Markov random fields and energy-based models

A: Markov random fields

Image restoration

Geman, Geman (1984). *Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images.*

Restoring degraded images:

- Images from spaces
- Blurry photos of license plates or crime scenes
- Noise in x-rays

Simplest model of degradation:

- Original $m \times n$ image $X(i,j)$
- Degraded version $Y(i,j)$ given by $Y = H \star X + Z$, i.e.,

$$Y(i,j) = \sum_{k,l} X(k,l)H(i-k, j-l) + Z(i,j)$$

where H is (known) shift-invariant blurring process, Z is Gaussian noise

Examples of blurring processes

- Original $m \times n$ image $X(i,j)$
- Degraded version $Y(i,j)$ given by $Y = H \star X + Z$, i.e.,

$$Y(i,j) = \sum_{k,l} X(k,l)H(i-k, j-l) + Z(i,j)$$

Handling linear models of degradation

So far, simple degradation process: **linear**, $Y = HX + Z$.

Can reconstruct X using (regularized) least-squares.

What about more sophisticated models of blurring?

What if $Y = \phi(H \star X) \odot Z$?

Bayesian approach:

- Prior distribution on X
- Probabilistic model of corruption process
- Given Y, determine posterior distribution over X
- Sample from this posterior or find the MAP (maximum a-posteriori) model

What prior distribution over images?

Think of each pixel as a random variable.

X_1	X_2	X_3	X_4
X_5	X_6	X_7	X_8
X_9	X_{10}	X_{11}	X_{12}

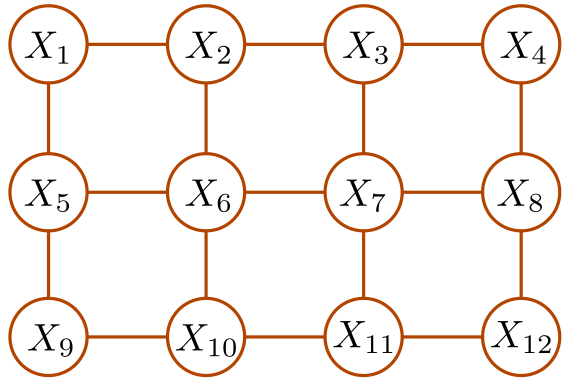
$X_1, \ldots, X_n$ are not independent, but the dependencies aren't arbitrary either.

Possible assumption:
Each pixel is conditionally independent of the others **given** its neighbors, e.g.

$$X_1 \perp\!\!\!\perp \{X_2, \ldots, X_{12}\} \mid X_2, X_5$$

Implication (Hammersley-Clifford Thm):

The distribution of $X = (X_1, \ldots, X_n)$ can be represented by a grid-shaped **Markov random field**.



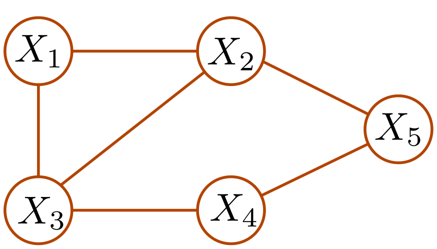
Markov random fields

Joint distribution over random variables $X_1, \ldots, X_n$ given by:

1. An undirected graph with nodes $X_1, \ldots, X_n$ and edges representing dependencies.
2. A distribution that *factors* over this graph:

$$P(X_1, \ldots, X_n) = \frac{1}{Z} \prod_C \Psi_C(\{X_i : i \in C\})$$

where the product is over maximal cliques in the graph, and the **clique potentials** Ψ_C are positive-valued functions.



Functional form of $P(X_1, X_2, X_3, X_4, X_5)$:

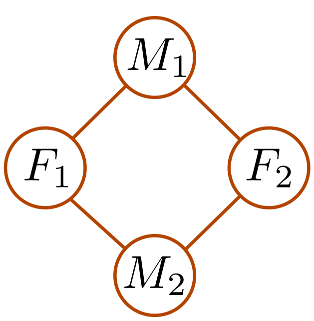
$$\frac{1}{Z}\Psi_{123}(X_1, X_2, X_3)\Psi_{25}(X_2, X_5)\Psi_{34}(X_3, X_4)\Psi_{45}(X_4, X_5)$$

B: Independence properties of MRFs

Example (from Pearl)

Four people engage in occasional pairwise activities. There is a disease going around.

Boolean variables (0/1): have disease?



