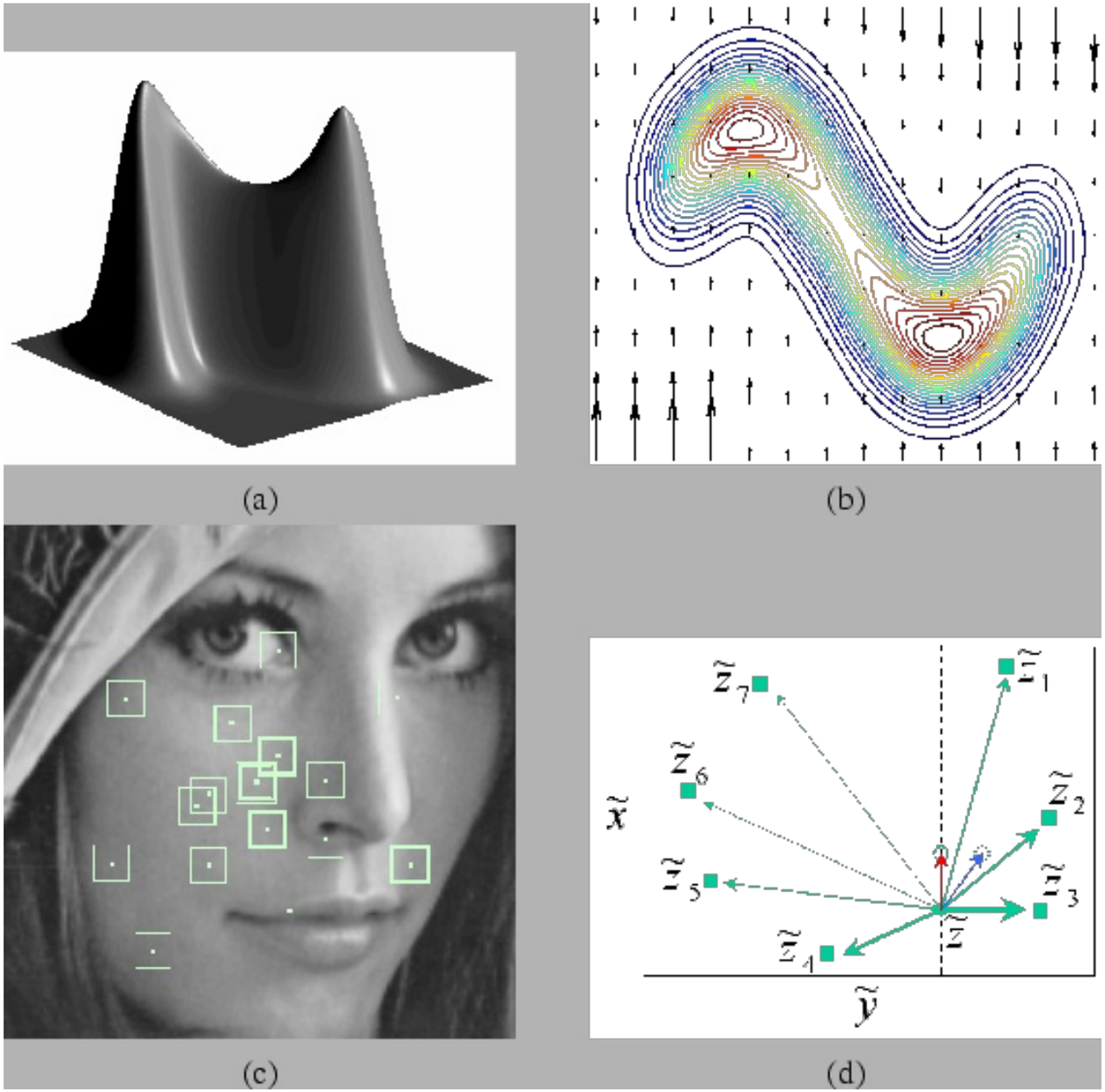
Markov Random Fields and Bayesian Image Analysis

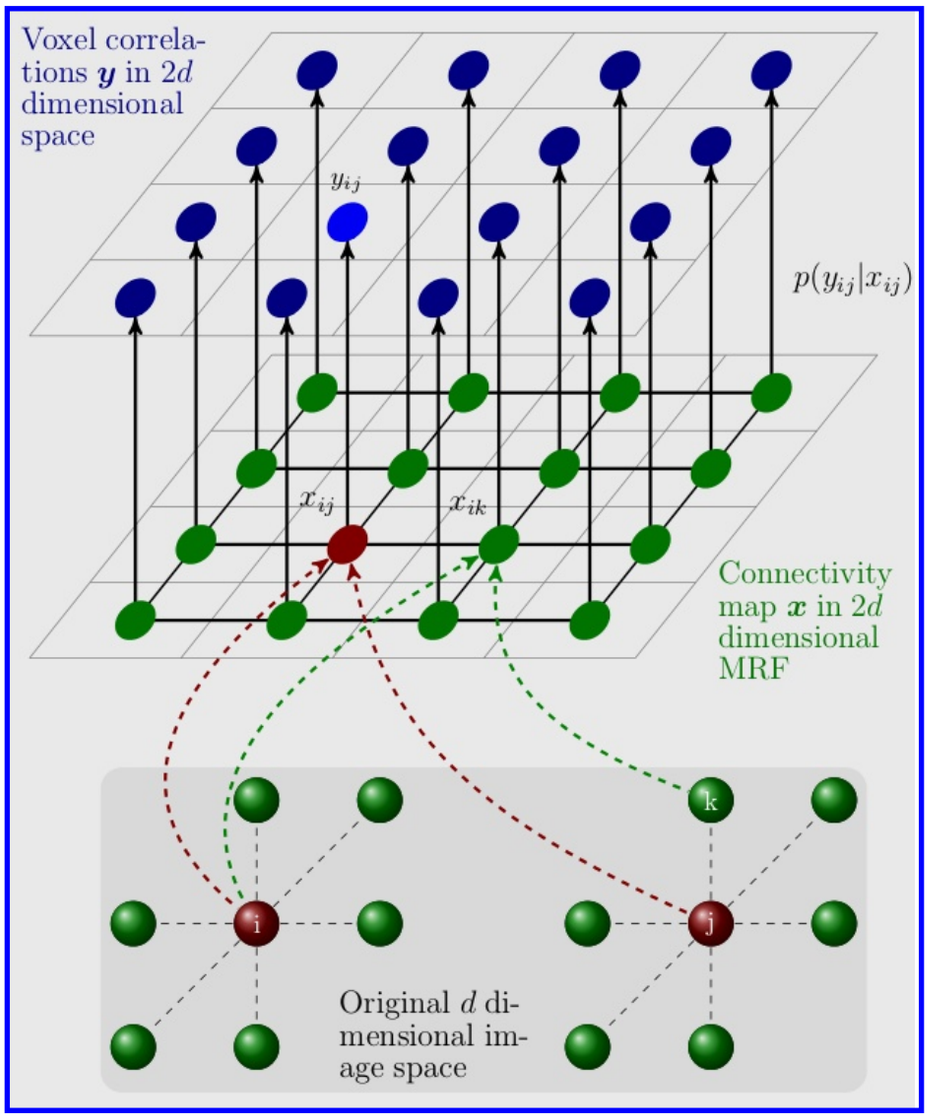
Wei Liu
Advisor: Tom Fletcher

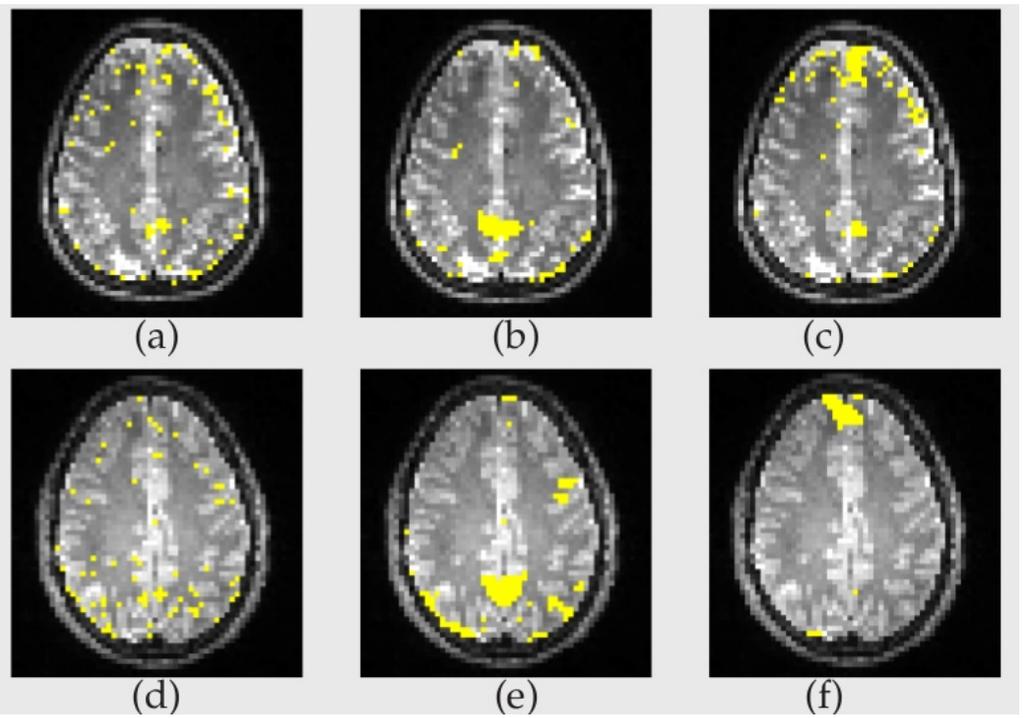
Markov Random Field: Application Overview



