.

Cluster Sampling and Its Applications in Image Analysis

Adrian Barbu[2] and Song-Chun Zhu[1,2]

(Selected sessions)

1 Background: Potts, SW, and interpretations

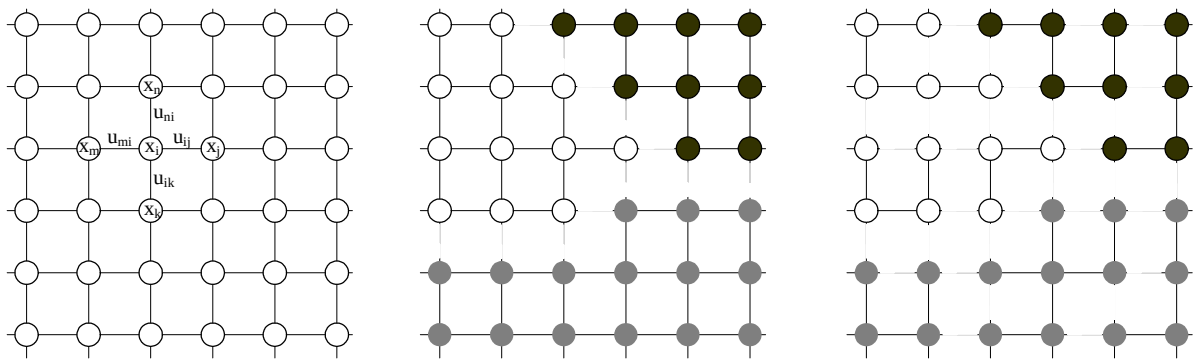
In this section, we review the Potts model, SW method and its two interpretations. The review is made concrete enough so that important results can be followed.

1.1 SW on Potts model

Let $\mathbf{G} =< V, E >$ be an adjacency graph, such as a lattice with 4 nearest neighbor connections. Each vertex $v_i \in V$ has a state variable x_i with finite number of labels (or colors), $x_i \in \{1, 2, ..., \mathrm{L}\}$. The total number of label L is pre-defined, and the Potts model for a homogeneous Markov field is,

$$\pi_{\mathrm{PTS}}(\mathbf{X}) = \frac{1}{Z} \exp\{\beta \sum_{<i,j>\in E} \mathbf{1}(x_i = x_j)\}. \quad (1)$$

$\mathbf{1}(x_i = x_j)$ is a Boolean function. It is equal to 1 if its condition $x_i = x_j$ is observed, and is 0 otherwise. In more general cases, $\beta = \beta(v_i, v_j)$ may be position dependent. Usually we consider $\beta > 0$ for a ferro-magnetic system which prefers same colors for neighboring vertices. The Potts models and its extensions are used as *a priori* probabilities in many Bayesian inference tasks.



(a) G_0 (b) G (c) CCP

Figure 1: Illustating the SW method. (a) An adjacency graph $\mathbf{G}$ and each edge $< i, j >$ is augmented with a binary variable $\mu_{ij} \in \{1, 0\}$. (b) A labeling of the Graph $\mathbf{G}$ where the edges connecting vertices of different colors are removed. (c). A number of connected component after turning off some edges in (b) probabilistically.

As Fig.1.(a) illustrates, the SW method introduces a set of auxiliary variables on the edges.

$$\mathbf{U} = \{\mu_{ij} : \ \mu_{ij} \in \{0, 1\}, \ \forall < i, j > \in E\}. \tag{2}$$

The edge $< i, j >$ is disconnected (or turned off) if and only if $\mu_{ij} = 0$. μ_{ij} follows a Bernoulli distribution conditioning on x_i, x_j.

$$\mu_{ij}|(x_i, x_j) \sim \text{Bernoulli}(\rho \mathbf{1}(x_i = x_j)), \quad \rho = 1 - e^{-\beta}. \tag{3}$$

$\mu_{ij} = 1$ with probability ρ if $x_i = x_j$, and $\mu_{ij} = 1$ with probability 0 if $x_i \neq x_j$. The SW method iterates two steps.