F_1	M_1	$\Psi_{11}(F_1, M_1)$
0	0	100
0	1	20
1	0	20
1	1	50

F_1	M_2	$\Psi_{12}(F_1, M_2)$
0	0	100
0	1	20
1	0	20
1	1	50

F_2	M_1	$\Psi_{21}(F_2, M_1)$
0	0	200
0	1	10
1	0	100
1	1	50

F_2	M_2	$\Psi_{22}(F_2, M_2)$
0	0	100
0	1	20
1	0	20
1	1	1

- What is the most likely configuration?
- What are the conditional independence relationships here?

Conditional independence in MRFs

Let G be an undirected graph with nodes $X_1, \ldots, X_n$.
Let $N_G(X_i)$ denote the neighbors of X_i in G.

1. Any MRF over G satisfies, for all i, the **local independence property**

$$X_i \perp\!\!\!\perp \{X_j : j \neq i\} \mid N_G(X_i).$$

Easy proof: Algebraic manipulation of functional form of MRF.

2. **Global independence property**: for any subsets of nodes S, T, U such that removing U separates S from T,

$$X_S \perp\!\!\!\perp X_T \mid X_U.$$

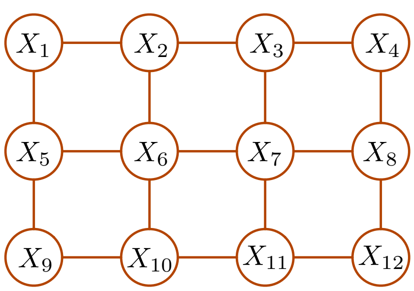
3. **Hammersley-Clifford Thm.** Let P be a distribution on $(X_1, \ldots, X_n)$ such that
   - $P(x) > 0$ for all x, and
   - P satisfies the local independence properties.

   Then P can be expressed as an MRF over G.

C: Inference by sampling

Back to image restoration

Recall: $Y = \phi(H \star X) \odot Z$. For prior on X, use a grid-shaped MRF.



Distribution

$$P(X) = \frac{1}{Z} \prod_{\text{edges } (i,j)} \Psi_{ij}(X_i, X_j).$$

E.g. $\Psi_{ij}(X_i, X_j) = \alpha^{|X_i - X_j|}$.

"Energy-based" form:

$$P(X) = \frac{1}{Z} e^{-U(X)}, \quad \text{where } U(x) = \sum_{(i,j) \in E} U_{ij}(x_i, x_j).$$

What is U_{ij} in the example above, and what is the lowest energy configuration?

Posterior distribution

Say $Y = \phi(H \star X) + Z$, where $Z \sim N(0, \sigma^2 I_n)$.
What is the posterior on X, and does it correspond to some MRF?

Inference: three algorithmic tasks

Suppose the posterior distribution is $P(x) \propto \exp(-U(x))$.

(1) Sample from the posterior.
We'll see how to do this using **Gibbs sampling**.

(2) Compute posterior expectations, e.g. $\mathbb{E}X_i$.
Easy to estimate using (1).

(3) Find the maximum a-posteriori (MAP) image.
Problem: The landscape of $U(x)$ is typically riddled with local optima.

Simulated annealing:

- Introduce a **temperature** $T > 0$ and define $P_T(x) \propto \exp(-U(x)/T)$.
    - High temp $T \to \infty$: $P_T \to$ uniform.
    - Low temp $T \to 0$: P_T concentrates near low-energy configurations.
- Simulated annealing: Run sampler for P_T, gradually letting T go to zero.
- If this is done slowly, it ultimately yields the MAP solution.

Gibbs sampler

Note: rejection sampling would be horrendously slow in this setting.

To sample from a distribution P over $(x_1, \ldots, x_n)$:

- Start with any x in the support
- Repeat:
  - Pick a feature $i \in \{1, 2, \ldots, n\}$
  - Resample x_i from $P(X_i = x_i | x_{\setminus i})$

E.g. if the X_i are $0 - 1$ valued then in each step:

- pick a feature i
- set $x_i = 1$ with probability

$$\frac{P(x_i = 1, x_{\setminus i})}{P(x_i = 0, x_{\setminus i}) + P(x_i = 1, x_{\setminus i})}$$

Guaranteed to converge to the right distribution!

Other approaches to inference

Recall three types of query: (1) **conditional probability query**, (2) **most probable explanation**, (3) **maximum a posteriori**.

Similar landscape to Bayes nets:

- All three types of query are NP-hard.
- Efficient exact inference for trees, or more generally, for bounded tree-width.
- Approximate inference using sampling, variational methods, belief propagation.

D: Energy-based models

Energy-based formalism

Density of the form $p(x) \propto \exp(-U(x))$

- $U(x)$ is the *energy function*
- E.g., $U(x)$ could be a neural network
- Give up on computing the normalization factor!

