Machine Learning in Computer Vision: Markov Random Fields – Part III

Oren Freifeld
Computer Science, Ben-Gurion University

22/3/18, 28/3/18, and 11/4/2018
1. Image Restoration as a Generic Example

2. Ising Model

3. Introduction to MCMC and Gibbs Sampling
   - Gibbs Sampling
   - MCMC

4. Generalizations
Image Restoration

- Consider an $n \times n$ image
- Pixels define a regular 2D lattice
- $x = (x_s)_{s \in S}$, a latent ("clean" / "uncorrupted") image
- $y = (y_s)_{s \in S}$, an observed ("corrupted-by-noise") image
- Problem: Recover $x$ from $y$. 
The Bayesian approach is based on

\[ p(x|y) = \frac{p(x, y)}{p(y)} = \frac{p(y|x)p(x)}{p(y)} \propto p(y|x)p(x) \]

- \( p(x|y) \): posterior
- \( p(y|x) \): likelihood
- \( p(y) \): evidence
- \( p(x) \): prior
Consider Binary Latent Images with an MRF prior

- \( x_s \in \{-1, 1\} \) (binary, with -1 instead of 0); i.e., \( x \in \{-1, 1\}^{|S|} \)

- Prior model:

\[
p(x) \propto \prod_{c \in C} H_c(x_c) \tag{1}
\]

- Observation/likelihood model 1 (\( \mathbb{R} \)-valued, additive):

\[
\mathbb{R} \ni y_s = x_s + n_s \quad s \in S \quad \mathbb{R} \ni n_s \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)
\]

- Observation/likelihood model 2 (binary, multiplicative):

\[
\{-1, 1\} \ni y_s = x_s n_s \quad s \in S \quad n_s \in \{-1, 1\} \quad (n_s + 1)/2 \overset{iid}{\sim} \text{Bernoulli}(\theta)
\]

- In both cases, the overall likelihood is:

\[
p(y|x) = \prod_{s \in S} p(y_s|x) = \prod_{s \in S} p(y_s|x_s)
\]

The first "\( = \)" : the \( y_s \)'s are conditionally independent given \( x \).

The second "\( = \)" is stronger: \( y_s \perp \perp (x_s, y_s)\mid x_s \ \forall s \in S \).

\( iid \) = independent and identically distributed
Posterior

Case 1:

\[ p(x|y) = \frac{\prod_{s \in S} \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left( -\frac{(y_s - x_s)^2}{2\sigma^2} \right) \prod_{c \in C} H_c(x_c)}{\sum_{(x'_s)_{s \in S} \in \{-1,1\}^{|S|}} \left[ \prod_{s \in S} \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left( -\frac{(y_s - x'_s)^2}{2\sigma^2} \right) \right] \prod_{c \in C} H_c(x'_c)} \]

\[ \propto \left[ \prod_{s \in S} \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left( -\frac{(y_s - x_s)^2}{2\sigma^2} \right) \right] \prod_{c \in C} H_c(x_c) \]

\[ \propto \prod_{c \in C} F_c(x_c) \]

(singletons – w.r.t. \( G_A \) – absorbed into the \( H_c \)'s; renamed the resulting clique functions \( F_c \)'s where \( F_c(x_c) = F_c(x_c, y) \))
Posterior

Case 2:

\[ p(x|y) \propto (1 - \theta)|\{s \in S: y_s = -x_s\}|\theta|\{s \in S: y_s = x_s\}| \prod_{c \in C} H_c(x_c) \]

\[ = \prod_{c \in C} F_c(x_c) \]

(singletons – w.r.t. \( G_A \) – absorbed into the \( H_c \)'s; renamed the resulting clique functions \( F_c \)'s where \( F_c(x_c) = F_c(x_c, y) \))
Digression: What if $y_s$ Depends on more than $x_s$?