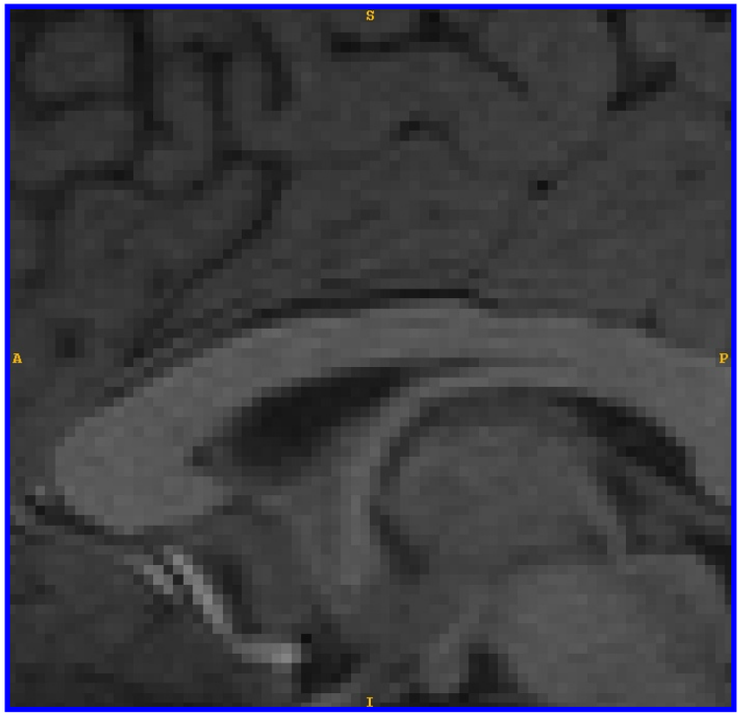
Awate and Whitaker 2006

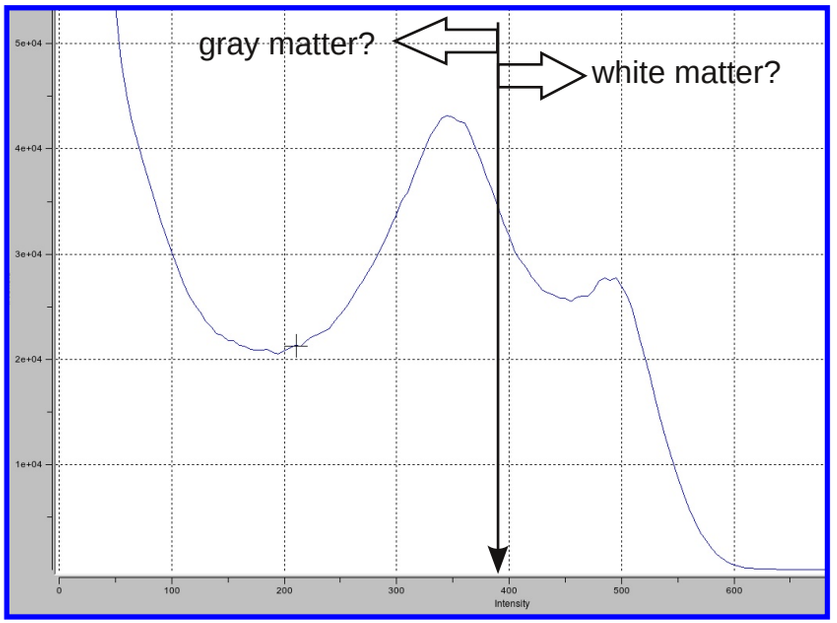
Markov Random Field: Application Overview

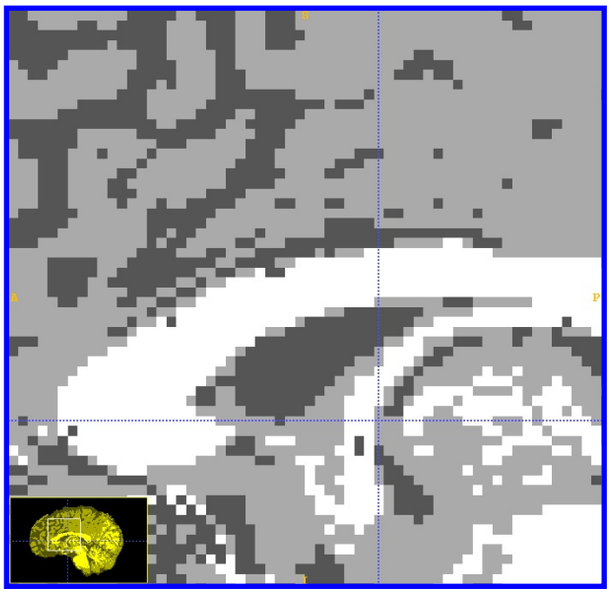




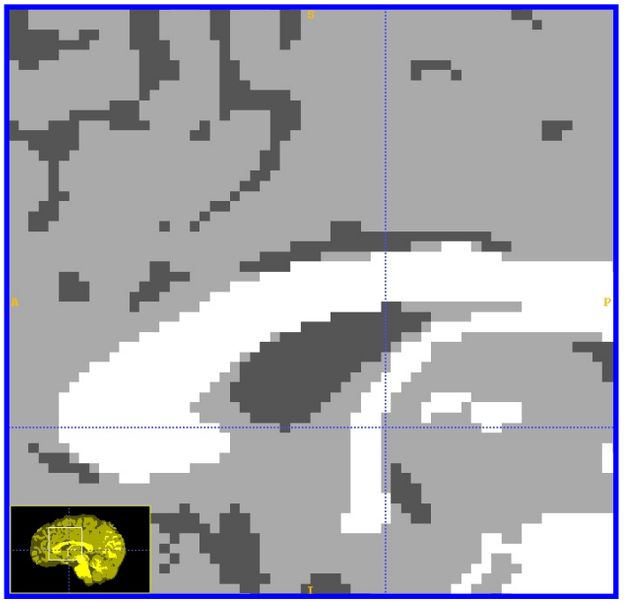
Markov Random Field: Application Overview







without spatial MRF prior



with spatial MRF prior

Bayesian Image Analysis

- Unknown 'true' image X
- observed data Y
- Model $\mathcal{M}$ and parameter set θ

Goal: Estimate X from Y based on some objective funciton.

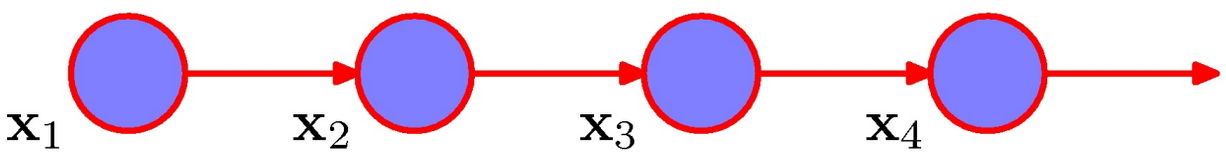
Review: Markov Chains

Definition 1. *A markov chain is a sequence of random variables* $X_1, X_2, X_3, \ldots$ *with the Markov property that given the present state, the future and past states are* ***conditionally*** *indepedent.*

$$P(X_{n+1}|X_1, X_2, \ldots, X_n) = P(X_{n+1}|x_n)$$

The joint probability of the sequence is given by

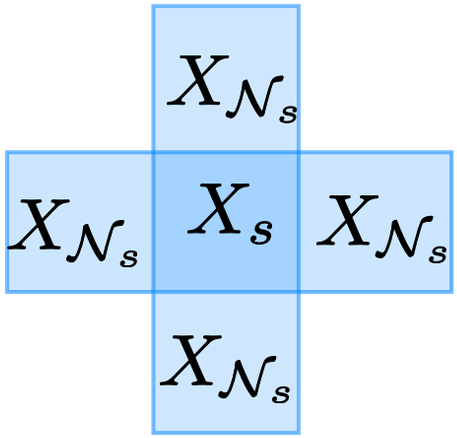
$$P(X) = P(X_0) \prod_{n=1}^{N} P(X_n|X_{n-1})$$



Markov Random Fields: Some Definition

Define

- $\mathcal{S} = \{1, \ldots, M\}$ the set of lattice points.
- $s \in \mathcal{S}$ a site in $\mathcal{S}$
- $\mathcal{L} = \{1, \ldots, L\}$ the set of labels
- X_s the random variable at s. $X_s = x_s \in \mathcal{L}$
- $\mathcal{N}_s$ the set of sites neighboring s. Properties of neighboring sites:
  - $s \notin \mathcal{N}_s$
  - $s \in \mathcal{N}_t \Leftrightarrow t \in \mathcal{N}_s$
- $\mathcal{S}$ and neighbor system $\mathcal{N}$ together defines a graph $(\mathcal{S}, \mathcal{N}) = \mathcal{G}$.



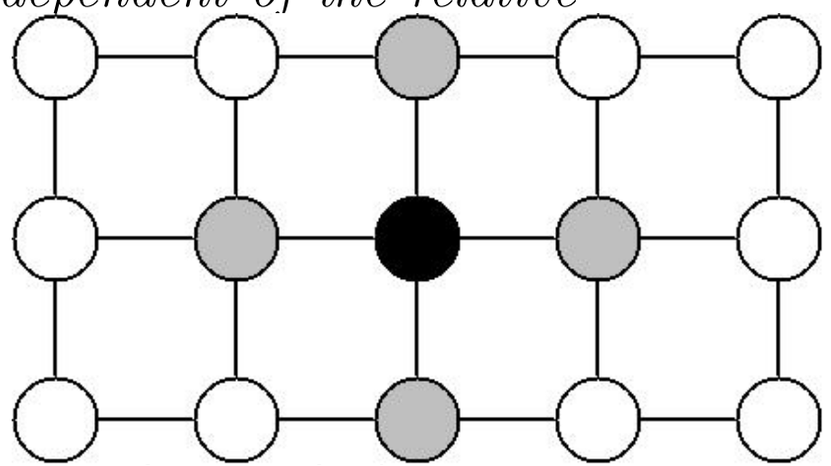
Markov Random Fields: Some Definition

Definition. *X is called a random field if $X = \{X_1, \ldots, X_N\}$ is a collection of random variables defined on the set $\mathcal{S}$, where each X_s takes a value x_s in $\mathcal{L}$. $x = \{x_1, \ldots, x_N\}$ is called a configuration of the field.*

Definition. *X is said to be a Markov random field on $\mathcal{S}$ with respect to a neighborhood system $\mathcal{N}$ if for all $s \in \mathcal{S}$*

$$P(X_s|X_{\mathcal{S}-s}) = P(X_s|X_{\mathcal{N}_s})$$

Definition. *X is homogeneous if $P(X_s|X_{\mathcal{N}_s})$ is independent of the relative location of site s in $\mathcal{S}$.*



Generalization of Markov chain:

- unilateral $\rightarrow$ bilateral
- 1D $\rightarrow$ 2D
- time domain $\rightarrow$ space domain. No natural ordering on image pixels.

