum_{B \subset A} (-1)^{|A \setminus B|} g(B) = \sum_{B \subset A} (-1)^{|A \setminus B|} \sum_{D \subset B} f(D)$$

$$= \sum_{D \subset A} \sum_{B \supset D} (-1)^{|A \setminus B|} f(D)$$

$$= \sum_{D \subset A} \sum_{E \subset A \setminus D} (-1)^{|A \setminus (E \cup D)|}$$

$$= \sum_{D \subset A} f(D) \sum_{E \subset A \setminus D} (-1)^{|A| - |E| - |D|}$$

$$= \sum_{D \subset A} f(D)(-1)^{|A| - |D|} \sum_{E \subset A \setminus D} (-1)^{|E|}$$

$$= \sum_{D \subset A} f(D)(-1)^{|A| - |D|} \sum_{k=0}^{\left\lfloor \frac{|A \setminus D|}{|D|} \right\rfloor} (-1)^k$$

$$= 0 \text{ if } |A \setminus D| > 0 \text{ and } 1 \text{ if } D = A$$

$$\sum_{k=0}^{\left\lfloor \frac{|A \setminus D|}{|D|} \right\rfloor} (-1)^k = f(A).$$
Proof of MRF with Positive Probability Are Gibbs Fields

Notation:

- Let $\lambda_0 \in \Lambda$ denote a fixed value in the phase space $\Lambda$.
- For a subset $A \subset V$, and a configuration $x \in \Omega = \Lambda^V$, we denote by $xA$ the configuration that is the same as $x$ at each site of $A$ and is $\lambda_0$ elsewhere, that is,

$$ (xA)_\alpha = \begin{cases} x_\alpha & \text{if } \alpha \in A, \\ \lambda_0 & \text{otherwise.} \end{cases} $$

- In particular, we have $xV = x$ and we denote $x_0 = x\emptyset$ the configuration with the value $\lambda_0$ at all the sites.

Candidate energy:

- Remark that if $P$ is a Gibbs distribution with energy $E$, then,

$$ P(x) = \frac{1}{Z} e^{-E(x)} \quad \Rightarrow \quad E(x) = \log \frac{P(x_0)}{P(x)} + E(x_0). $$

Besides, the constant $E(x_0)$ has no influence since it can be included in $Z$.

- Hence given our Markov distribution $P$, we now define a function $E : \Omega \rightarrow \mathbb{R}$ by

$$ E(x) = \log \frac{P(x_0)}{P(x)} \quad (P(x) > 0 \text{ by hypothesis}). $$
Proof of MRF with Positive Probability Are Gibbs Fields

**Candidate energy:** \( E(x) = \log \frac{P(x_0)}{P(x)} \) so that \( P(x) = P(x_0) e^{-E(x)} \).

**Goal:** Show that this energy \( E \) derives from a family of potentials on cliques \((U_C)_{C \in C}\): For all \( x \in \Omega \),

\[
E(x) = \sum_{C \in C} U_C(x).
\]

- Let \( x \) be a given configuration.
- For all subset \( A \subset V \), we define

\[
U_A(x) = \sum_{B \subset A} (-1)^{|A \setminus B|} E(xB).
\]

- By the Möbius inversion formula, we have, for all \( A \subset V \),

\[
E(xA) = \sum_{B \subset A} U_B(x).
\]

- In particular, taking \( A = V \), we get \( E(x) = \sum_{A \subset V} U_A(x) \) and thus

\[
P(x) = P(x_0) e^{-E(x)} = P(x_0) \exp \left( - \sum_{A \subset V} U_A(x) \right).
\]
Proof of MRF with Positive Probability Are Gibbs Fields

\[ P(x) = P(x_0) e^{-E(x)} = P(x_0) \exp \left( - \sum_{A \subset V} U_A(x) \right). \]

- \( P \) has nearly the targeted form!
- \( U_A(x) = \sum_{B \subset A} (-1)^{|A \setminus B|} E(xB) \) only depends on the values of \( x \) at the sites of \( A \).
- It only remains to show that \( U_A(x) = 0 \) if \( A \) is not a clique.
- Indeed, in this case the energy
  \[ E(x) = \sum_{A \subset V} U_A(x) = \sum_{C \in C} U_C(x) \]
  derives from a family of potential, that is, \( P \) is a Gibbs field.
- The proof of \( U_A(x) = 0 \) if \( A \) is not a clique will use the Markov property satisfied by \( P \)... (not used yet)
Proof of MRF with Positive Probability Are Gibbs Fields