E.g., $y$ obtained from $x$ by convolving it with a $3 \times 3$ filter. We get products of $p(y_s|3 \times 3 \text{ n’hood of } x_s)$

$\Rightarrow \ln p(x|y)$, each $3 \times 3$ block is a clique, with (partial) overlaps between nearby cliques.
What We Want

- Sample from $p(x|y)$, compute posterior expectations (e.g., $E(x|y)$ or $E(g(x)|y)$ for some function $g$), and posterior arg max$_x p(x|y)$.
- In fact, if we can sample somehow, we can use this for the other tasks as well; for example (and regardless of MRF’s):

$$
(x^i_s)_{s \in S} \overset{iid}{\sim} p((x_s)_{s \in S}|y) \quad p(x|y) \\
i = 1, 2, \ldots, N
$$

$$
\Rightarrow \frac{1}{N} \sum_{i=1}^{N} g((x^i_s)_{s \in S}) \xrightarrow{N \to \infty} E(g((X_s)_{s \in S})|Y = y) \\
by \text{LLN} \quad E(g(X|Y = y))
$$

$(x^i$ is a sample of an entire image)

Remark: usually $E(X|Y = y)$ is in $[-1, 1]^{|S|}$ (as opposed to $\{-1, 1\}^{|S|}$)

LLN: Law of Large Numbers
The Ising Model Prior

\[ p(x) = \frac{1}{Z} \exp \left( \beta \sum_{s \sim t} x_s x_t \right) \quad (s \sim t \text{ means } s \text{ is a neighbor of } t) \]

Samples from \( p(x) \) with three different betas where, from left to right, \( \beta_1 < \beta_2 < \beta_3 \)

---

Figure from Winkler’s monograph, “Image Analysis: Random Fields and Markov Chain Monte Carlo Methods”
If $\beta = \frac{1}{1.8}$ and $y_s = x_s + n_s$ where $n_s \overset{\text{iid}}{\sim} \mathcal{N}(0, 4)$, $s \in S$ then

$$p(x|y) \propto \exp \left( \frac{1}{1.8} \left( \sum_{s\sim t} x_s x_t \right) - \frac{1}{8} \sum_s (y_s - x_s)^2 \right)$$
Left: the true $x$ where $x \sim p(x)$ (using the Ising Model)
Right: $y$ – the noisy image.
Left: \( x \sim p(x) \) (using the Ising Model)
Right: a posterior sample, \( x \sim p(x|y) \)
Left: $x \sim p(x)$ (using the Ising Model)
Right: the maximum-a-posteriori (MAP) solution, $\arg\max_x p(x|y)$
Interpretations of the Ising Model

- WLOG, let $\beta = 1$.

\[
\frac{\exp \left( \sum_{s \sim t} x_s x_t \right)}{\sum_{x'} \exp \left( \sum_{s \sim t} x'_s x'_t \right)} = \frac{\exp \left( \sum_{s \sim t} x_s x_t + c \right)}{\sum_{x'} \exp \left( \sum_{s \sim t} x'_s x'_t + c \right)} \quad \forall c \in \mathbb{R}
\]  

(2)

\[
\Rightarrow p(x) \propto \frac{\exp \left( \sum_{s \sim t} x_s x_t + c \right)}{\sum_{x'} \exp \left( \sum_{s \sim t} x'_s x'_t + c \right)}
\]  

(3)

- Take $c = |S|$ (# pixels) $\Rightarrow$

\[
\sum_{s \sim t} x_s x_t + c =
\]

# like n’brs pairs - # unlike n’brs pairs + $|S| = 2 \times$ # like n’brs pairs

$\Rightarrow p(x) \propto \exp(2 \times$ # like n’brs pairs)

- Take $c = -|S|$ $\Rightarrow$

# like n’brs pairs - # unlike n’brs pairs - $|S| = -2 \times$ # unlike n’brs pairs

$= -2$ times of boundary length

$\Rightarrow p(x) \propto \exp(-2$ times of boundary length)
**When Sampling from the Conditionals is Easy**  
*Regardless of MRFs*

Assume the following setting:

- $X$, $Y$ and $Z$: 3 random vectors (of possibly-different dimensions) with a joint pdf/pmf $p(x, y, z)$
- Want to sample $x, y, z \sim p(x, y, z)$ – but it's hard or don’t know how.
- Can relatively easily sample from the conditionals:

$$p(x|y, z)$$
$$p(y|x, z)$$
$$p(z|x, y)$$
Gibbs Sampling [Geman and Geman, 1984]

