IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 24, no.8, August 2002. An early version appeared in CVPR 2000.

# Learning in Gibbsian Fields: How Accurate and How Fast Can It Be?

Song Chun Zhu and Xiuwen Liu

*Abstract*

Gibbsian fields or Markov random fields are widely used in Bayesian image analysis, but learning Gibbs models is computationally expensive. The computational complexity is pronounced by the recent minimax entropy (FRAME) models which use large neighborhoods and hundreds of parameters[22]. In this paper, we present a common framework for learning Gibbs models. We identify two key factors that determine the accuracy and speed of learning Gibbs models: The efficiency of likelihood functions and the variance in approximating partition functions using Monte Carlo integration. We propose three new algorithms. In particular, we are interested in a *maximum satellite likelihood estimator*, which makes use of a set of pre-computed Gibbs models called "satellites" to approximate likelihood functions. This algorithm can approximately estimate the minimax entropy model for textures in seconds in a HP workstation. The performances of various learning algorithms are compared in our experiments.

**Keywords**: Markov random fields, Minimax entropy learning, Texture modeling, Markov chain Monte Carlo, Maximum likelihood estimate, Importance sampling.

S.C. Zhu is with the Departments of Statistics and Computer Science, UCLA, Los Angeles 90045. sczhu@stat.ucla.edu. X.W. Liu is with the Department of Computer Science at Florida State University.

# 1 Introduction

Gibbsian fields or Markov random fields (MRF) are widely used in Bayesian image analysis for characterizing stochastic patterns or prior knowledge, but learning accurate Gibbs models is computationally expensive. In the literature, many learning algorithms are focused on auto-models, such as auto-binomial models[7] and Gaussian Markov random fields[5] whose parameters can be computed analytically. For MRF models beyond the auto-model families, there are three algorithms. The first is a *maximum pseudo-likelihood estimator* (MPLE)[4]. The second is a *stochastic gradient algorithm*[18, 21, 22]. The third is *Markov chain Monte Carlo maximum likelihood estimator* (MCMCMLE)[12, 13, 9]. The second and third methods approximates partition functions by importance sampling techniques — a topic studied extensively in the literature[15, 16, 17, 8]. There are also methods[10, 11] that estimate Gibbs parameters with pre-computed derivatives of log-partition functions. These algorithms were used primarily for learning MRF models with pair cliques, such as Ising models and Potts models. The computational complexity is pronounced by the recent minimax entropy models (FRAME) for texture[22], which involve large neighborhoods (e.g. filters with window sizes of $33 \times 33$ pixels) and large number of parameters (e.g. $100 - 200$).

Recently four algorithms were proposed to speedup minimax entropy learning with mixed success: 1). A minuteman minimax algorithm[6], 2). A variational method[2, 1], 3). A method of learning by diffusion[19], 4). A generalized MPLE (Private communication with Y.N. Wu). Besides the new algorithms, an ensemble equivalence theorem enables the separation of the model selection procedure from the parameter learning step[20].

In this article, we study a common statistical framework for learning the parameters of Gibbs models with an emphasis on computational efficiency. There are two factors that determine the accuracy and speed of these learning algorithms. The first is the efficiency of the formulated likelihood functions measured by the Fisher's information. The second is the approximation of partition functions by importance sampling. This analysis leads to three new learning algorithms: 1). A maximum partial likelihood estimator, 2). A maximum patch likelihood estimator, and 3). A maximum satellite likelihood estimator. Our emphasis will be on the third algorithm which approximates partition functions by a set of pre-computed reference Gibbs models in a similar spirit to [10, 11]. This algorithm can approximately compute Gibbs models in about 10 seconds on a HP workstation.

The paper is organized as follows. Section (2) presents a common framework for Gibbs learning. Section (3) presents three new algorithms. Section (4) demonstrates experiments on texture synthesis. We conclude with a discussion in Section (5).

## 2 Learning in Gibbsian fields — A Common Framework

Let $\mathbf{I}_\Lambda$ be an image defined on a lattice $\Lambda$, and $\mathbf{I}_{\partial\Lambda}$ its boundary conditions. $\partial\Lambda$ is the neighborhood of $\Lambda$. Let $\mathbf{h}(\mathbf{I}_\Lambda|\mathbf{I}_{\partial\Lambda})$ be the feature statistics of $\mathbf{I}_\Lambda$ under boundary conditions $\mathbf{I}_{\partial\Lambda}$. For example, $\mathbf{h}()$ are histograms of filtered images[23]. Without loss of generality, a Gibbs model is of the following form (see [22]),