Proof of $U_A(x) = 0$ if $A$ is not a clique:

- Let $A \subset V$ such that $A$ is not a clique, that is $A \notin C$.
- Then, $A$ contains two vertices $\alpha$ and $\beta$ that are not neighbors in the graph $(V, \mathcal{E})$.

\[
U_A(x) = \sum_{B \subseteq A \setminus \{\alpha, \beta\}} (-1)^{|A \setminus B|} E(xB) + \sum_{B \subseteq A \setminus \{\alpha, \beta\}} (-1)^{|A \setminus B|} E(xB) \\
+ \sum_{B \subseteq A \setminus \{\alpha, \beta\}} (-1)^{|A \setminus B|} E(xB) + \sum_{B \subseteq A \setminus \{\alpha, \beta\}} (-1)^{|A \setminus B|} E(xB) \\
= \sum_{B \subseteq A \setminus \{\alpha, \beta\}} (-1)^{|A \setminus B|} E(x(B \cup \{\alpha, \beta\})) + \sum_{B \subseteq A \setminus \{\alpha, \beta\}} (-1)^{|A \setminus B|} E(xB) \\
- \sum_{B \subseteq A \setminus \{\alpha, \beta\}} (-1)^{|A \setminus B|} E(x(B \cup \{\alpha\})) - \sum_{B \subseteq A \setminus \{\alpha, \beta\}} (-1)^{|A \setminus B|} E(x(B \cup \{\beta\})) \\
= \sum_{B \subseteq A \setminus \{\alpha, \beta\}} (-1)^{|A \setminus B|} \left( E(x(B \cup \{\alpha, \beta\})) + E(xB) - E(x(B \cup \{\alpha\})) - E(x(B \cup \{\beta\})) \right) \cdot \left( \begin{array}{c} E(x(B \cup \{\alpha, \beta\})) + E(xB) - E(x(B \cup \{\alpha\})) - E(x(B \cup \{\beta\})) \\ \end{array} \right) .
\]

Let us show that it is $= 0$. 

\[
\]
Proof of MRF with Positive Probability Are Gibbs Fields

Proof of $U_A(x) = 0$ if $A$ is not a clique:

**Goal:**

$$E(x(B \cup \{\alpha, \beta\})) + E(xB) - E(x(B \cup \{\alpha\})) - E(x(B \cup \{\beta\})) = 0$$

Since

$$E(x) = \log \frac{P(x_0)}{P(x)}$$

$$E(x(B \cup \{\alpha, \beta\})) + E(xB) - E(x(B \cup \{\alpha\})) - E(x(B \cup \{\beta\}))$$

$$= \log \frac{P(x(B \cup \{\alpha\}))P(x(B \cup \{\beta\}))}{P(x(B \cup \{\alpha, \beta\}))P(xB)}.$$

► Let us use the **conditional independence** (eq. to Markov property) with the partition: $V_1 = \{\alpha\}$, $V_2 = V \setminus \{\alpha, \beta\}$, and $V_3 = \{\beta\}$.

► Remark that the four states $x(B \cup \{\alpha\})$, $x(B \cup \{\beta\})$, $x(B \cup \{\alpha, \beta\})$, and $xB$ have the same restriction $x_2$ to $V_2 = V \setminus \{\alpha, \beta\}$.

$$P(x(B \cup \{\alpha, \beta\})) = P(x_\alpha = x_\alpha, x_\beta = x_\beta, x_2 = x_2)$$

$$= P(x_\alpha = x_\alpha, x_\beta = x_\beta | x_2 = x_2)P(x_2 = x_2)$$

$$= P(x_\alpha = x_\alpha | x_2 = x_2)P(x_\beta = x_\beta | x_2 = x_2)P(x_2 = x_2).$$

Similarly

$$P(xB) = P(x_\alpha = \lambda_0 | x_2 = x_2)P(x_\beta = \lambda_0 | x_2 = x_2)P(x_2 = x_2)$$

$$P(x(B \cup \{\alpha\})) = P(x_\alpha = x_\alpha | x_2 = x_2)P(x_\beta = \lambda_0 | x_2 = x_2)P(x_2 = x_2)$$

$$P(x(B \cup \{\beta\})) = P(x_\alpha = \lambda_0 | x_2 = x_2)P(x_\beta = x_\beta | x_2 = x_2)P(x_2 = x_2)$$
Proof of $U_A(x) = 0$ if $A$ is not a clique:

**Goal:** $E(x(B \cup \{\alpha, \beta\})) + E(xB) - E(x(B \cup \{\alpha\})) - E(x(B \cup \{\beta\})) = 0$

Hence,

$$\frac{P(x(B \cup \{\alpha\}))P(x(B \cup \{\beta\}))}{P(x(B \cup \{\alpha, \beta\}))P(xB)} = 1$$

which shows that

$$E(x(B \cup \{\alpha, \beta\})) + E(xB) - E(x(B \cup \{\alpha\})) - E(x(B \cup \{\beta\}))$$

$$= \log \frac{P(x(B \cup \{\alpha\}))P(x(B \cup \{\beta\}))}{P(x(B \cup \{\alpha, \beta\}))P(xB)} = 0$$

and thus $U_A(x) = 0$. 

Proof of $U_A(x) = 0$ if $A$ is not a clique:
Outline

The Ising Model

Markov Random Fields and Gibbs Fields

Sampling Gibbs Fields
Considered Problems

Gibbs fields pose several difficult computational problems

1. Compute samples from this distribution,
2. Compute the mode, the most probable state $x = \text{argmax } P(x)$ (if it is unique...),
3. Compute the marginal distribution on each component $\mathcal{X}_\alpha$ for some fixed $\alpha \in V$.

- The simplest method to sample from a Gibbs field is to use a type of Monte Carlo Markov Chain (MCMC) known as the **Metropolis Algorithm**.
- The main idea of Metropolis algorithms are to let evolve a Markov chain taking values in the space of configurations $\Omega = \Lambda^V$.
- Running a sampler for a reasonable length of time enables one to estimate the marginals on each component $\mathcal{X}_\alpha$ or the mean of other random functions of the state.

- **Simulated annealing**: Samples at very low temperatures will cluster around the mode of the field. But the Metropolis algorithm takes a very long time to give good samples if the temperature is too low. Simulated annealing consists in starting the Metropolis algorithm at sufficiently high temperatures and gradually lowering the temperature.
Recalls on Markov Chains

We refer to the textbook [Brémaud 1998] for a complete reference on Markov chains. The main reference for this section is [Mumford and Desolneux 2010].

Definition (Markov Chain)

Let $\Omega$ be a finite set of states. A sequence $(X_n)_{n \in \mathbb{N}}$ of random variables taking values in $\Omega$ is said to be a **Markov chain** if it satisfies the Markov condition: For all $n \geq 1$ and $x_0, x_1, \ldots, x_n \in \Omega$,

$$P(X_n = x_n | X_0 = x_0, X_1 = x, \ldots, X_{n-1} = x_{n-1}) = P(X_n = x_n | X_{n-1} = x_{n-1}),$$

whenever both conditional probabilities are well-defined, that is, whenever $P(X_0 = x_0, X_1 = x, \ldots, X_{n-1} = x_{n-1}) > 0$.

Moreover the Markov chain $(X_n)_{n \in \mathbb{N}}$ is said **homogeneous** if the probability does not depend on $n$, that is, for all $n \geq 1$, and $x, y \in \Omega$,

$$P(X_n = y | X_{n-1} = x) = P(X_1 = y | X_0 = x).$$

In what follows Markov chains will always be assumed homogeneous. In this case, the **transition matrix** $Q$ of the chain is defined as the $|\Omega| \times |\Omega|$ matrix of all transition probabilities

$$Q(x, y) = q_{x \rightarrow y} = P(X_1 = y | X_0 = x) \quad [ = P(X_n = y | X_{n-1} = x), \ \forall n].$$
Recalls on Markov Chains

Proposition
For all $n \geq 1$, the matrix $Q^n = Q \times \cdots \times Q$ gives the law of $X_n$ given the initial state $X_0$, that is,

$$
P(X_n = y | X_0 = x) = Q^n(x, y).
$$

Proof.
By induction. This is true for $n = 1$. Suppose it is true for some $n \geq 1$. Then,

$$
P(X_{n+1} = y | X_0 = x) = \sum_{z \in \Omega} P(X_{n+1} = y, X_n = z | X_0 = x)
$$

$$
= \sum_{z \in \Omega} \frac{P(X_{n+1} = y, X_n = z, X_0 = x)}{P(X_0 = x)}
$$

$$
= \sum_{z \in \Omega} \frac{P(X_{n+1} = y, X_n = z, X_0 = x)}{P(X_n = z, X_0 = x)} \frac{P(X_n = z, X_0 = x)}{P(X_0 = x)}
$$

$$
= \sum_{z \in \Omega} P(X_{n+1} = y | X_n = z, X_0 = x) P(X_n = z | X_0 = x)
$$

$$
= \sum_{z \in \Omega} P(X_{n+1} = y | X_n = z) Q^n(x, z)
$$

$$
= \sum_{z \in \Omega} Q^n(x, z) Q(z, y) = Q^{n+1}(x, y).
$$
Recalls on Markov Chains

Definition
We say that the Markov chain is **irreducible** if for all \( x, y \in \Omega \), there exists \( n \geq 0 \) such that \( Q^n(x, y) > 0 \).