Markov Random Fields: Issues

Advantage of MRF's

- Can be isotropic or anistropic depending on the definition of neighbor system $\mathcal{N}$.
- Local dependencies

Disadvantages of MRF's

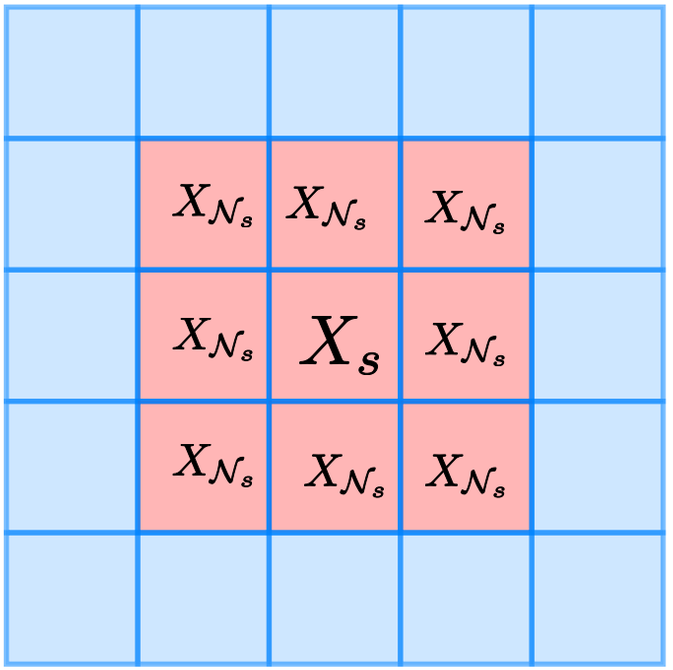
- difficult to compute $P(X)$ from local dependency $P(X_s|X_{\mathcal{N}_s})$
- Parameter estimation is difficult

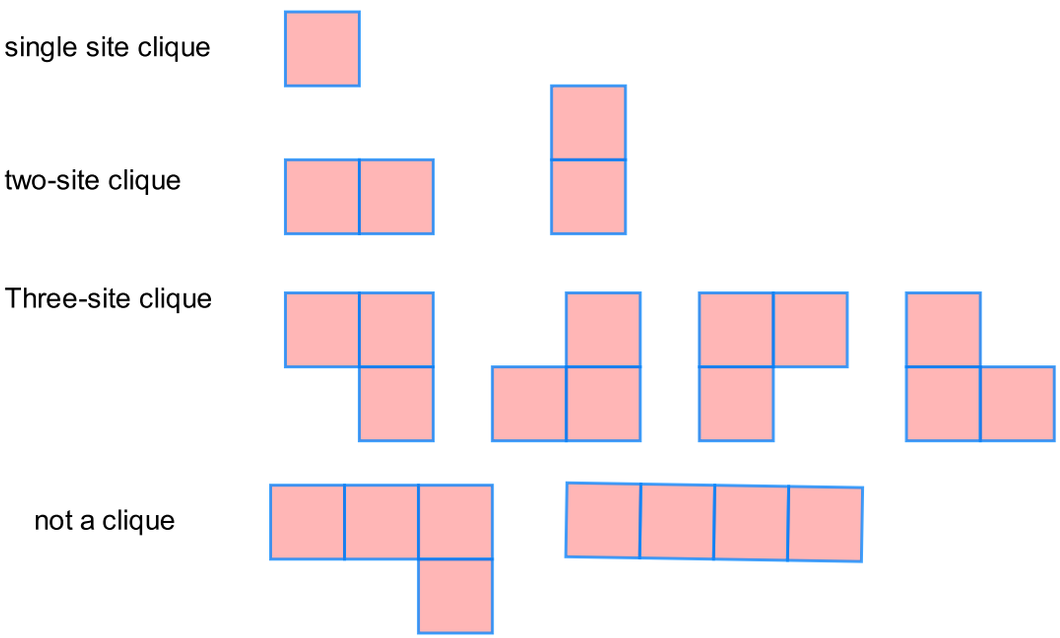
Hammersley-Clifford theorem build the relationship between local properties $P(X_s|X_{\mathcal{N}_s})$ and global properties $P(X)$.

Gibbs Random fields: Definition

Definition. *A clique $\mathcal{C}$ is a set of points, which are all neighbors of each other*

$$\mathcal{C}_1 = \{s | s \in \mathcal{S}\}$$
$$\mathcal{C}_2 = \{(s,t) | s \in \mathcal{N}_t, \quad t \in \mathcal{N}_s\}$$
$$\mathcal{C}_3 = \dots$$





Gibbs Random Fields: Definition

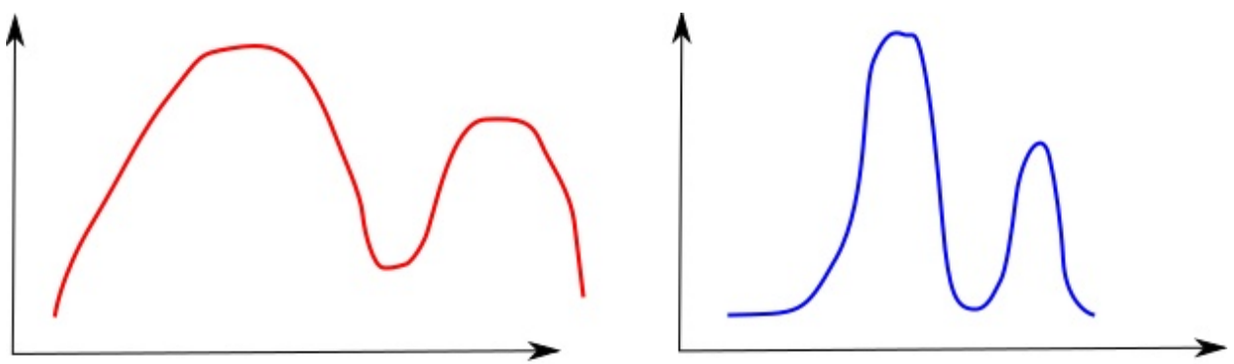
Definition. *A set of random variable X is said to be a Gibbs random field (GRF) on $\mathcal{S}$ with respect to $\mathcal{N}$ if and only if its configurations obey a Gibbs distribution*

$$P(X) = \frac{1}{Z}\exp\{-\frac{1}{T}U(X)\}$$

- $U(X)$ – energy function. Configurations with lower energy are more probable.

$$U(X) = \sum_{c \in \mathcal{C}} V_c(X)$$

- T – temperature. *Sharpness* of the distribution.
- Z – normalization constant. $Z = \sum_{X \in \mathcal{X}} \exp\{\frac{1}{T}U(X)\}$, $\mathcal{X} = \mathcal{L}^N$



High temperature

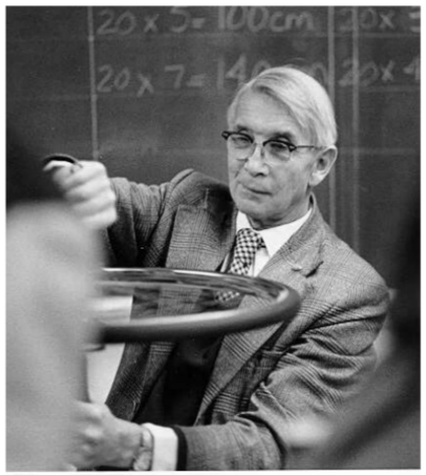
low temperature

Gibbs-Markov Equivalence

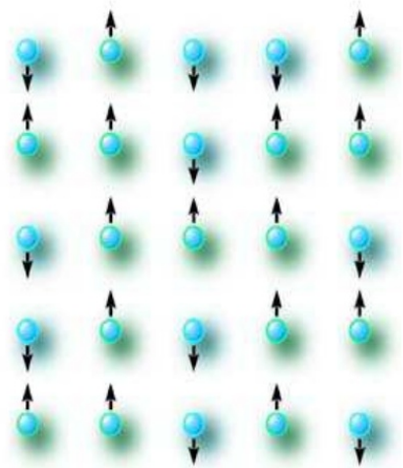
Theorem. *X is an Markov random field on $\mathcal{S}$ if and only if X is a Gibbs field on $\mathcal{S}$ with respect to $\mathcal{N}$.*

- Gives a method to specify joint probability by specifying the clique potential $V_c(X)$.
- Different clique potential gives different MRFs.
- Z is still difficult to compute.

Ising Model

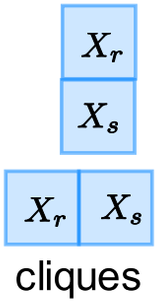


(a)



(b)

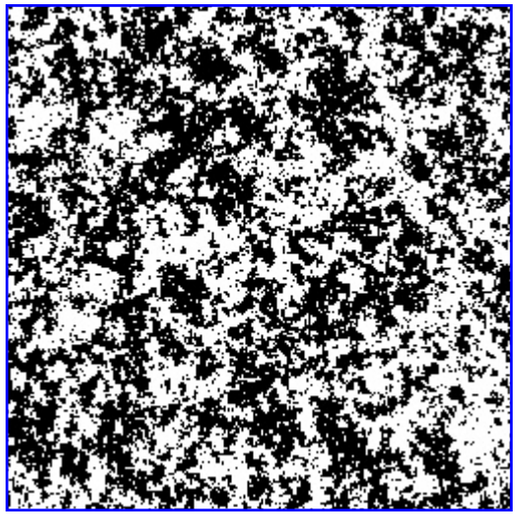
1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1
1	1	1	1	1	-1	-1	1
1	1	1	1	-1	-1	-1	1
1	1	1	-1	-1	-1	-1	1
1	1	1	1	-1	-1	1	1
1	1	1	-1	-1	-1	-1	1
1	1	1	1	1	1	1	1



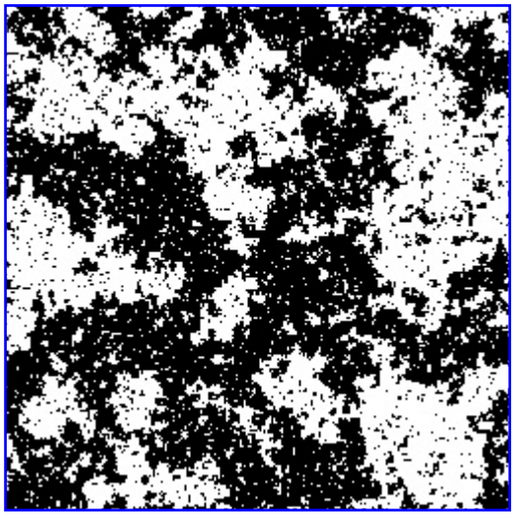
- Two state: $\mathcal{L} = \{-1, +1\}$
- Clique potential $V(X_r, X_s) = -\beta X_r X_s$
- $$U(X) = \sum_{c \in \mathcal{C}} V_c(X) = -\beta \sum_{(r,s) \in \mathcal{C}_2} X_r X_s, \quad P(X) = \frac{1}{Z} \exp\{-\frac{U(X)}{kT}\}$$
- Conditional distribution at site X_s:

$$P(X_s | X_{\mathcal{N}_s}) = \frac{\exp\{\beta X_s \sum_{r \in \mathcal{N}_s} X_r\}}{2 \cosh(\beta \sum_{r \in \mathcal{N}_s} X_r)}$$