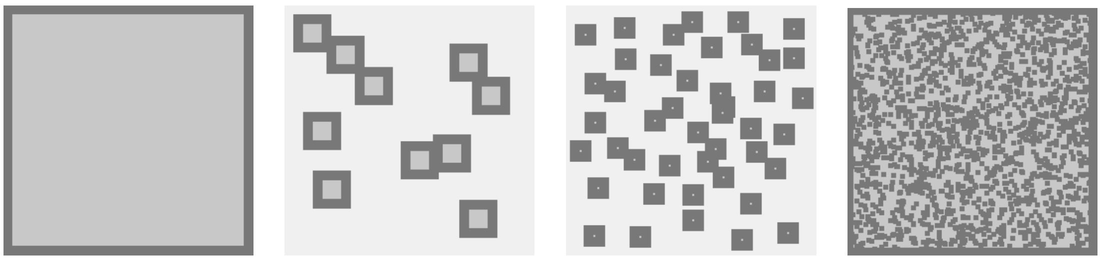
$$p(\mathbf{I}_\Lambda|\mathbf{I}_{\partial\Lambda};\boldsymbol{\beta}) = \frac{1}{Z(\mathbf{I}_{\partial\Lambda},\boldsymbol{\beta})}\exp\{-<\boldsymbol{\beta},\mathbf{h}(\mathbf{I}_\Lambda|\mathbf{I}_{\partial\Lambda})>\}. \quad (1)$$

In equation (1), $\boldsymbol{\beta}$ is a vector valued parameter corresponding to the Julesz ensemble on infinite images[20], and it is invariant to the sizes and shapes of $\Lambda$. So $\boldsymbol{\beta}$ can be estimated on an arbitrary $\Lambda$.

In learning Gibbs models, we are given an observed image $\mathbf{I}_\Lambda^{\text{obs}}$, where $\Lambda$ may have many disconnected components to account for multiple observations. Equally we may break $\Lambda$ into smaller patches $\mathbf{I}_{\Lambda_i}^{\text{obs}}, i = 1,2,...,M$ on lattices $\Lambda_i$ of arbitrary shapes and sizes. These patches may overlap with each other. Then $\boldsymbol{\beta}$ is learned by maximizing a log-likelihood,

$$\boldsymbol{\beta}^* = \arg\max_{\boldsymbol{\beta}} \mathcal{G}(\boldsymbol{\beta}), \quad \text{with } \mathcal{G}(\boldsymbol{\beta}) = \sum_{i=1}^{M}\log p(\mathbf{I}_{\Lambda_i}^{\text{obs}}|\mathbf{I}_{\partial\Lambda_i}^{\text{obs}};\boldsymbol{\beta}). \quad (2)$$

We show that existing Gibbs learning algorithms are unified as ML-estimators and they differ in the following two choices.



a b c d

Figure 1: Various choices of $\Lambda_i$, $i = 1,2,...,M$. The bright pixels are in foreground $\Lambda_i$ which are surrounded by dark background pixels in $\partial\Lambda_i$. a). likelihood, b). patch likelihood (or satellite) likelihood, c). pseudo likelihood, d). partial likelihood.

**Choice I: The number, sizes, and shapes of the *foreground* patches $\Lambda_i, i = 1,...,M$.**

Figure 1 displays four typical choices for $\Lambda_i$. The bright pixels are in the *foreground* $\Lambda_i, i = 1,2,...,M$, which are surrounded by dark pixels in the *background* $\partial\Lambda_i, i =$

$1, 2, ..., M$. In the first three cases, $\Lambda_i$ are square patches with $m \times m$ pixels. In one extreme, Figure 1.a chooses one largest patch denoted by $\Lambda_1$, i.e. $M = 1$ and $m = N - 2w$ with $w$ being the width of the boundary. $\mathcal{G}(\boldsymbol{\beta})$ is called the *log-likelihood*, and it is adopted by the stochastic gradient[21, 22] and MCMCMLE[12, 13, 9] methods. In the other extreme, Figure 1.c chooses the minimum patch size $m = 1$ and $\mathcal{G}(\boldsymbol{\beta})$ is called the *log-pseudo-likelihood*, used in the maximum pseudo-likelihood estimation (MPLE)[4]. Figure 1.b is an example between the two extremes and $\mathcal{G}(\boldsymbol{\beta})$ is called the *log-patch-likelihood*. In the fourth case, Figure 1.d chooses only one ($M = 1$) irregular-shaped patch, denoted by $\Lambda_1$, where $\Lambda_1$ is a set of randomly selected pixels with the rest pixels being the background $\partial\Lambda_1$, and $\mathcal{G}(\boldsymbol{\beta})$ is called the *log-partial-likelihood*. In Figures 1.b and c, a foreground pixel may serve as background in different patches. It is straightforward to prove that maximizing $\mathcal{G}(\boldsymbol{\beta})$ leads to a consistent estimator for all four choices[14].