1. The clustering step. Given the current state $\mathbf{X}$, it samples the auxiliary variables in $\mathbf{U}$ according to eqn. (3). It first turns off all edges $< i, j >$ deterministically if $x_i \neq x_j$, as Fig.1.(b) shows. Then it turns off the remain edges with probability ρ. The edge $< i, j >$ is divided into the "on" and "off" sets respectively depending on $\mu_{ij} = 1$ or 0.

$$E = E_{\text{on}}(\mathbf{U}) \cup E_{\text{off}}(\mathbf{U}). \tag{4}$$

The edges in $E_{\text{on}}(\mathbf{U})$ form a number of connected components shown in Fig. 1.(c). We denote all the connected components given $E_{\text{on}}(\mathbf{U})$ by,

$$\text{CP}(\mathbf{U}) = \{\text{cp}_i : \ i = 1, 2, ..., K, \ \text{with} \sum_{i=1}^{K} \text{cp}_i = V\}. \tag{5}$$

Vertices in each connected component cp_i have the same color.

2. The flipping step. It selects one connected component cp $\in$ CP at random and assign a common color y to all vertices in cp. y follows a uniform probability,

$$x_i = y \ \forall v_i \in cp, \quad y \sim \text{unif}\{1, 2, ..., \text{L}\}. \tag{6}$$

In this step, one may choose to repeat the random color flipping for all the connected components in CP($\mathbf{U}$) independently, as they are decoupled given the edges in $E_{\text{on}}(\mathbf{U})$.

In one modified version by Wolff (1989), one may choose a vertex $v \in V$ and grow a connected component following the Bernoulli trials on edges around v. This saves some computation in the clustering step, and thus bigger components have higher chance to be selected.

1.2 SW Interpretation 1: data augmentation and RCM

The SW method described above is far from what was presented in the original paper (Swendsen and Wang 1987). Instead our description follows the interpretation by Edward and Sokal (1988), who augmented the Potts model to a joint probability for both $\mathbf{X}$ and $\mathbf{U}$,

$$\begin{aligned} p_{\text{ES}}(\mathbf{X}, \mathbf{U}) &= \frac{1}{Z} \prod_{<i,j>\in E} [(1-\rho)\mathbf{1}(\mu_{ij} = 0) + \rho \mathbf{1}(\mu_{ij} = 1) \cdot \mathbf{1}(x_i = x_j)] & (7) \\ &= \frac{1}{Z} [(1-\rho)^{|E_{\text{off}}(\mathbf{U})|} \cdot \rho^{E_{\text{on}}(\mathbf{U})}] \cdot \prod_{<i,j>\in E_{\text{on}}(\mathbf{U})} \mathbf{1}(x_i = x_j). & (8) \end{aligned}$$

The second factor $\prod_{<i,j>\in E_{\text{on}}(\mathbf{U})} \mathbf{1}(x_i = x_j)$ is in fact a hard constraint on $\mathbf{X}$ and $\mathbf{U}$. Let the space of $\mathbf{X}$ be

$$\Omega = \{1, 2, ..., \text{L}\}^{|V|}. \tag{9}$$

Under this hard constraint, the labeling $\mathbf{X}$ is reduced to a quotient space $\frac{\Omega}{\text{CP}(\mathbf{U})}$ where each connected component must have the same label,

$$\prod_{<i,j>\in E_{\text{on}}(\mathbf{U})} \mathbf{1}(x_i = x_j) = \mathbf{1}(\mathbf{X} \in \frac{\Omega}{\text{CP}(\mathbf{U})}). \tag{10}$$

The joint probability $p_{\text{ES}}(\mathbf{X}, \mathbf{U})$ observes two nice properties, and both are easy to verify.

Proposition 1 *The Potts model is a marginal probability of the joint probability,*

$$\sum_{\mathbf{U}} p_{\mathrm{ES}}(\mathbf{X}, \mathbf{U}) = \pi_{\mathrm{PTS}}(\mathbf{X}). \tag{11}$$

The other marginal probability is the random cluster model π_{RCM}*,*

$$\sum_{\mathbf{X}} p_{\mathrm{ES}}(\mathbf{X}, \mathbf{U}) = \pi_{\mathrm{RCM}}(\mathbf{U}) = \frac{1}{Z}(1-\rho)^{|E_{\mathrm{off}}(\mathbf{U})|} \cdot \rho^{E_{\mathrm{on}}(\mathbf{U})} \mathrm{L}^{|\mathrm{CP}(\mathbf{U})|}. \tag{12}$$