- The algorithm:
  1. \( x^{[0]}, y^{[0]}, z^{[0]} \sim g(x, y, z) \) using some easy-to-sample-from \( g(x, y, z) \) of the same support as \( p(x, y, z) \) (e.g., take \( g(x, y, z) = g(x)g(y)g(z) \)).
  2. Iterate:
     - \( x^{[i]} \sim p(x|y^{[i-1]}, z^{[i-1]}) \)
     - \( y^{[i]} \sim p(y|x^{[i]}, z^{[i-1]}) \)
     - \( z^{[i]} \sim p(z|x^{[i]}, y^{[i]}) \)

\[
p(x^{[n]}, y^{[n]}, z^{[n]}) \xrightarrow{n \to \infty} p(x, y, z) \text{ (in some sense)}
\]

- The resulting sequence is clearly a Markov Chain:

\[
p(x^{[i]}, y^{[i]}, z^{[i]}|x^{[0:(i-1)]}, y^{[0:(i-1)]}, z^{[0:(i-1)]}) = p(x^{[i]}, y^{[i]}, z^{[i]}|x^{[i-1]}, y^{[i-1]}, z^{[i-1]})
\]

- Gibbs sampling is a particular case of MCMC
Gibbs Sampling [Geman and Geman, 1984]

It is OK even if we know $p$ only up to a multiplicative constant:

- Know $g(x, y, z)$, a non-normalized nonnegative function, where

$$p(x, y, z) = \frac{1}{Z} g(x, y, z)$$

$$Z = \int g(x, y, z) \, dx \, dy \, dz$$

but can’t (or don’t want to) compute the integral.

$$p(x | y, z) = \frac{p(x, y, z)}{p(y, z)} = \frac{p(x, y, z)}{\int p(x', y, z) \, dx'}$$

$$= \frac{1}{Z} \frac{g(x, y, z)}{\int \frac{1}{Z} g(x', y, z) \, dx'} = \frac{g(x, y, z)}{\int g(x', y, z) \, dx'}$$

or, in the discrete case:

$$p(x | y, z) = \frac{g(x, y, z)}{\sum_{x'} g(x', y, z)}$$
Gibbs Sampling [Geman and Geman, 1984]

More generally: want \((x_1, \ldots, x_n) \sim p(x_1, \ldots, x_n)\).

Gibbs sampling \(n\) RV’s:

1. \(x_1^{[0]}, x_2^{[0]}, \ldots, x_n^{[0]} \sim g(x_1, \ldots, x_n)\) using some easy-to-sample-from \(g(x_1, \ldots, x_n)\) of the same support as \(p(x_1, \ldots, x_n)\)

2. Iterate:
   - Update \(x_1\) by sampling \(x_1\) given the rest
   - Update \(x_2\) by sampling \(x_2\) given the rest
   - 
   - 
   - Update \(x_n\) by sampling \(x_n\) given the rest

One such iteration over all \(n\) variables is called a sweep.
Motivation

Assume

- \( p \) is “local” (MRF with modest n’hood size)
- \( p(y|x) \) is “local”

\[ p(x|y) \] is MRF with “local” n’hood structure.

But: in an \( n \times n \) lattice structure (e.g., images), max boundary is of order \( n \) – can’t use Dynamic Programming. So try MCMC.

- Gibbs sampling [Geman and Geman, 1984] is a particular case of MCMC methods.
- In principle, Gibbs sampling is applicable in general distributions, not just in Gibbs distributions (i.e., MRFs), but it is particularly easy to do Gibbs sampling in Gibbs distributions.
Gibbs Sampling in MRFs [Geman and Geman, 1984]

- Pick a (possibly-random) site-visitation scheme, and then, when visiting \( s \), sample
  \[
  x_s \sim p(x_s | x_{\neq s}) \overset{\text{MRF}}{=} p(x_s | x_{\eta_s})
  \]
- \( p(x_s | x_{\eta_s}) \), which involves only the local neighborhood, is usually easy to sample from.
- It is ok if know \( p \) only up to a multiplicative constant (which is often the case with MRFs):
  \[
  p(x) \propto \prod_{c \in C} F_c(x_c)
  \]
- Asymptotically great, but takes a long time if the graph is too large/complicated
- Often, particularly for images, it can be massively parallelized (but even then this can take a long time)
Gibbs Sampling in the Ising Model: Sampling from $p(x)$