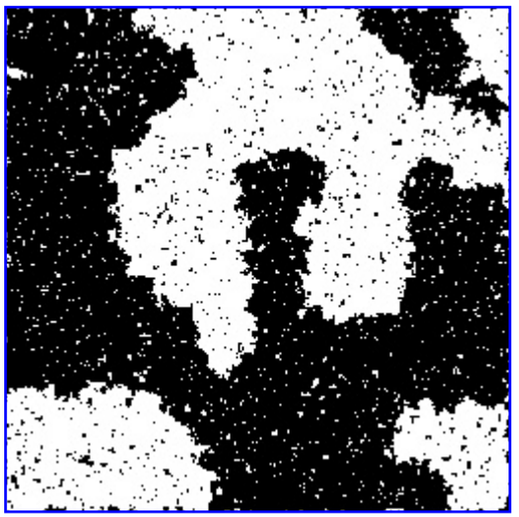
Ising Model



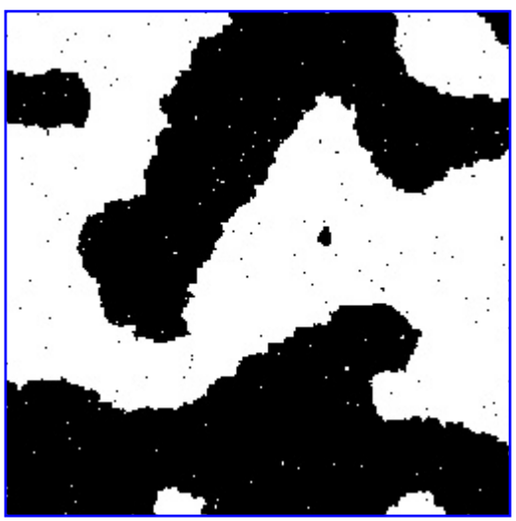
Beta = 0.8



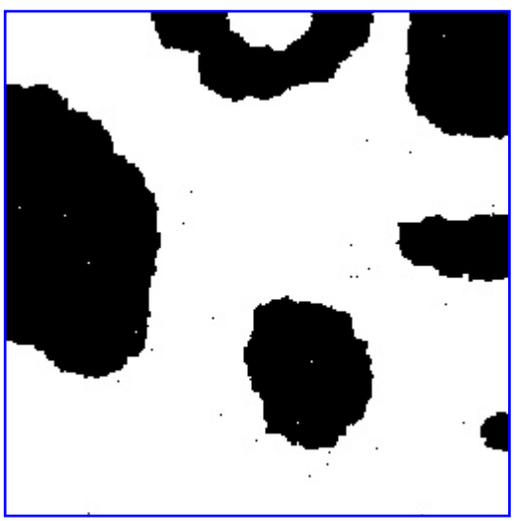
Beta = 0.88



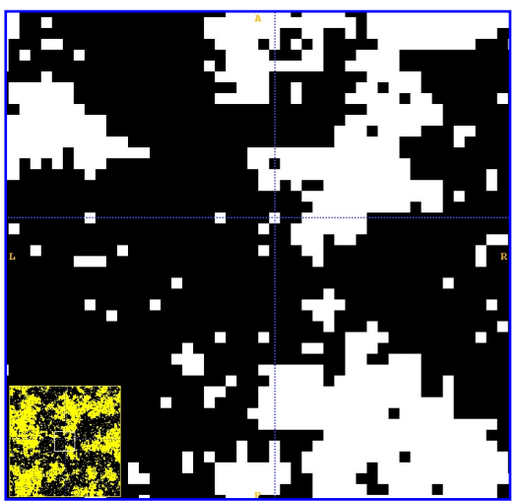
Beta = 1.0



Beta = 1.5



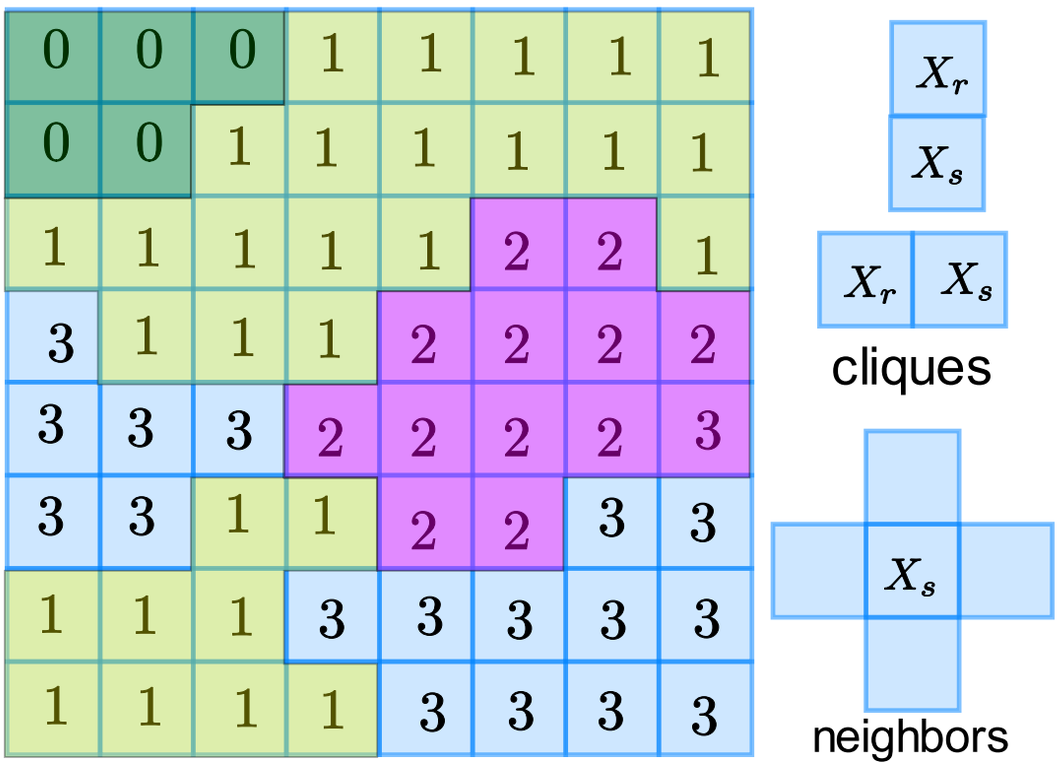
Beta = 2.0



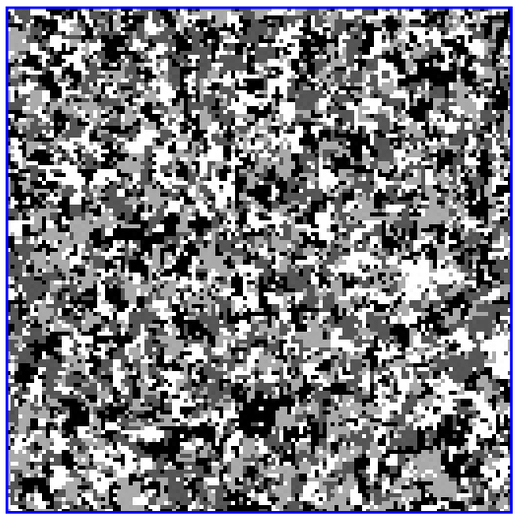
Beta = 0.88. detailed view

Potts Model

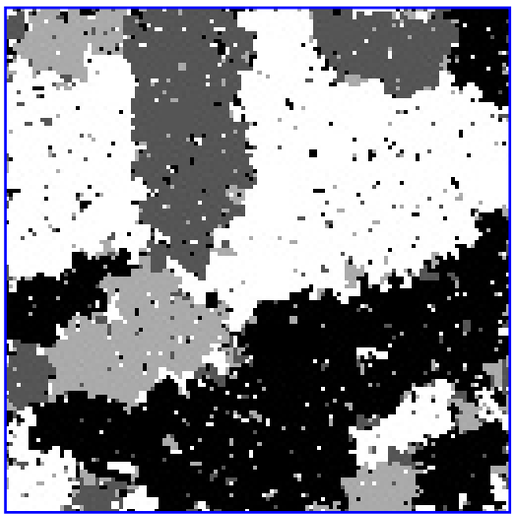
- Multiple state: $X_s = x_s \in \mathcal{L}, \quad \mathcal{L} = \{1, 2, \ldots, L\}$
- 4-neighbor or 8-neighbor system
- $V_1(X_s = l) = \alpha_l, \quad l \in \mathcal{L}$
- $V_2(X_r, X_s) = \begin{cases} \beta & X_r \neq X_s \\ 0 & X_r = X_s \end{cases}$