Proposition 2 *The conditional probabilities of* $p_{\mathrm{ES}}(\mathbf{X}, \mathbf{U})$ *are*

$$p_{\mathrm{ES}}(\mathbf{U}|\mathbf{X}) = \prod_{<i,j>\in E} p(\mu_{ij}|x_i, x_j), \quad \text{with } p(\mu_{ij}|x_i, x_j) = \mathrm{Bernoulli}(\rho \mathbf{1}(x_i = x_j)), \tag{13}$$

$$p_{\mathrm{ES}}(\mathbf{X}|\mathbf{U}) = \mathrm{unif}[\frac{\Omega}{\mathrm{CP}(\mathbf{U})}] = (\frac{1}{\mathrm{L}})^{|\mathrm{CP}(\mathbf{U})|} \text{ for } \mathbf{X} \in \frac{\Omega}{\mathrm{CP}(\mathbf{U})}; \ = 0 \ \text{ otherwise.} \tag{14}$$

Therefore the two SW steps can be viewed as sampling the two conditional probabilities.

1. Clustering step: $\mathbf{U} \sim p_{\mathrm{ES}}(\mathbf{U}|\mathbf{X})$, i.e. $\mu_{ij}|(x_i, x_j) \sim \mathrm{Bernoulli}(\rho \mathbf{1}(x_i = x_j))$.

2. Flipping step: $\mathbf{X} \sim p_{\mathrm{ES}}(\mathbf{U}|\mathbf{X})$, i.e. $\mathbf{X}(\mathrm{cp}_i) \sim \mathrm{Unif}\{1, 2, ..., \mathrm{L}\}$, $\forall \mathrm{cp}_i \in \mathrm{CP}(\mathbf{U})$.

As $(\mathbf{X}, \mathbf{U}) \sim p_{\mathrm{ES}}(\mathbf{X}, \mathbf{U})$, discarding the auxiliary variables $\mathbf{U}$, we have $\mathbf{X}$ following the marginal of $p_{\mathrm{ES}}(\mathbf{X}, \mathbf{U})$. The goal is achieved,

$$\mathbf{X} \sim \pi_{\mathrm{PTS}}(\mathbf{X}). \tag{15}$$

The beauty of this data augmentation method (Tanner and Wong 1987) is that the labeling of the connected components are completely decoupled (independent) given the auxiliary variables. As $\rho = 1 - e^{-\beta}$, it tends to choose smaller clusters if the temperature ($T \propto \frac{1}{\beta}$) in the Potts model is high, and in low temperature it chooses large clusters. So it can overcome the coupling problem with single site Gibbs sampler.

1.3 SW Interpretation 2: slice sampling and decoupling

In the presence of external field (data), the SW method can be interpreted and extended by the auxiliary method proposed by Higdon (1998). Suppose we write the target probability in a more general form,

$$\pi(\mathbf{X}) = \frac{1}{Z} \prod_{v_i \in V} \phi_i(x_i) \cdot \prod_{<i,j>\in E} \psi(x_i, x_j), \quad \phi() > 0, \psi() > 0. \tag{16}$$

For the Potts model above, we have $\psi(x_i, x_j) = e^{\beta \mathbf{1}(x_i = x_j)}$. Higdon (1998) introduced a continuous variable on the edges as the *bond strength*,

$$W = \{\omega_{ij} : \ \omega_{ij} \in [0, +\infty), \forall < i, j > \in E\} \tag{17}$$

In contrast to the Bernoulli probability for the binary variable μ_{ij} in eqn. (3), the bond variables follow uniform probabilities, depending on $\mathbf{X}$,

$$\omega_{ij}|(x_i, x_j) \sim \text{Unif}[0, \psi(x_i, x_j)] = \psi^{-1}(x_i, x_j)\mathbf{1}(0 \leq \omega_{ij} \leq \psi(x_i, x_j)). \tag{18}$$