\[ p(x) \propto \exp \left( \beta \sum_{s \sim t} x_s x_t \right) \]

\[ p(x_s | s x) = \frac{p(x_s, s x)}{p(s x)} = \frac{p(x)}{p(s x)} = \frac{1}{Z} \exp \left( \beta \sum_{s \sim t} x_s x_t \right) \]

\[ \sum_{x'} \frac{1}{Z} \exp \left( \beta \sum_{s \sim t} x'_s x_t \right) \]

\[ = \frac{\exp \left( \beta \sum_{t:t \in \eta_s} x_s x_t \right)}{\sum_{x'} \exp \left( \beta \sum_{t:t \in \eta_s} x'_s x_t \right)} \]

\[ = \frac{\exp \left( \beta \sum_{t:t \in \eta_s} x_s x_t \right)}{\exp \left( -\beta \sum_{t:t \in \eta_s} x_t \right) + \exp \left( \beta \sum_{t:t \in \eta_s} x_t \right)} \]

\[ \Rightarrow \]

\[ \begin{cases} 
  p(x_s = 1 | s x) = \frac{\exp \left( \beta \sum_{t:t \in \eta_s} x_t \right)}{\exp \left( -\beta \sum_{t:t \in \eta_s} x_t \right) + \exp \left( \beta \sum_{t:t \in \eta_s} x_t \right)} \propto \exp \left( \beta \sum_{t:t \in \eta_s} x_t \right) \\
  p(x_s = -1 | s x) = \frac{\exp \left( -\beta \sum_{t:t \in \eta_s} x_t \right)}{\exp \left( -\beta \sum_{t:t \in \eta_s} x_t \right) + \exp \left( \beta \sum_{t:t \in \eta_s} x_t \right)} \propto -\exp \left( \beta \sum_{t:t \in \eta_s} x_t \right)
\end{cases} \]
Gibbs Sampling in the Ising Model: Sampling from $p(x|y)$

iid Additive Gaussian Noise

\[ p(x|y) \propto \exp \left( \beta \left( \sum_{s \sim t} x_s x_t \right) - \frac{1}{2\sigma^2} \sum_s (y_s - x_s)^2 \right) \]

\[ p(x_s|x, y) = \frac{p(x_s, s|x, y)}{p(s|x, y)} = \frac{p(x|y)}{p(s|x, y)} \]

\[ = \frac{\exp \left( \beta \left( \sum_{s \sim t} x_s x_t \right) - \frac{1}{2\sigma^2} \sum_s (y_s - x_s)^2 \right)}{\sum_{x_s'} \exp \left( \beta \left( \sum_{s \sim t} x_s' x_t \right) - \frac{1}{2\sigma^2} \sum_s (y_s - x_s')^2 \right)} \]

\[ \propto \exp \left( \beta \left( \sum_{t: t \in \eta_s} x_s x_t \right) - \frac{1}{2\sigma^2} (y_s - x_s)^2 \right) \]

\[ p(x_s = 1|x, y) \propto \exp \left( \beta \left( \sum_{t: t \in \eta_s} x_t \right) - \frac{1}{2\sigma^2} (y_s - 1)^2 \right) \]

\[ p(x_s = -1|x, y) \propto \exp \left( \beta \left( - \sum_{t: t \in \eta_s} x_t \right) - \frac{1}{2\sigma^2} (y_s + 1)^2 \right) \]

Nothing more than flipping a biased coin.
Gibbs Sampling in the Ising Model: Sampling from $p(x|y)$

*Multiplicative flips (AKA binary symmetric channel)*

\[
p(x|y) \propto \exp \left( \beta \sum_{s \sim t} x_s x_t \right) \theta^{|\{s \in S : y_s = x_s\}|} (1 - \theta)^{|\{s \in S : y_s = -x_s\}|}
\]

\[
p(x_s|x, y) = \frac{p(x_s, x|y)}{p(x|x|y)} = \frac{p(x|y)}{p(x|x|y)}
\]

\[
\propto \exp \left( \beta \sum_{t : t \in \eta_s} x_t \right) \theta^{1_{y_s = x_s}} (1 - \theta)^{1_{y_s = -x_s}}
\]

\[\Rightarrow \]

\[
p(x_s = 1|x, y) \propto \exp \left( \beta \sum_{t : t \in \eta_s} x_t \right) \theta^{1_{y_s = 1}} (1 - \theta)^{1_{y_s = -1}}
\]

\[
p(x_s = -1|x, y) \propto \exp \left( -\beta \sum_{t : t \in \eta_s} x_t \right) \theta^{1_{y_s = -1}} (1 - \theta)^{1_{y_s = 1}}
\]