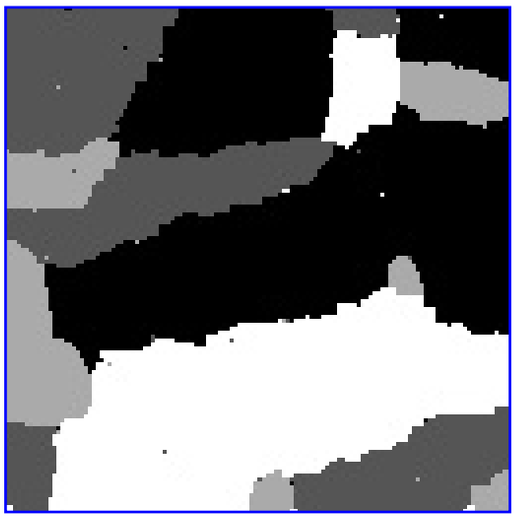
Potts Model example



Beta = 0.88

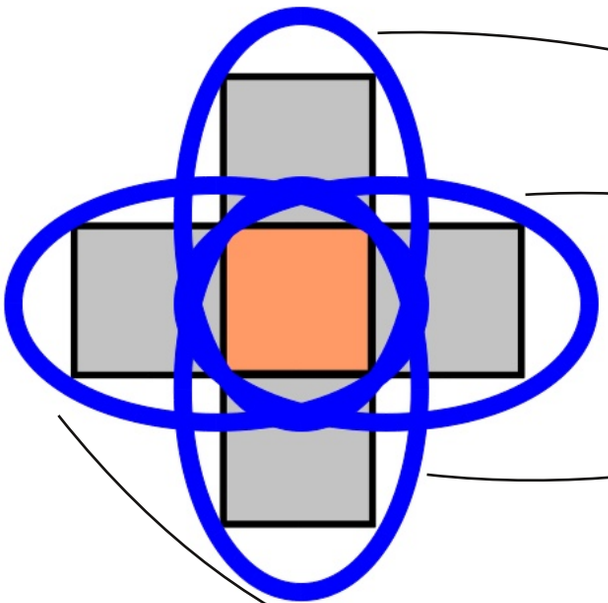


Beta = 1.2



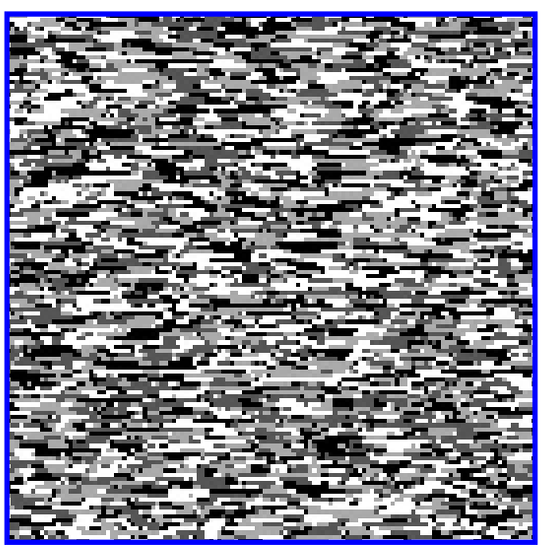
Beta = 2.0

Potts Model

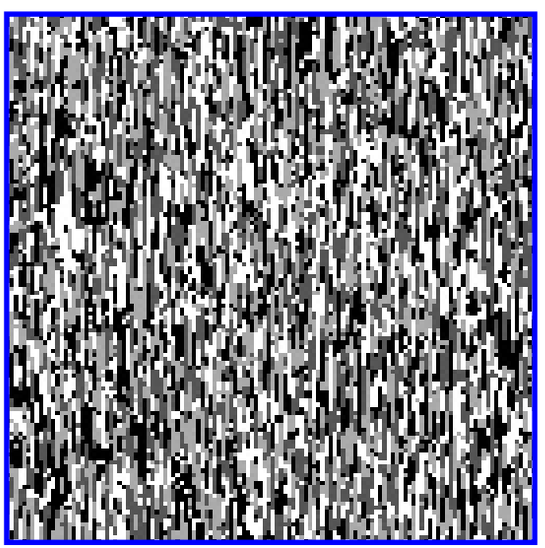


β can be different at different directions $\Rightarrow$ anistropic field.

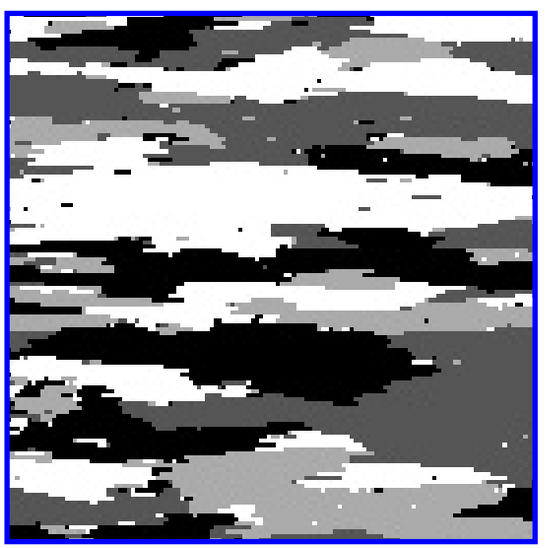
- β_u
- β_r
- β_d
- β_l



$\beta_u = \beta_d = 0, \beta_l = \beta_r = 2$

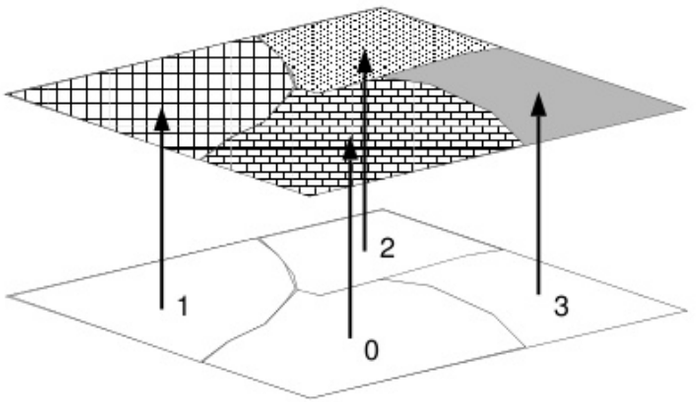


$\beta_u = \beta_d = 2, \beta_l = \beta_r = 0$



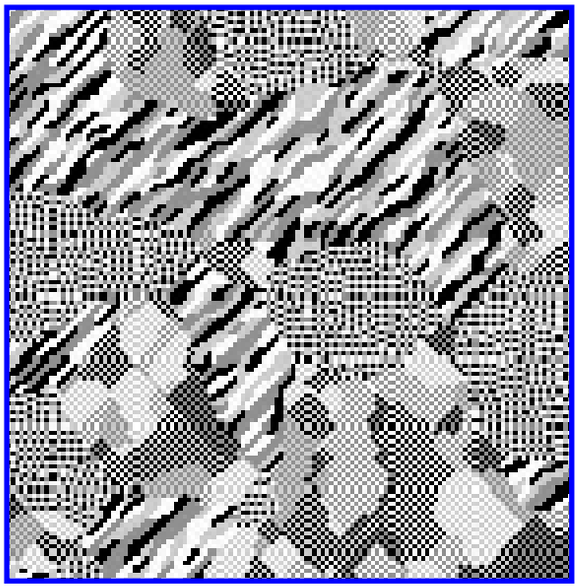
$\beta_u = \beta_d = 1, \beta_l = 4, \beta_r = 2$

Hierarchical MRF Model



Y - Texture feature vectors observed from image.

X - Unobserved field containing the class of each pixel



- $X \in \mathcal{L}^N$ is MRF – region configuration.
- $P(Y_s|X_s)$ depneds on $Y_{\mathcal{N}_s}$.
- Given X_s, $\{s, \mathcal{N}_s\}$ has same texture type.

Simulation of MRFs

Why do we want draw a sample of MRFs (or Gibbs distribution)?

$$P(X) = \frac{1}{Z} \exp\{-U(X)\}$$

- Compare simulatd image with real image $\Rightarrow$ Model is good?
- Texture synthesis
- Model verification.
- Monte Carlo integration

Review of Monte Carlo integration. Consider the generic problem of evaluating the integral

$$\mathbb{E}_{f(x)} = \int_{\mathcal{X}} h(x) f(x) dx$$

We can use a set of samples $(x_1, x_2, \ldots, x_M)$ generated from density $f(x)$ to approximate above integral by the empirical average

$$\overline{h} = \frac{1}{M} \sum_{m=1}^{M} h(x_m)$$

Simulation of MRFs

- Metropolis sampler. Used when we know $P(X)$ up to a constant
- Gibbs Sampler. Used when we know exactly $P(X)$

Metropolis Sampling: Review

Goal: draw samples from some distribution $P(X)$ where $P(X) = f(X)/K$.

- Start with any initial value X_0 satisfying $f(X_0) > 0$.
- Sample a candidate point X^* from distribution $g(X)$ (proposal distribution).
- Calculate the
$$\alpha = \frac{P(X^*)}{P(X_{t-1})} = \frac{f(X^*)}{f(X_{t-1})}$$
- If $\alpha > 1$, accept cadidate point and set $X_t = X^*$. Otherwise accept X^* with probability α.

We don't have to know the constant K!

Metropolis Sampling of MRFs

Goal: draw samples from Gibbs distribution $P(X) = \frac{1}{Z}\exp\{-U(X)\}$.

1. Randomly init X^0 satisfying $f(X^0) > 0$ (X is the whole iamge)
2. For $s \in \mathcal{S}$ do step 3, 4, 5
3. Generate a *univariate* sample X_s^* from proposal probability $Q(X_s^*|X^{t-1})$ (Q can be uniform distribution), and replace X_s with X_s^* to get candidate X^*. X^{t-1} and X^* differs only at X_s.
4. Calculate the

$$\Delta U(X^*) = U(X^*) - U(X^{t-1}) = U(X_s^*) - U(X_s)$$

5. If $\Delta U(X^*) < 0$, accept cadidate point and set $X^t = X^*$. Otherwise accept X^* with probability $\exp\{-\Delta U(X^*)\}$.
6. Repeat above steps M times.

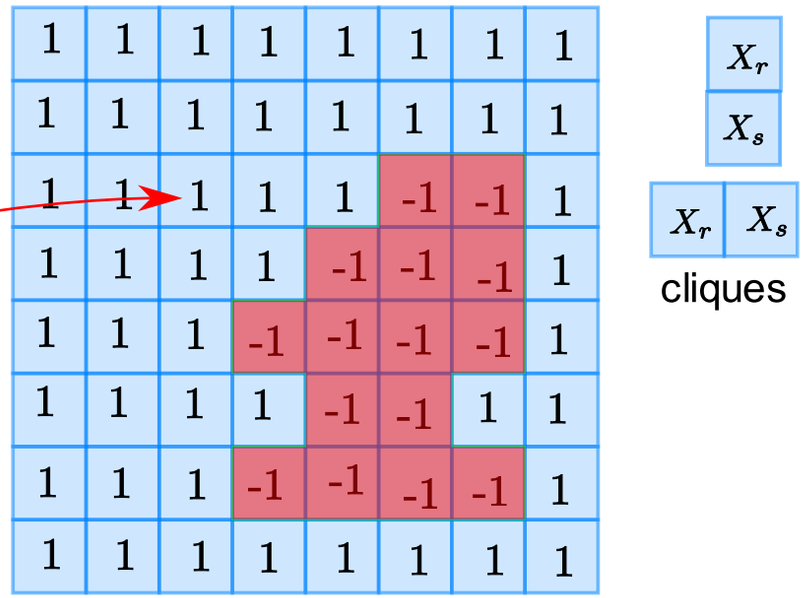
The sequence of random fields X^t (after burn-in period) is a Markov chain.

Gibbs Sampling of MRFs

Goal: draw samples from Gibbs distribution $P(X) = \frac{1}{Z}\exp\{-U(X)\}$.