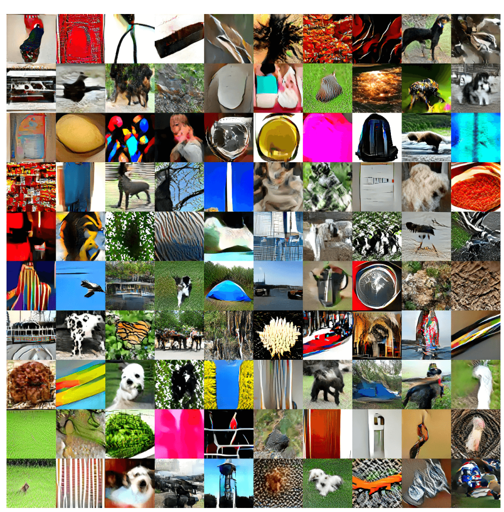
What can we do without normalization?

- Compute likelihoods?
- Sample?
- Generate most likely explanation/completion?
- Learn?

Example

Du, Mordatch (2019). *Implicit generation and modeling with energy-based models.*

Conditional generation after training on Imagenet128:



Other experiments with **compositionality**.

Sampling from an energy-based model

For $p(x) \propto \exp(-U(x))$, can use Gibbs sampling.

Alternative: **Langevin sampler**.

- Initialize $x \in \mathbb{R}^d$
- Repeat:
  - Sample $Z \sim N(0, I_d)$
  - Set $x \leftarrow x - \gamma \nabla_x U(x) + \sqrt{2\gamma} Z$

If $\nabla_x U(x)$ is well-behaved (e.g., Lipschitz), this gets close to $p(\cdot)$.

Learning 1: Maximum likelihood

Let $U_\theta(x)$ be the energy function with (e.g., neural net) parameters θ.

$$p_\theta(x) = e^{-U_\theta(x)}/Z_\theta$$
$$Z_\theta = \int e^{-U_\theta(x)} dx$$

Objective: given data $x_1, \ldots, x_n$, maximize likelihood

$$LL(\theta) = \sum_{i=1}^{n} \ln p_\theta(x_i).$$

Key fact: $\nabla_\theta \ln p_\theta(x) = -\nabla_\theta U_\theta(x) + \mathbb{E}_{X \sim p_\theta}[\nabla_\theta U_\theta(X)]$.

- Thus, can use gradient descent
- Estimate $\mathbb{E}_{X \sim p_\theta}[\cdot]$ by sampling from p_θ

We have $p_\theta(x) = e^{-U_\theta(x)}/Z_\theta$ where $Z_\theta = \int e^{-U_\theta(x)} dx$.

Check: $\nabla_\theta \ln p_\theta(x) = -\nabla_\theta U_\theta(x) + \mathbb{E}_{X \sim p_\theta}[\nabla_\theta U_\theta(X)]$.

Learning 2: Noise-contrastive estimation

Gutmann, Hyvarinen (2010). *Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics*.

- True data distribution $p_{\rm data}$ that we want to fit
- We have a family of **unnormalized** densities $\{\exp(-U_\theta(x)) : \theta \in \Theta\}$. These produce densities $p_\theta(x) = \exp(-U_\theta(x))/Z_\theta$, but normalizers Z_θ not known.

High-level scheme:

- Define an **augmented family** that has all multiples of the unnormalized densities:

$$q_{\tilde{\theta}}(x) = \exp(-U_\theta(x))/c \quad \text{for } \tilde{\theta} = (\theta, c) \in \Theta \times \mathbb{R}^+$$

- We will learn $\tilde{\theta}$, the model as well as its normalizer!
- We'll do this by maximizing a likelihood-type objective function $J(\tilde{\theta})$

Noise-contrastive estimation

- Data distribution: p_{data}
- Choose a **noise distribution** p_n, e.g. $N(0, I)$

Define objective function

$$J(\tilde{\theta}) = \mathbb{E}_{x \sim p_{\text{data}}} \left[\ln \frac{q_{\tilde{\theta}}(x)}{q_{\tilde{\theta}}(x) + p_n(x)} \right] + \mathbb{E}_{x \sim p_n} \left[\ln \frac{p_n(x)}{q_{\tilde{\theta}}(x) + p_n(x)} \right].$$

This is binary cross-entropy for separating p_{data} from p_n.

Claim: If $p_{\text{data}} = p_{\theta^*}$ for some $\theta^* \in \Theta$, then $J(\tilde{\theta})$ is maximized by $\tilde{\theta} = (\theta^*, Z_{\theta^*})$.