Again, this is just flipping a biased coin.
Markov Chain Monte Carlo (MCMC)

Regardless of MRFs

- Let $p$ be the pdf/pmf of interest, known as the target distribution.
- Find a Markov Chain, $X(0), X(1), \ldots$ with $X(t) \in \mathcal{R}^{|S|}$ such that:
  - $\Pr(X(t) = x) \to p(x)$ as $t \to \infty$ ("sampling")
  - or
  - $\Pr(X(t) \in M) \to 1$ as $t \to \infty$ where
    \[ M = \{ x : p(x) = \max_{x'} p(x') \} \]
    ("annealing")
  - or
  - \[
  \frac{1}{T} \sum_{t=1}^{T} H(X(t)) \xrightarrow{t \to \infty} E_p H(X)
  \]
    ("ergodicity")

for some function of interest $H$
Stochastic matrix

Definition

A square matrix $P$ of nonnegative values is called a stochastic matrix if each of its row sums to 1.

- Remark: It is called doubly stochastic if both $P$ and $P^T$ are stochastic matrices.
- Remark: a stochastic vector is a vector of nonnegative values whose sum is 1.
- This terminology can be confusing – we often use “random” and “stochastic” interchangeably, but here neither a stochastic matrix nor a stochastic vector are “random”.

Markov Chain Monte Carlo (MCMC) – Regardless of MRFs

- For simplicity, assume finite-state space (but more generally, MCMC is also applicable for the countable or continuous settings)
- The MC is said to be stationary if the transition probability
  \[ P_{x,y} \triangleq \Pr(X(t) = y | X(t-1) = x) \]
  is independent of \( t \) \( \forall x, y \in \mathcal{R}^{\mathcal{S}} \)
- The transition probability matrix is
  \[ P = \{ P_{x,y} \} \in \mathbb{R}^{\mathcal{R}^{\mathcal{S}} \times \mathcal{R}^{\mathcal{S}}} \]
  (this might be huge)
- \( P \) is a stochastic matrix.
- Consider pmf’s over \( \mathcal{R}^{\mathcal{S}} \) as row vectors of length \( \mathcal{R}^{\mathcal{S}} \). Such a vector has nonnegative entries summing to 1.
- Let \( \pi(t) \) be the pmf of \( X(t) \) \( \Rightarrow \pi(t+1) = \pi(t)P \) is the pmf of \( X(t+1) \):
  \[ \pi_i(t+1) = \sum_{j=1}^{\mathcal{R}^{\mathcal{S}}} \pi_j(t)P_{x,y}(j, i) = \sum_{j=1}^{\mathcal{R}^{\mathcal{S}}} \Pr(X(t+1) = i, X(t) = j) \]
Markov Chain Monte Carlo (MCMC)

Regardless of MRFs

- \( \pi(t + 2) = \pi(t + 1)P = \pi(t)PP = \pi(t)P^2 \)

- more generally \( \pi(t + n) = \pi(t)P^n \)

- If \( \pi P = \pi \), i.e., \( \pi \) is a left eigenvector of \( P \) with eigenvalue 1, then \( \pi \) is called an equilibrium distribution of the MC.

- If \( \pi(t) \xrightarrow{t \to \infty} \pi \) then \( \pi \) is an equilibrium distribution, called the stationary distribution of the MC.

If, in addition, \( \exists \lim_{t \to \infty} P^t \), then \( P^t \xrightarrow{t \to \infty} \begin{bmatrix} \pi \\ \pi \\ \vdots \\ \pi \end{bmatrix} \) (i.e., identical rows)

This is consistent with the fact that regardless what \( \pi(0) \) is,

\( \pi(t) = \pi(0)P^t \xrightarrow{t \to \infty} \pi. \)

- Not every MC has an equilibrium distribution.