- If a Markov chain is irreducible then every pair of states can be “connected” by the chain.
- Some Markov chains can have periodic behavior in visiting in a prescribed order all the states of \( \Omega \). This is generally a behavior that is best avoided.

Definition
We say that a Markov chain is **aperiodic** if for all \( x \in \Omega \) the greatest common divisor of all the \( n \geq 1 \) such that \( Q^n(x, x) > 0 \) is 1.
Equilibrium probability distribution

Definition
A probability distribution $\Pi$ on $\Omega$ is an equilibrium probability distribution of $Q$ if
\[
\forall y \in \Omega, \quad \sum_{x \in \Omega} \Pi(x)Q(x, y) = \Pi(y).
\]

Interpretation: If the initial state $X_0$ follows the distribution $\Pi$, then all the random variables $X_n$ also follow the distribution $\Pi$:
\[
P(X_1 = y) = \sum_{x \in \Omega} P(X_1 = y | X_0 = x)P(X_0 = x) = \sum_{x \in \Omega} Q(x, y)\Pi(x) = \Pi(y).
\]

Theorem
If $Q$ is irreducible, then there exists a unique equilibrium probability distribution $\Pi$ for $Q$. If moreover $Q$ is aperiodic, then
\[
\forall x, y \in \Omega, \quad \lim_{n \to +\infty} Q^n(x, y) = \Pi(y).
\]

Consequence: If $X_0$ follows any distribution $P_0$, then, for all $y \in \Omega,$
\[
P(X_n = y) = \sum_{x \in \Omega} P(X_n = y | X_0 = x)P(X_0 = x)
\]
\[
= \sum_{x \in \Omega} Q^n(x, y)P_0(x) \xrightarrow{n \to +\infty} \sum_{x \in \Omega} \Pi(y)P_0(x) = \Pi(y).
\]
Equilibrium probability distribution

Consequence:

\[ \mathbb{P}(X_n = y) \xrightarrow{n \to +\infty} \Pi(y) \]

- Whatever the distribution of the initial state \( X_0 \), as \( n \) tends to \( +\infty \), the sequence \( (X_n)_{n \in \mathbb{N}} \) converges in distribution to \( \Pi \).

Approximate simulation of the distribution \( \Pi \):

1. Draw randomly an initial state \( x_0 \) (with some distribution \( P_0 \)).
2. For \( n = 1 \) to \( N \) (\( N \) “large”), compute \( x_n \) by simulating the discrete distribution

\[ \mathbb{P}(X_n | X_{n-1} = x_{n-1}) = Q(x_{n-1}, \cdot). \]

3. Return the random variate \( x_N \).

This procedure is called a Monte Carlo Markov Chain (MCMC) method and is at the heart of the Metropolis algorithm to simulate Gibbs fields.
Metropolis algorithm

Framework:

▶ $\Omega$ a finite state space ($\Omega = \Lambda^V$ in our case).

▶ Given any energy function $E : \Omega \to \mathbb{R}$, we define a probability distribution on $\Omega$ at all temperatures $T$ via the formula:

$$p_T(x) = \frac{1}{Z_T} e^{-\frac{1}{T}E(x)}, \text{ where } Z_T = \sum_{x \in \Omega} e^{-\frac{1}{T}E(x)}.$$  

▶ Idea: Simulate a Markov chain $(\mathcal{X}_n)_{n \geq 0}$, $\mathcal{X}_n \in \Omega$, such that the equilibrium probability distribution is $P_T$, so that, as $n$ tends to $+\infty$, the distribution of $\mathcal{X}_n$ will tend to $P_T$.

▶ We start with some simple Markov chain, given by a transition matrix $Q(x, y) = (q_{x \to y})_{(x, y) \in \Omega^2}$, $q_{x \to y}$ being the probability of the transition from a state $x \in \Omega$ to a state $y \in \Omega$.