The flexibility of likelihood function distinguishes Gibbs learning from the problem of estimating partition functions[15, 16, 17]. The latter computes the "pressure" on a large lattice in order to overcome boundary effects.

**Choice II: The reference models used for estimating the partition functions.**

For a chosen foreground and log-likelihood function, the second step is to approximate the partition functions $Z(\mathbf{I}^{\text{obs}}_{\partial\Lambda_i}, \boldsymbol{\beta})$ for each $\Lambda_i, i = 1, ..., M$ by Monte Carlo integration using a reference model at $\boldsymbol{\beta}_o$.

$$
\begin{aligned}
Z(\mathbf{I}^{\text{obs}}_{\partial\Lambda_i}, \boldsymbol{\beta}) &= \int \exp\{- < \boldsymbol{\beta}, \mathbf{h}(\mathbf{I}_{\Lambda_i} | \mathbf{I}^{\text{obs}}_{\partial\Lambda_i}) >\} d\mathbf{I}_{\Lambda_i}, \\
&= Z(\mathbf{I}_{\partial\Lambda_i}, \boldsymbol{\beta}_o) \int \exp\{- < \boldsymbol{\beta} - \boldsymbol{\beta}_o, \mathbf{h}(\mathbf{I}_{\Lambda_i} | \mathbf{I}^{\text{obs}}_{\partial\Lambda_i}) >\} dp(\mathbf{I} | \mathbf{I}^{\text{obs}}_{\partial\Lambda_i}; \boldsymbol{\beta}_o) \\
&\approx \frac{Z(\mathbf{I}^{\text{obs}}_{\partial\Lambda_i}, \boldsymbol{\beta}_o)}{L} \sum_{j=1}^{L} \exp\{- < \boldsymbol{\beta} - \boldsymbol{\beta}_o, \mathbf{h}(\mathbf{I}^{\text{syn}}_{ij} | \mathbf{I}^{\text{obs}}_{\partial\Lambda_i}) >\}.
\end{aligned} \tag{3}
$$

$\mathbf{I}^{\text{syn}}_{ij}, j = 1, 2, ..., L$ are *typical* samples from the reference model $p(\mathbf{I}_{\Lambda_i} | \mathbf{I}^{\text{obs}}_{\partial\Lambda_i}; \boldsymbol{\beta}_o)$ for each patch $i = 1, 2, ..., M$.

Since $\frac{Z(\mathbf{I}^{\text{obs}}_{\partial\Lambda_i}, \boldsymbol{\beta}_o)}{L}, i = 1, 2, ...M$ are independent of $\boldsymbol{\beta}$, we can maximize the estimated log-likelihood $\mathcal{G}(\boldsymbol{\beta})$ by gradient descent. This leads to

$$
\frac{d\boldsymbol{\beta}}{dt} = \sum_{i=1}^{M} \{ \sum_{j=1}^{L} \omega_{ij} \mathbf{h}(\mathbf{I}^{\text{syn}}_{ij} | \mathbf{I}^{\text{obs}}_{\partial\Lambda_i}) - \mathbf{h}(\mathbf{I}^{\text{obs}}_{\Lambda_i} | \mathbf{I}^{\text{obs}}_{\partial\Lambda_i}) \}. \tag{4}
$$

$\omega_{ij}$ is the weight for sample $\mathbf{I}^{\text{syn}}_{ij}$,

$$
\omega_{ij} = \frac{\exp\{- < \boldsymbol{\beta} - \boldsymbol{\beta}_o, \mathbf{h}(\mathbf{I}^{\text{syn}}_{ij} | \mathbf{I}^{\text{obs}}_{\partial\Lambda_i}) >\}}{\sum_{j'=1}^{L} \exp\{- < \boldsymbol{\beta} - \boldsymbol{\beta}_o, \mathbf{h}(\mathbf{I}^{\text{syn}}_{ij'} | \mathbf{I}^{\text{obs}}_{\partial\Lambda_i}) >\}}.
$$