1. Randomly init X^0 satisfying $f(X^0) > 0$ (X is the whole iamge)
2. For $s \in \mathcal{S}$ do step 3, 4
3. Compute $P(X_s|X^{t-1}) = P(X_s|X^{t-1}_{\mathcal{N}_s})$ and draw sample X^*_s from it.
4. Accept X^*_s, i.e. replace X_s with X^*_s and obtain X^t.
5. Repeat above steps M times.

The sequence of random fields X^t (after burn-in period) is a Markov chain.



Gibbs v.s. Metropolis Sampling

Gibbs:

- Always accepted.
- Have to compute $P(X_s = l|X_{\mathcal{N}_s})$ for all $l \in \mathcal{L}$.

Metropolis:

- Expected acceptance rate is $1/L$ – low when L is large $\Rightarrow$ more burn-in time.
- No need to compute $P(X_s = l|X_{\mathcal{N}_s})$ for all $l \in \mathcal{L}$. Only compute $U(X_s^*|X_{\mathcal{N}_s})$ for candidate X^*.

Bayesian Image Analysis

Image Segmentation:

- $X \in \mathcal{L}^N$: Image labels we're interested
- $Y \in \mathcal{R}^N$: noise data (observed image)

Goal: Estimate X from Y.

Image Denoising:

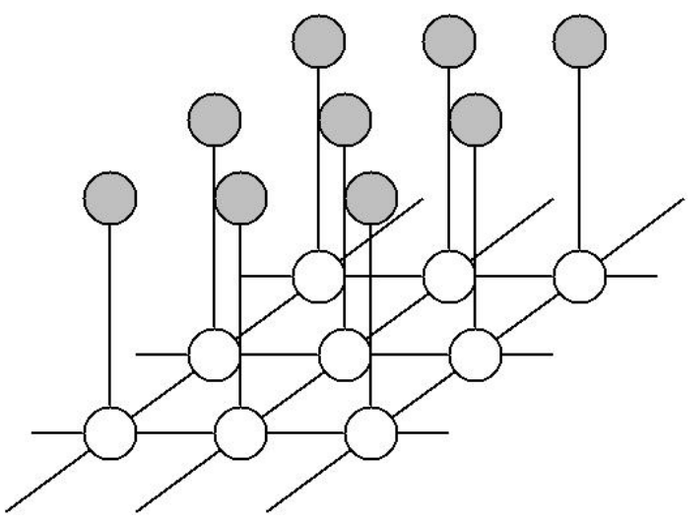
- $X \in \mathcal{R}^N$: True image intensity.
- $Y \in \mathcal{R}^N$: noise data (observed image)

Goal: Recover X from Y.

$$P(X|Y) \propto P(X) + P(Y|X)$$

MRF prior

$P(X) = \frac{1}{Z} \exp U(X)$



Conditional likelihood

For Segmenttation:

$P(Y|X) = \sum_{s \in \mathcal{S}} P(Y_s|X_s = l) = \mathcal{N}(\mu_l, \sigma_l^2)$

For denoising:

$P(Y|X) = \sum_{s \in \mathcal{S}} P(Y_s|X_s = x_s) = \mathcal{N}(x_s, \sigma_l^2)$

Bayesian Image Segmentation

- Define a model.

$$P(X) = \frac{1}{Z}\exp\{U(X)\}$$
$$P(Y|X) = \sum_{s\in\mathcal{S}} P(Y_s|X_s = l) = \mathcal{N}(\mu_l, \sigma_l^2)$$

- Formulation of objective function. Optimal Criteria.
- Search solution in the admissible space.

Bayesian Risk

- Bayesian Risk is defined as

$$R(X^*) = \int_{X \in \mathcal{X}} C(X, X^*) P(X|Y) dX$$

- $C(X, X^*)$: cost function. X: true value. X^*: estimated value.
  - $C(X, X^*) = ||X - X^*||^2 \Rightarrow X^* = \int_{X \in \mathcal{X}} P(X|Y) \mathrm{d}X$ (Posterior mean)
  - $C(X, X^*) = \begin{cases} 0 & ||X - X^*|| \leq \delta \\ 1 & \text{otherwise} \end{cases} \Rightarrow X^* = \text{argmax}_{X \in \mathcal{X}} P(X|Y) = \text{argmax}_{X \in \mathcal{X}} \left(P(X) + P(Y|X)\right)$. This is mode of posterior.

MRF-MAP: Case Study

Image Segmentation. Two classes $\mathcal{L} = \{-1, 1\}$

- Prior is Ising model
  - $P(X) = \frac{1}{Z}\exp\{U(X)\}$, $U(X) = -\beta \sum_{(r,s)\in\mathcal{C}_2} X_r X_s$. Assume T and K is 1.
  - $P(X_s|X_{\mathcal{N}_s}) = \frac{\exp\{\beta X_s \sum_{r\in\mathcal{N}_s} X_r\}}{2\cosh(\beta \sum_{r\in\mathcal{N}_s} X_r)}$
- Conditional likelihood $P(Y|X) = \prod_{s\in\mathcal{S}} P(Y_s|X_s), \quad P(Y_s|X_s = l) = \mathcal{N}(\mu_l, \sigma_l^2)$
- objective function:

$$\begin{aligned}\log P(X|Y) &\propto \log P(X) + \log P(Y|X) \\ &= -\beta \sum_{(r,s)\in\mathcal{C}_2} X_r X_s - \log(Z) + \sum_{s\in\mathcal{S}} \frac{(Y_s - \mu_l)^2}{2\sigma_l^2} - \log(\sigma_l) + \text{const}\end{aligned}$$

- Combinatorial optimization problem. NP hard.

Posterior Optimization

(Approximation) Optimization method:

- Iterated Conditional Modes
- Simulated Annealing
- Graph-cuts

Strategy:

- constrained minimization $\Rightarrow$ unconstrained minimization (Lagrange multiplier).
- discrete labels $\Rightarrow$ continuous labels (Relaxation labeling).

1. Init X by maximum likelihood $X^0 = \text{argmax}_{X \in \mathcal{X}} P(Y|X)$

2. For $s \in \mathcal{S}$, update X_s by

$$X_s^{t+1} = \text{argmax}_{X_s \in \mathcal{L}} \log P(X_s | X_{\mathcal{N}_s}^t, Y_s).$$

   For the Ising-Gaussian case, this is

$$\begin{aligned} X_s^{t+1} &= \text{argmax}_{X_s \in \mathcal{L}} \log P(X_s | X_{\mathcal{N}_s}) + \log P(Y_s | X_s) \\ &= \text{argmin}_{X_s \in \mathcal{L}} \left\{ -\beta X_s \sum_{r \in \mathcal{N}_s} X_r^t + \frac{(Y_s - \mu_l)^2}{2\sigma^2} + \log(\sigma_l) \right\}. \end{aligned}$$

   Note μ_l and σ_l is function of X_s, and $\log(Z_s) = \log(2\cosh(\beta \sum_{r \in \mathcal{N}_s} X_r^t))$ is not a function of X_s.

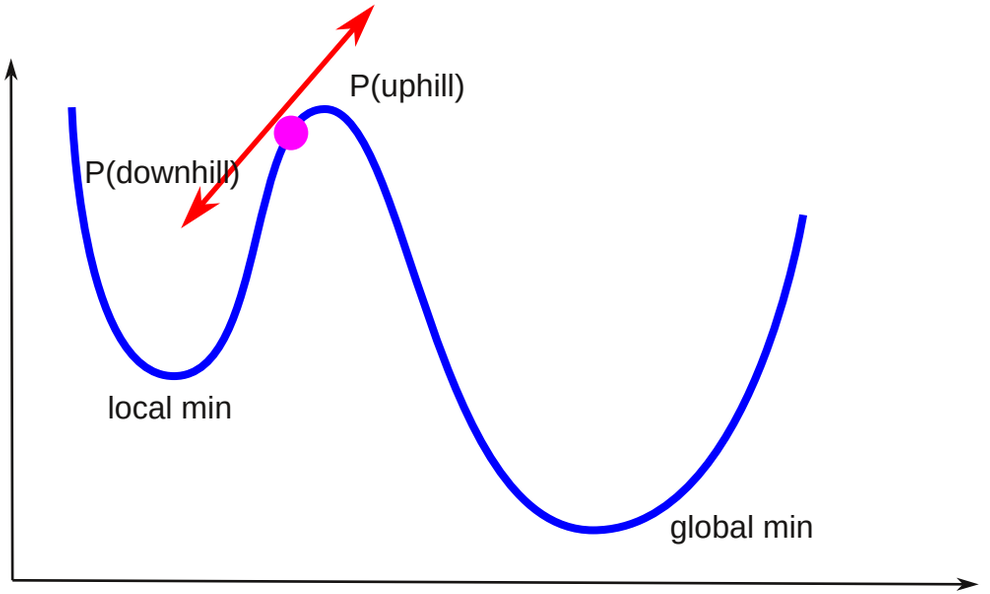
3. Do above step for all $s \in \mathcal{S}$.

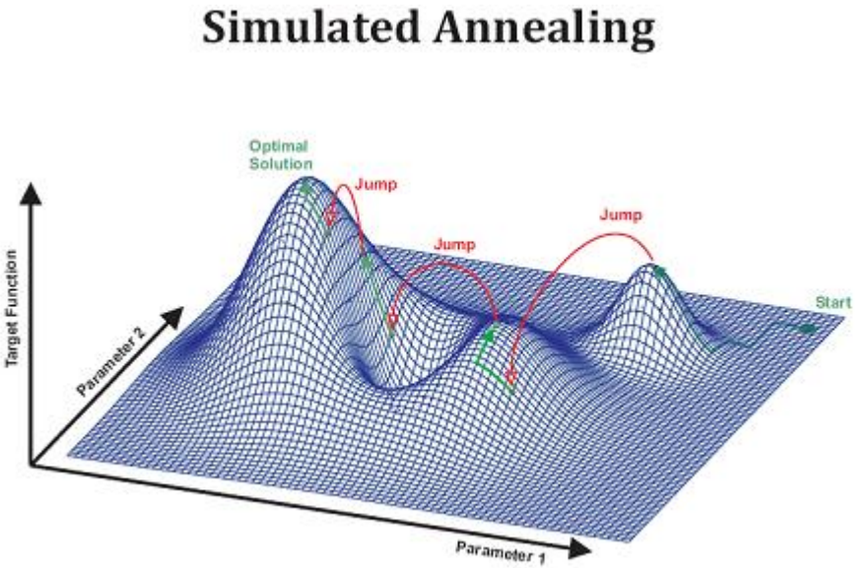
4. Repeat 2 and 3 until converge.

ICM cont.

- Greddy algorithm $\Rightarrow$ local minimum.
- Sensitive to initialization.
- Quick convergence.