- MCMC methods are based on finding an MC whose equilibrium distribution is the target distribution.
**Theorem**

If \((X(t))_{t \geq 0}\) is an MC on \(\mathcal{R}^S\) and if

1. \(\mathcal{R}^{|S|}\) is connected under \(\{P_{x,y}\}\), namely, \(\forall x, y \in \mathcal{R}^{|S|}\),
   \(\exists x(0), x(1), \ldots, x(t)\) such that \(x(0) = x, x(t) = y\) and \(P_{x(n+1),x(n)} > 0\)
   with \(n = 0, 1, \ldots, t - 1\) (in MC terminology, there is one communication class)

2. \(P_{x,x} > 0\) for some \(x \in \mathcal{R}^{|S|}\) (or more generally, in terms of MC terminology, \(P_{x,y}\) is aperiodic)

Then:

- \(\exists\) a unique \(\tilde{p}\) on \(\mathcal{R}^{|S|}\) such that \(\sum_x \tilde{p}P_{x,y} = \tilde{p}(y) \forall y \in \mathcal{R}^{|S|}\)
  (\(\tilde{p}\) = “equilibrium distribution”)

- \(\Pr(X(t) = x | X(0) = x_0) \xrightarrow{t \to \infty} \tilde{p}(x)\), indep. of \(x_0\) (the initial state)

- \(\frac{1}{T} \sum_{i=1}^T H(X(t)) \xrightarrow{T \to \infty} \sum_x \tilde{p}(x)H(x) = E_{\tilde{p}}H(X)\) with \(H : \mathcal{R}^{|S|} \to \mathbb{R}\)

But what we want is to sample from \(p\), not \(\tilde{p}\). Trick: find \(P_{x,y}\) such that \(p = \tilde{p}\).
Fact (Detailed Balance)

If for some $\pi$, probability on $\mathcal{R}^{|S|}$,

$$
\pi(x)P_{x,y} = \pi(y)P_{y,x} \quad \forall x, y \in \mathcal{R}^{|S|}
$$

then $\sum_x \pi(x)P_{x,y} = \pi(y)$, i.e., $\pi$ is the equilibrium distribution.

Proof.

$$
\sum_x \pi(x)P_{x,y} = \sum_x \pi(y)P_{y,x} = \pi(y)\sum_x \Pr(X(t+1) = x | X(t) = y) = \pi(y)
$$

Remark

Note that $\pi(x)P_{x,y} = \pi(y)P_{y,x}$ is just a short way for writing that

$$
\Pr(X(t) = x, X(t + 1) = y) = \Pr(X(t) = y, X(t + 1) = x)
$$
Putting the theorem and the fact together: If $P_{x,y}$ satisfies the condition of the theorem and $\pi = p$ satisfies detailed balance, then $p$ is the unique equilibrium probability of the chain and (2) and (3) from the theorem hold with $\tilde{p} = p$. 
Detailed balance is also called reversibility since it implies

$$
\Pr(X(t) = y | X(t+1) = x) = \underbrace{\Pr(X(t+1) = y | X(t) = x)}_{P_{x,y} \text{ (by definition)}};
$$

$$\pi(x)$$

i.e., the backward dynamics is the same as the forward dynamics.

**Proof.**

$$
\Pr(X(t) = y | X(t+1) = x) = \frac{\Pr(X(t) = y, X(t+1) = x)}{\Pr(X(t+1) = x)}
$$

$$\pi(x)$$

$$\pi(y)$$

$$
= \frac{\Pr(X(t+1) = x | X(t) = y) \Pr(X(t) = y)}{\pi(x)} = \frac{P_{y,x} \pi(y)}{\pi(x)} \equiv P_{x,y}
$$
Suppose the visiting schedule is done by, $q_v$, a probability on $S$. Given $X(t) = x \in \mathcal{R}^{|S|}$:

- Choose a site $v$ using $v \sim q_v$.
- Change $x_v$ to a sample from $p(x_v|x_{\eta_v})$

Checking the conditions of the theorem:

- We can get from any state $x$ to any state $y$ by changing one site at the time. So the state space is connected.
- $P_{x,x} > 0$ for some $x \in \mathcal{R}^{|S|}$? Yes. In fact, it holds for every $x$.
- Gibbs Sampling also satisfies Detailed Balance (see next slide).
Gibbs Sampling satisfies Detailed Balance.