Thus a conditional probability is constructed as

$$p_{\text{HGD}}(W|\mathbf{X}) = \prod_{<i,j>\in E} p(\omega_{ij}|x_i, x_j) = \prod_{<i,j>\in E} \psi^{-1}(x_i, x_j)\mathbf{1}(0 \leq \omega_{ij} \leq \psi(x_i, x_j)). \tag{19}$$

This formula is chosen to cancel the internal field in a joint probability,

$$p_{\text{HGD}}(\mathbf{X}, W) = \pi(\mathbf{X})p(W|\mathbf{X}) = \frac{1}{Z}[\prod_{v_i \in V} \phi_i(x_i)] \cdot [\prod_{<i,j>\in E} \mathbf{1}(0 \leq \omega_{ij} \leq \psi(x_i, x_j))]. \tag{20}$$

We have the second conditional probability by the Bayes rule,

$$p_{\text{HGD}}(\mathbf{X}|W) = \frac{1}{Z'}[\prod_{v_i \in V} \phi_i(x_i)] \cdot [\prod_{<i,j>\in E} \mathbf{1}(0 \leq \omega_{ij} \leq \psi(x_i, x_j))] \tag{21}$$

That is, given the bond strength ω_{ij}, x_i and x_j must achieve higher probability factor so that the condition $\psi(x_i, x_j) \geq \omega_{ij}$ is observed. This idea is called "slice sampling". In case of the Potts model, this becomes,

$$p(\mathbf{X}|W) = \frac{1}{Z'}[\prod_{v_i \in V} \phi_i(x_i)] \cdot [\prod_{<i,j>\in E} \mathbf{1}(0 \leq \omega_{ij} \leq e^{\beta \mathbf{1}(x_i = x_j)}] \tag{22}$$

Given W, the second product imposes a hard constraint on $\mathbf{X}$. If $\omega_{ij} \leq 1$, $\mathbf{1}(0 \leq \omega_{ij} \leq e^{\beta \mathbf{1}(x_i = x_j)}) = 1$ is satisfied for any x_i, x_j, because $\beta > 0$ and $e^{\beta \mathbf{1}(x_i = x_j)} \geq 1$. Thus it imposes no constraints on x_i, x_j. If $\omega_{ij} > 1$, then it imposes the constraint that $x_i = x_j$. Thus the auxiliary variables μ_{ij} and ω_{ij} are linked by the following equation,

$$\mu_{ij} = \mathbf{1}(\omega_{ij} > 1), \ \ \forall < i, j > \in E. \tag{23}$$

Thus we have to turn on the edges if $\omega_{ij} > 1$, otherwise we turn it off.

$$E_{\text{on}}(W) = \{e =< ij >: \ \omega_{ij} > 1, < i, j > \in E\}. \tag{24}$$

Given W, we have the set of connected components and the vertices in each component receive the same color.

$$\mathrm{CP}(W) = \{\mathrm{cp}_k : k = 1, 2, ..., K, \cup_{i=1}^{K} \mathrm{cp}_k = V\}. \tag{25}$$

As the hard constraints are absorbed by the connected component, the conditional probability in eqn. (22) becomes

$$p_{\mathrm{HGD}}(\mathbf{X}|W) = \prod_{k=1}^{K} \prod_{v_i \in \mathrm{cp}_k} \phi_i(x_i). \tag{26}$$

As we can see, the coloring of each connected component is independent of other vertices (completely decoupled !). In the special case when $\phi_i(x_i) = 1$, it reduces to the RCM model in the previous subsection.

In summary $p_{\mathrm{HGD}}(\mathbf{X}, W)$, like $p_{\mathrm{ES}}(\mathbf{X}, \mathbf{U})$ in eqn.(7), has marginal probability being the target $\pi(\mathbf{X})$ and has two conditional probabilities that are easy to sample. There are two problems with this design.

Firstly, although the decoupling idea with conditional probability $p_{\mathrm{HGD}}(W|\mathbf{X})$ in eqn. (21) is valid for any pair clique Markov random field models and thus goes beyond the Potts model, the hard constraints may become impractical to compute for non-Potts model. That is, given W, the constraint conditions on $\mathbf{X}$ are no longer expressed as clustering. Many slice sampling methods suffer from this problem.