The selection of the reference models $p(\mathbf{I}_{\Lambda_i}|\mathbf{I}^{\text{obs}}_{\partial\Lambda_i};\boldsymbol{\beta}_o)$ depends on the sizes of the patches $\Lambda_i, i = 1, ..., M$. In general importance sampling is only valid when the two distributions $p(\mathbf{I}_{\Lambda_i}|\mathbf{I}^{\text{obs}}_{\partial\Lambda_i};\boldsymbol{\beta}_o)$ and $p(\mathbf{I}_{\Lambda_i}|\mathbf{I}^{\text{obs}}_{\partial\Lambda_i};\boldsymbol{\beta})$ heavily overlap. In one extreme case $m = 1$, the MPLE method[4] selects $\boldsymbol{\beta}_o = 0$ and $p(\mathbf{I}_{\Lambda_i}|\mathbf{I}^{\text{obs}}_{\partial\Lambda_i};\boldsymbol{\beta}_o)$ a uniform distribution. In this case $Z(\mathbf{I}^{\text{obs}}_{\partial\Lambda_i},\boldsymbol{\beta})$ can be computed exactly. In the other extreme case for a large foreground $m = N - 2w$, the stochastic gradient and the MCMCMLE methods have to choose $\boldsymbol{\beta}_o = \boldsymbol{\beta}$ in order to obtain sensible approximations. Thus, both methods must sample $p(\mathbf{I};\boldsymbol{\beta})$ iteratively starting from $\boldsymbol{\beta}_0 = 0$. This is the algorithm adopted in learning the FRAME models[22].

To summarize, Figure 2 illustrates two factors that determine the accuracy and speed of learning $\boldsymbol{\beta}$. These curves are verified through experiments in section 4 (see Figure 7). The horizontal axis is the size of an individual foreground lattice $|\Lambda_i|$.

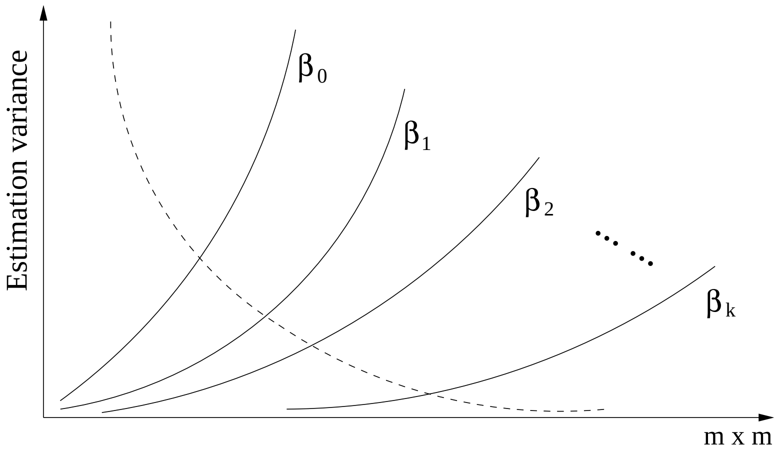



Figure 2: Estimation variances for various selections of patch sizes $m \times m$ and reference models $\boldsymbol{\beta}_o$. The dashed curve shows the inverse Fisher's information which decreases as $m \times m$ increases. The solid curves show the variances in the importance sampling for a sequence of models approaching $\boldsymbol{\beta}$.

1. *The variances of MLE or inverse Fisher information.* Let $\hat{\boldsymbol{\beta}}(\mathbf{I}^{\text{obs}})$ be the estimator maximizing $\mathcal{G}(\boldsymbol{\beta})$ and let $\boldsymbol{\beta}^*$ be the optimal solution. The dashed curve in Figure 2 illustrates the variance,

$$E_f[(\hat{\boldsymbol{\beta}}(\mathbf{I}^{\text{obs}}) - \boldsymbol{\beta}^*)^2],$$

where $f(\mathbf{I})$ is a underlying distribution representing the Julesz ensemble. For choices shown in Figure 1, if we fix the total number of foreground pixels $\sum_{i=1}^{M} |\Lambda_i|$, then the variance (or estimation error) decreases as the patch size (diameter of the hole) increases.

2. *The variance of estimating $Z$ by Monte Carlo integration $E_p[(\hat{Z} - Z)^2]$.* For a given

reference model $\boldsymbol{\beta}_o = \boldsymbol{\beta}_i, i = 1, 2, ..., k$ (see solid curves in Figure 2), this estimation error increases with the lattice sizes. Therefore, for very large patches, such as $m = 200$, we must construct a sequence of reference models to approach $\boldsymbol{\beta}$, $\boldsymbol{\beta}_0 = 0 \to \boldsymbol{\beta}_1 \to \boldsymbol{\beta}_2 \to ..., \to \boldsymbol{\beta}_k \to \boldsymbol{\beta}$. This is the major reason for why the stochastic gradient algorithm was so slow in FRAME[22]!