Simulated Annealing





- Not always downhill moving.
- Global minimum with enough scan.

Simulated Annealing

- Assuming Ising+Gaussian model

$$
\begin{aligned}
P(X|Y) &\propto P(X) \cdot P(Y|X) \\
&= \frac{1}{Z} \exp\{\beta \sum_{(r,s)\in\mathcal{C}_2} X_r X_s\} \cdot \prod_{s\in\mathcal{S}} \exp\left\{-\frac{(X_s - \mu_l(X_s))^2}{2\sigma_l^2(X_s)} - \log(\sqrt{2\pi}\sigma_l(X_s))\right\} \\
&= \frac{1}{Z_P} \exp\{-U_P(X|Y)\} \\
U_P(X|Y) &= -\beta \sum_{(r,s)\in\mathcal{C}_2} X_r X_s + \frac{(X_s - \mu_l(X_s))^2}{2\sigma_l^2(X_s)} + \sqrt{2\pi}\sigma_l(X_s)
\end{aligned}
$$

Posterior distribution $P(X|Y)$ is also Gibbs.

Simulated Annealing cont.

Goal: Find $\text{argmax}_{X \in \mathcal{X}} P(X|Y)$

- Introduce temperature T:

$$P(X|Y) = \frac{1}{Z_P} \exp\{U_P(X|Y)\} \Rightarrow P(X|Y) = \frac{1}{Z_P} \exp\left\{\frac{U_P(X|Y)}{T}\right\}$$

1. Init with X^0 and a high temperature T.
2. Draw samples form $P(X|Y)$ by Gibbs Sampling or Metropolis Sampling (by sample from $P(X_s|Y_s), \forall s \in \mathcal{S}$.
3. Decrease T and repeat step 2.
4. Repeat step 2 and 3 until T is low enough.

Why this works?

Energy minimization for Segmentation

	Variational methods (optimization in $\mathcal{R}^\infty$)	Combinatorial methods (optimization in $\mathcal{Z}^n$)
explicit boundary representation	*Snakes &Balloons* (variational formulations) e.g. (Kass et al., 1988; Cohen, 1991)	*Dynamic Programming* and "*path-based*" *graph methods* (2D only) (e.g., Amir et al., 1990; Geiger et al., 1995; Mortensen and Barrett, 1998; Falcão et al., 1998; Jermyn and Ishikawa, 1999)
implicit boundary representation	*Level-sets* (e.g., Sethian, 1999; Osher and Fedkiw, 2002; Sapiro, 2001; Osher and Paragios, 2003)	**Combinatorial Graph Cuts** (as originally outlined in Boykov and Jolly, 2001)

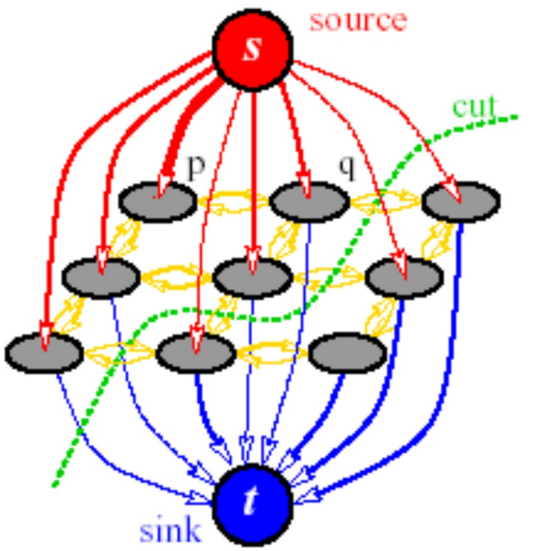
Boykov et. al. 2006

Graph Cuts for Ising Model

- Different with Normalized Cuts.
- For two class labeling problem, find the *global* minimum of

$$\log P(X|Y) = \sum_{s\in\mathcal{S}} \lambda_s X_s + \sum_{(r,s)\in\mathcal{C}_2} \beta_{(r,s)}(X_s X_r + (1-X_s)(1-X_r)),$$

where $\lambda_s = \log P(Y_s|X_s = 1)/P(Y_S|Y_s = 0))$.



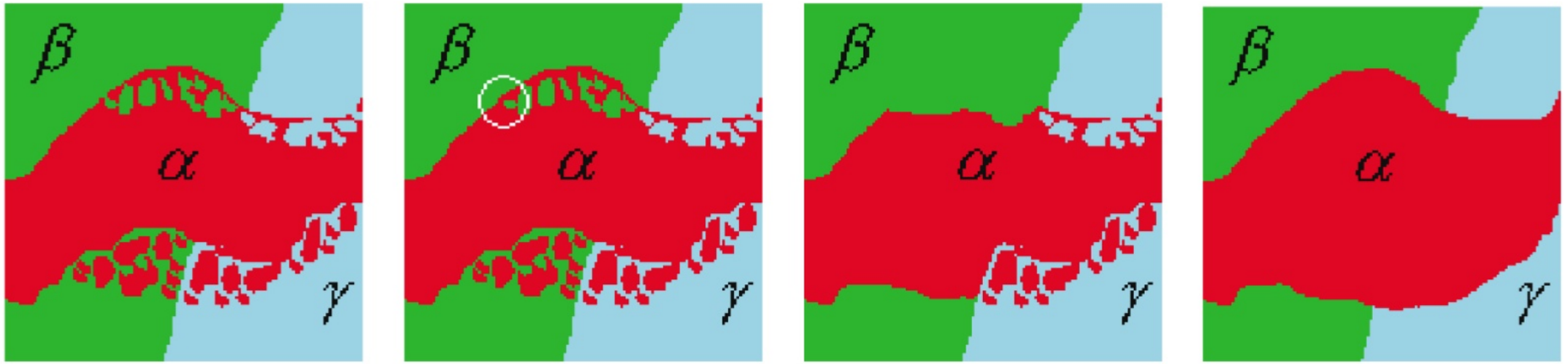
Boykov ICCV, 2005

- Define a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. $\mathcal{V} = \{\mathcal{S}, u, t\}$

$$c_{ws} = \begin{cases} \lambda_s & \lambda_s > 0 \\ -\lambda_s & \lambda_s < 0 \end{cases}, \qquad c_{sr} = \beta_{(s,r)}$$

- Define partition $B = \{u\} \bigcup \{s : X_s = 1\}$, $W = \{t\} \bigcup \{s : X_s = 0\}$ and cut $C(X) = \sum_{s\in B} \sum_{r\in W} c_{rs}$.
- It can be proved $C(x) = \log P(X|Y) + \text{const}$. In words, finding a min-cut is equivalent to find the minimum of posterior $P(X|Y)$.
- Ford-Fulkerson algorithm and Push-Relabeling method can be used to find such a cut quickly.

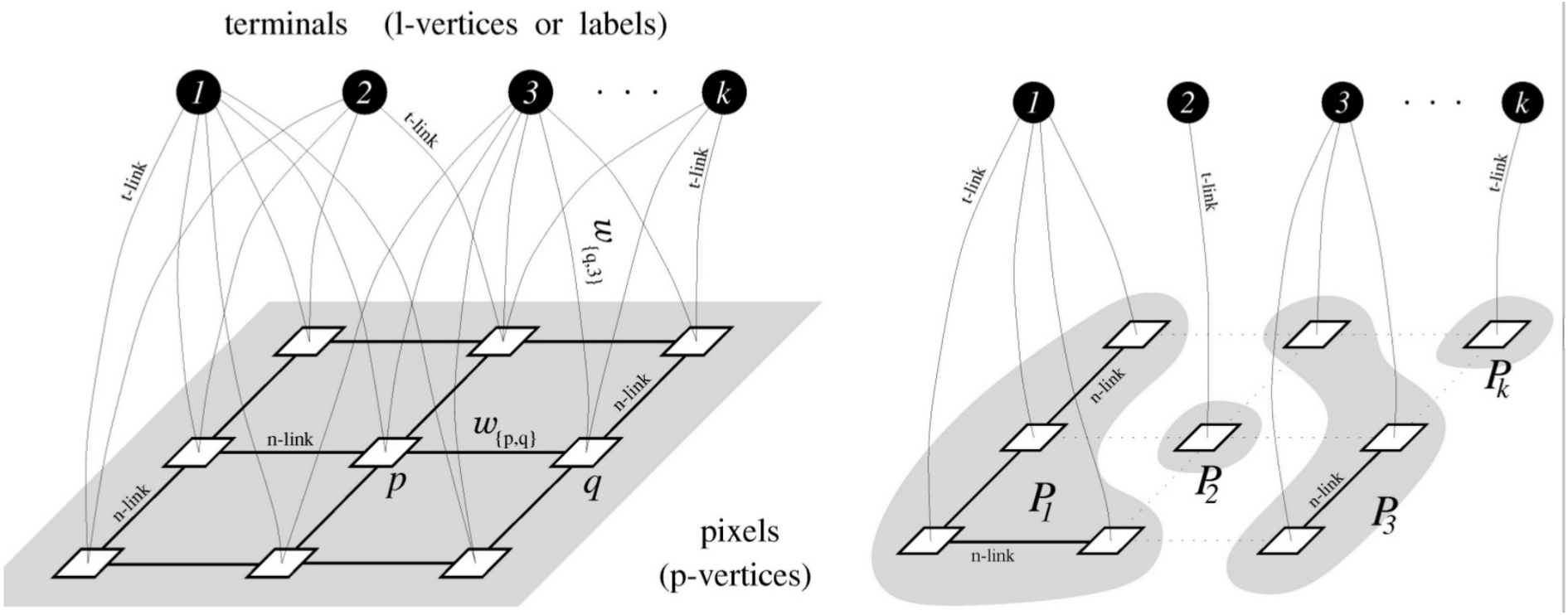
Graph Cuts for Multi-Labeling



From Left 1. Initial image. 2. standard move (ICM), 3. strong moves of alpha-beta swap. 4. strong moves of alpha expansion. (Boykov 2002).

- Convert the multi-labeling problem to 2-labeling problem by $\alpha - \beta$ swap and α expansion.
- Approximation method, but with strong sense of local minima.
- Answer questions like: if results is not good, is that due to bad modeling or bad optimization algorithm?

Graph Cuts for Multi-Labeling



- For label $\{\alpha, \beta\} \in \mathcal{L}$
  - Find $\hat{X} = \text{argmin} E(X')$ among X' within one $\alpha - \beta$ swap of X.
  - If $E(\hat{X} < E(X)$, accept $\hat{X}$
- Repeat above step for all pair of labels $\{\alpha, \beta\}$.