Secondly, although the flipping step in eqn.(26) makes use of the data, the clustering step in eqn. (19) does not. It is similar to the original SW method. This in practice often make the formed cluster ineffective.

1.4 SW Interpretation 3: the Metropolis-Hastings perspective

Proposition 3 *If we set the edge probability to a constant $q_{ij} = 1 - e^{-\beta}$, then*

$$\frac{q(R|\mathbf{X})}{q(R|\mathbf{X}')} = \frac{\prod_{<i,j> \in \mathcal{C}(R,V_\ell)}(1 - q_{ij})}{\prod_{<i,j> \in \mathcal{C}(R,V_{\ell'})}(1 - q_{ij})} = \exp\{\beta(|\mathcal{C}(R, V_{\ell'})| - |\mathcal{C}(R, V_\ell)|)\}, \tag{27}$$

where $|\mathcal{C}|$ is the cardinality of the set.

As $\mathbf{X}$ and $\mathbf{X}'$ only differ in labeling R, the potentials for the Potts model only differs at the "cracks" between R and V_ℓ and $V_{\ell'}$ respectively.

Proposition 4 *For the Potts model* $\pi(\mathbf{X}) = p_o(\mathbf{X}) = \pi_{\text{PTS}}(\mathbf{X})$,

$$\frac{\pi_{\text{PTS}}(\mathbf{X}_R = \ell' | \mathbf{X}_{\partial R})}{\pi_{\text{PTS}}(\mathbf{X}_R = \ell | \mathbf{X}_{\partial R})} = \exp\{\beta(|\mathcal{C}(R, V_\ell)| - |\mathcal{C}(R, V_{\ell'})|)\} \tag{28}$$

Therefore, following eq. (**??**) (where the proposal probabilities for the labels are uniform), the acceptance probability for the Potts model is always one, due to cancellation.

$$\alpha(\mathbf{X} \to \mathbf{X}') = 1. \tag{29}$$

Therefore the third acceptance step is always omitted. This interpretation is related to the Wolff (1989) modification (see also Liu 2001, p157).

References

[1] Barbu, A. and Zhu, S.C. (2003). "Graph partition by Swendsen-Wang cuts", *Proc. Int'l Conf. on Computer Vision*, Nice, France.

[2] Barbu, A. and Zhu, S.C. (2004). "Multigrid and multi-level Swendsen-Wang cuts for hierarchic graph partition", *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Washington DC, 2004.

[3] Cooper, C. and Frieze, A. (1999). "Mixing properties of the Swendsen-Wang process in classes of graphs", *Random Structures and Algorithms* **15**, no. 3-4, 242-261.

[4] Edwards, R.G. and Sokal, A.D. (1988). "Generalization of the Fortuin-Kasteleyn-Swendsen-Wang representation and Monte Carlo algorithm", *Phys. Rev. Lett.* **38**, 2009-2012.

[5] Geman, S. and Geman, D. (1984),"Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images", *IEEE Trans. on PAMI* **6**, 721-741.

[6] Hastings, W.K. (1970). "Monte Carlo sampling methods using Markov chains and their applications", *Biometrika* **57**, 97-109.

[7] Higdon, D.M. (1998). "Auxiliary variable methods for Markov chain Monte Carlo with applications", *J. Am. Statist. Assoc.* **93**, 585-595.

[8] Ising, E (1925). "Beitrag zur theorie des ferromagnetismus", *Zeitschrift für Physik* **31**, 253-258.

[9] Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. and Teller, E. (1953). "Equations of the state calculations by fast computing machines", *J. Chem. Physics* **22**, 1087-1091.

[10] Potts, R.B. (1953) "Some generalized order-disorder transformations", *Proceedings of the Cambridge Philosophic Society* **48**, 106-109.

[11] Swendsen, R.H. and Wang, J.S. (1987), "Nonuniversal critical dynamics in Monte Carlo simulations", *Physical Review Letters* **58** no. 2, 86-88.

[12] Tanner, M. A. and Wong, W.H. (1987), "The calculation of posterior distributions by data augmentation (with discussion)", *J. Amer. Stat. Assoc.*, 82(398):528-540.

[13] Wolff, U. (1989). "Collective Monte Carlo updating for spin systems", *Physical Review Letters* **62**, no. 4, 361-364.