\[ P_{x,y} = \Pr(X(t+1) = y | X(t) = x) = \sum_{v \in S} \Pr(X(t+1) = y | X(t) = x, V = v) q_v \]

where \( \Pr(X(t+1) = y | X(t) = x, V = v) \) is \( p(y_v | v x) \) if \( v y = v x \) and 0 otherwise. Thus:

\[
P_{x,y} = \sum_{v \in S} 1_{v y = v x} q_v p(y_v | v x)
\]

\[
\Rightarrow p(x) P_{x,y} = \sum_{v \in S} 1_{v y = v x} q_v p(y_v | v x) p(x_v | v x) p(v x)
\]

Similarly:

\[
p(y) P_{y,x} = \sum_{v \in S} 1_{v x = v y} q_v p(x_v | v y) p(y_v | v y) p(v y)
\]

\[
= \sum_{v \in S} 1_{v x = v y} q_v p(x_v | v x) p(y_v | v x) p(v x)
\]

\[
= \sum_{v \in S} 1_{v x = v y} q_v p(y_v | v x) p(x_v | v x) p(v x)
\]

\[
\Rightarrow p(y) P_{y,x} = p(x) P_{x,y}, \text{ i.e., detailed balance is achieved.}
\]
Remarks on Gibbs Sampling

- If the visitation schedule is deterministic we still have
  \[ \Pr(X(t) = x | X(0) = x(0)) \xrightarrow{t \to \infty} p(x) \] provided that every site is visited infinitely often.
- Important to observe that we don’t need \( Z \).
- Recall: 
  \[ p(x) = \frac{1}{Z} \exp \left( - \sum_{c \in C} E_c(x_c) \right) \]
  (similarly if working with \( p(x|y) \)) Write
  \[ p(x) = p_T(x) \bigg|_{T=1} \triangleq \frac{1}{Z_T} \exp \left( - \frac{1}{T} \sum_{c \in C} E_c(x_c) \right) \bigg|_{T=1} \]

Fact (Simulated Annealing)

\[ p_T(x) \xrightarrow{T \downarrow 0} \text{probability concentrated on } \{ x : x = \arg \max_{x'} p(x') \} \]

as long as \( T \) doesn’t decay “too fast”.
- Similar results if working with \( p(x|y) \) instead of \( p(x) \)
**Modeling: Beyond the Ising Model**

- $x_s \in \{0, 1, \ldots, m\}$. The Potts model is

$$p(x) \propto \exp \left( \gamma \sum_{s \sim t} 1_{x_s \neq x_t} \right)$$

- A modification in case the ordering of the labels matters:

$$p(x) \propto \exp \left( \gamma \sum_{s \sim t} (x_s - x_t)^2 \right)$$

(can also replace the $\ell_2$ loss with a robust error function)

- Gaussian MRF: $x \sim \mathcal{N}(\mu, Q^{-1})$ where $Q$ is SPD and sparse.

- More complicated models and/or higher-order cliques (see examples in the presentation “MRFs, part 4”)

Inference: Beyond Dynamic Programming and Gibbs Sampling

There are many other approaches for searching for the argmax. For example:

- Graph-cut methods
- Belief propagation and loopy belief propagation
- Mean-field approximations
- ...
Applications: Beyond Image Restoration

- Inpainting ($y$ has missing pixels)
- Image segmentation
- Spatial coherence in various problems such as optical-flow estimation, depth-estimation, stereo, etc.
- Pose estimation
- Texture synthesis
- Interactions between superpixels (irregular grid)
- HMM-like models are useful in CV applications such as tracking
- Many other applications
Learning MRFs

- MRFs often have parameters (e.g., $\beta$ in the Ising model, or $Q$ in a GMRF) that can/should be learned; we won’t get into this too much, but we will touch upon some examples in the next presentation.
- We can also learn the structure of the graph (i.e., the existence, or lack thereof, edges in the graph); this is called structural inference.
Books (see course website for details)

- Prince
- Szeliski
- Markov random fields for vision and image processing; editors: Blake and Rother (nice mix of tutorials and applications)
- Winkler’s Image Analysis, Random Fields and Markov Chain Monte Carlo (mathematically advanced, focused on MCMC)