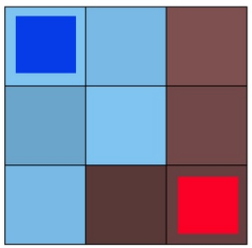
Graph Cuts Pros and Cons

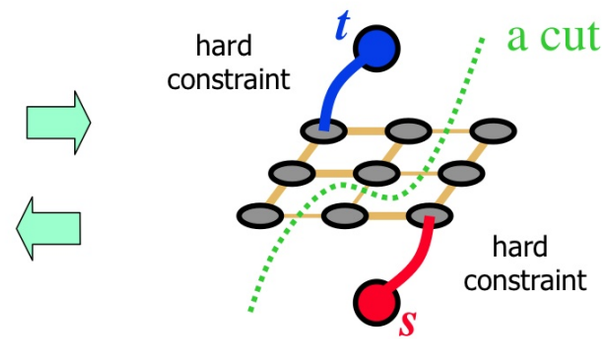
Pros:

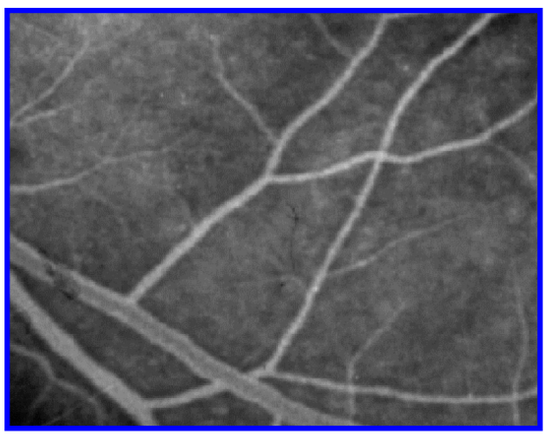
- Break the multi-cut problem to a sequence of binary $s-t$ cuts by $\alpha-\beta$ swap and α expansion.
- Approximation method, but with strong sense of local minima.
- Easy to add hard constraints.
- Answer questions like: if results is not good, is that due to bad modeling or bad optimization algorithm?
- Parallel algorithm $\Rightarrow$ Push-Relabeling algorithm.

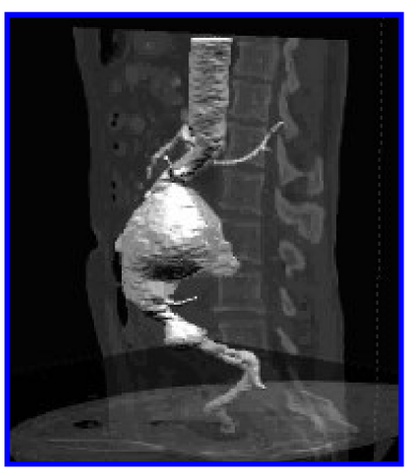
Cons:

- minimize boundary $\Rightarrow$ tends to fail for structures that are not blob shape, like vessels,





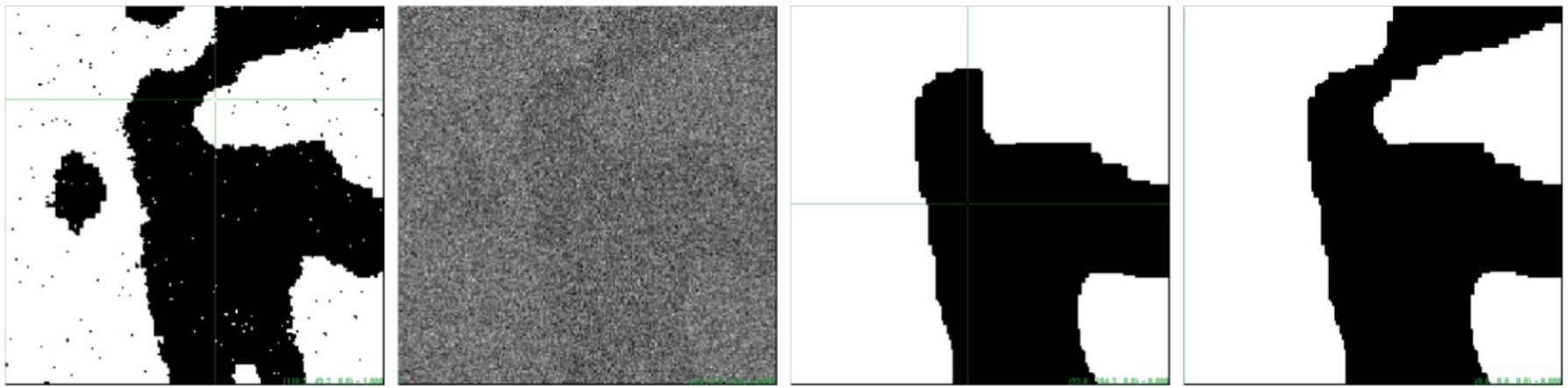




Vessels and aneurism. (kolmogorov, ICCV 2006)

MRF Parameter Estimation

- Correct model and correct parameters $\Rightarrow$ good result.
- Correct model, and incorrect parameters $\Rightarrow$ bad result.



(a) true (b) with noise (c) recovered, $\beta = 0.7$ (d) recovered, $\beta = 0.5$

(e) recovered, $\beta = 0.4$ (f) recovered, $\beta = 0.3$ (g) recovered, $\beta = 0.2$ (h) recovered, $\beta = 0.15$

MRF Parameter Estimation

`Problem 1:`

Given data $X \sim MRF$, we assume model $\mathcal{M}$ with unknown parameter set θ.

Goal: Estimate θ.

`Problem 2:`

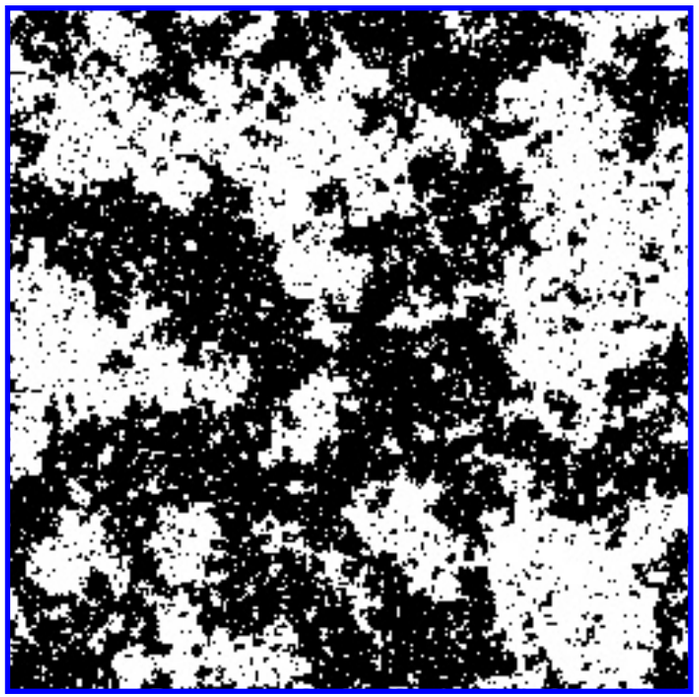
Given noised data Y, we assume model $\mathcal{M}$ with unknown parameter set θ.

Goal: Estimate θ and hidden MRF X simultaneously.

Problem 2 is significantly harder and for now we focus on problem 1.

Given an image shown on the right and suppose we know it is generated from Ising model

$$P(X) = \frac{1}{Z} \exp\{-\beta \sum_{(r,s)\in\mathcal{C}_2} X_r X_s\}.$$



Question: what is the is best estimation of β?

MRF Parameter Estimation

- Least square estimation
- Pseudo-likelihood
- Coding method

Least Square Estimation

For Ising model,

- $U(X) = -\beta \sum_{(r,s)\in\mathcal{C}_2} X_r X_s, \quad P(X) = \frac{1}{Z}\exp\{-U(X)\}, \quad P(X_s|X_{\mathcal{N}_s}) = \frac{\exp\{\beta X_s \sum_{r\in\mathcal{N}_s} X_r\}}{2\cosh(\beta \sum_{r\in\mathcal{N}_s} X_r)}$
- The ratio of observed states

$$\log\left(\frac{P(Xs=1|X_{\mathcal{N}_s})}{P(Xs=0|X_{\mathcal{N}_s})}\right) = 2\beta \sum_{r\in\mathcal{N}_s} X_r$$

- For each set of neighboring pixel value $\mathcal{N}_s$, we compute
  - The observed rate of $\log\left(\frac{P(Xs=1|X_{\mathcal{N}_s})}{P(Xs=0|X_{\mathcal{N}_s})}\right)$
  - The value of $\sum_{r\in\mathcal{N}_s} X_r$.
- We have a est of over-determined linear equations and β can be solved with standard least square method.
- Easy implementation.

Pseudo-likelihood

- Review ML estimation.
- ML estimation of θ: $\theta = \text{argmax} P(X;\theta) = \text{argmax} \frac{1}{Z(\theta)} \exp\{U(X;\theta)\}$. Intractable $Z(\theta)$
- Pseudo-likelihood:

$$PL(X) = \prod_{s \in \mathcal{S}} P(X_s | X_{\mathcal{N}_s})$$

  does not have $Z(\theta)$ anymore.
- Solve θ by standard method, $\frac{\partial \ln PL(X;\theta)}{\partial \theta} = 0$
- For full Bayesian, if we know $P(\theta)$, the estimation is

$$\hat{\theta} = \arg\max P(\theta|X) \propto P(\theta) \cdot P(X|\theta)$$

Least Square Estimation

For Ising model,

- $U(X) = -\beta \sum_{(r,s)\in\mathcal{C}_2} X_r X_s, \quad P(X) = \frac{1}{Z}\exp\{-U(X)\}, \quad P(X_s|X_{\mathcal{N}_s}) = \frac{\exp\{\beta X_s \sum_{r\in\mathcal{N}_s} X_r\}}{2\cosh(\beta \sum_{r\in\mathcal{N}_s} X_r)}$
- The ratio of observed states

$$\log\left(\frac{P(Xs=1|X_{\mathcal{N}_s})}{P(Xs=0|X_{\mathcal{N}_s})}\right) = 2\beta \sum_{r\in\mathcal{N}_s} X_r$$

- For each set of neighboring pixel value $\mathcal{N}_s$, we compute
  - The observed rate of $\log\left(\frac{P(Xs=1|X_{\mathcal{N}_s})}{P(Xs=0|X_{\mathcal{N}_s})}\right)$
  - The value of $\sum_{r\in\mathcal{N}_s} X_r$.
- We have a est of over-determined linear equations and β can be solved with standard least square method.
- Easy implementation.