▶ Hypotheses on $Q$:

- $Q$ is symmetric ($q_{x \to y} = q_{y \to x}$),
- $Q$ is irreducible,
- $Q$ is aperiodic.
Metropolis algorithm

Definition
The Metropolis transition matrix associated with $P_T$ and $Q$ is the matrix $P$ defined by

$$P(x, y) = p_{x \rightarrow y} = \begin{cases} 
q_{x \rightarrow y} & \text{if } x \neq y \text{ and } P_T(y) \geq P_T(x), \\
\frac{P_T(y)}{P_T(x)} q_{x \rightarrow y} & \text{if } x \neq y \text{ and } P_T(y) \leq P_T(x), \\
1 - \sum_{z \neq x} p_{x \rightarrow z} & \text{if } y = x.
\end{cases}$$

Theorem
The Metropolis transition matrix $P(x, y) = (p_{x \rightarrow y})$ is irreducible and aperiodic if $Q(x, y) = (q_{x \rightarrow y})$ is. In addition, the equilibrium probability distribution of $P(x, y) = (p_{x \rightarrow y})$ is $P_T$, that is,

$$\forall x, y \in \Omega, \lim_{n \to +\infty} P^n(x, y) = P_T(y).$$
Metropolis algorithm

**Proof:** Since $P_T(y) \neq 0$ for all $y \in \Omega$,

$$Q(x, y) \neq 0 \Rightarrow P(x, y) \neq 0,$$

and by induction,

$$Q^n(x, y) \neq 0 \Rightarrow P^n(x, y) \neq 0.$$

- As $Q$, $P$ is irreducible and aperiodic.
- In consequence, $P$ has a unique equilibrium probability distribution.

We now prove that there is “detailed balance”, which will imply that $P_T$ is the equilibrium probability distribution of $P$:

$$\forall x, y \in \Omega, \quad P_T(x) p_{x \rightarrow y} = P_T(y) p_{y \rightarrow x}.$$

Let $x, y \in \Omega$. If $x = y$, there is obvious. If $x \neq y$, we can suppose that $P_T(x) \geq P_T(y)$. Then,

$$P_T(x) p_{x \rightarrow y} = P_T(x) \frac{P_T(y)}{P_T(x)} q_{x \rightarrow y} = P_T(y) q_{x \rightarrow y} = P_T(y) p_{y \rightarrow x}.$$

Summing this equality over all $x$, one gets for all $y \in \Omega$,

$$\sum_{x \in \Omega} P_T(x) p_{x \rightarrow y} = P_T(y) \sum_{x \in \Omega} p_{y \rightarrow x} = P_T(y),$$

i.e. $P_T$ is the equilibrium probability distribution of the transition matrix $P$. 
Metropolis algorithm

Practical aspect: Choice of the transition matrix $Q$

- The uniform change satisfies the hypotheses

$$\forall x, y \in \Omega, \quad q_{x\rightarrow y} = \frac{1}{|\Omega|}.$$ 

However, this choice is not practical since one attempts at randomly draw a structured image from a white noise image.

- A more practical transition matrix is the one corresponding to the following procedure: one selects randomly a pixel/vertex $\alpha \in V$, and one draws uniformly a new value $\lambda \in \Lambda \setminus \{x_\alpha\}$ for this vertex. Hence,

$$q_{x\rightarrow y} = \begin{cases} 
\frac{1}{|V|(|\Lambda|-1)} & \text{if } x \text{ and } y \text{ differs on exactly one vertex/pixel,} \\
0 & \text{otherwise}.
\end{cases}$$

- One of the main problem of the Metropolis algorithm is that many proposed moves $x \rightarrow y$ will not be accepted.

- The Gibbs sampler attempts to reduce this problem.
Let $\Omega = \Lambda^V$ be the Gibbs state space.

The **Gibbs sampler** selects one site $\alpha \in V$ at random and then picks a new value $\lambda$ for $x_\alpha$ with probability distribution given by $P_T$, conditioned on the values of the state at all other vertices:

$$p_{x \rightarrow y} = \begin{cases} 
\frac{1}{|V|} P_T(x_\alpha = y_\alpha | x_{\Gamma \setminus \alpha} = x_{\Gamma \setminus \{\alpha\}}) & \text{if } y \text{ differs from } x \text{ only at } \alpha, \\
\sum_{\alpha \in V} \frac{1}{|V|} P_T(x_\alpha = x_\alpha | x_{\Gamma \setminus \alpha}) & \text{if } y = x, \\
0 & \text{otherwise.}
\end{cases}$$

The transition matrix $P$ satisfies the same property as the one of the Metropolis algorithm (**Exercice**).
Bibliographic references I