## 3 Three New Algorithms

The analysis in previous section leads to three new algorithms by selecting likelihoods that trade-off between the two factors, and the third algorithm improve accuracy by pre-computed reference models.

**Algorithm I: Maximizing partial likelihood.**

We choose a lattice shown in Figure 1.d by choosing at random a certain percentage (say 30%) of pixels as foreground $\Lambda_1$ and the rest are treated as background $\Lambda/\Lambda_1$. We define a *log-partial-likelihood*

$$\mathcal{G}_1(\boldsymbol{\beta}) = \log p(\mathbf{I}^{\text{obs}}_{\Lambda_1} | \mathbf{I}^{\text{obs}}_{\Lambda/\Lambda_1}; \boldsymbol{\beta}).$$

Maximizing $\mathcal{G}_1(\boldsymbol{\beta})$ by gradient descent, we update $\boldsymbol{\beta}$ iteratively.

$$\frac{d\boldsymbol{\beta}}{dt} = E_{p(\mathbf{I}_{\Lambda_1} | \mathbf{I}^{\text{obs}}_{\Lambda/\Lambda_1}; \boldsymbol{\beta})}[\mathbf{h}(\mathbf{I}_{\Lambda_1} | \mathbf{I}^{\text{obs}}_{\Lambda/\Lambda_1})] - \mathbf{h}(\mathbf{I}^{\text{obs}}_{\Lambda_1} | \mathbf{I}^{\text{obs}}_{\Lambda/\Lambda_1}). \tag{5}$$

This algorithm follows the same procedure as the original method in FRAME[22]. It trades off between accuracy and speed in a better way than the original algorithm in FRAME[22]. The log-partial-likelihood has lower Fisher information than the log-likelihood, however our experiments demonstrate that it is about 25 times faster than the original minimax learning method without losing much accuracy. We observed that the reason for this speedup is that the original sampling method[22] spends a major portion of its time synthesizing $\mathbf{I}^{\text{syn}}_{\Lambda_1}$ under "non-typical" boundary conditions starting with white noise images. In contrast, the new algorithm works on typical boundary condition $\mathbf{I}^{\text{obs}}_{\Lambda/\Lambda_1}$ where the probability mass of the Gibbs model $p(\mathbf{I}; \boldsymbol{\beta})$ is focused on. The speed appears to be decided by the diameter of the foreground lattice measured by the maximum circle that can fit in the foreground lattice.

**Algorithm II: Maximizing patch likelihood.**

Algorithm II chooses a set of $M$ overlapping patches from $\mathbf{I}^{\text{obs}}_{\Lambda}$ and "dig" a hole $\Lambda_i$ on each patch as Figure 1.b shows. Thus we define a *patch log-likelihood*

$$\mathcal{G}_2(\boldsymbol{\beta}) = \sum_{i=1}^{M} \log p(\mathbf{I}^{\text{obs}}_{\Lambda_i} | \mathbf{I}^{\text{obs}}_{\partial\Lambda_i}; \boldsymbol{\beta}).$$

Maximizing $\mathcal{G}_2(\boldsymbol{\beta})$ by gradient descent, we update $\boldsymbol{\beta}$ iteratively as Algorithm I does.

$$\frac{d\boldsymbol{\beta}}{dt} = \sum_{i=1}^{M} \mathbf{h}(\mathbf{I}^{\text{syn}}_{\Lambda_i} | \mathbf{I}^{\text{obs}}_{\Lambda/\Lambda_i}) - \sum_{i=1}^{M} \mathbf{h}(\mathbf{I}^{\text{obs}}_{\Lambda_i} | \mathbf{I}^{\text{obs}}_{\Lambda/\Lambda_i}). \tag{6}$$

In comparison with algorithm I, the diameters of the lattices are evenly controlled. Algorithm I has similar performance as algorithm I.

**Algorithm III: Maximizing satellite likelihood**

Both algorithms I and II still need to synthesize images on-line, which is a computationally intensive task. Now we propose an third algorithm which may approximately compute $\boldsymbol{\beta}$ in the speed of a few seconds without synthesizing images on-line.

We select a set of reference models in the exponential family $\Omega$ to which the Gibbs model $p(\mathbf{I}; \boldsymbol{\beta})$ belongs,

$$\mathcal{R} = \{p(\mathbf{I}; \boldsymbol{\beta}_j) \ : \ \boldsymbol{\beta}_j \in \Omega, j = 1, 2, ..., s.\}$$

We sample (or synthesize) one large typical image $\mathbf{I}^{\text{syn}}_j \sim p(\mathbf{I}; \boldsymbol{\beta}_j)$ for each reference model off-line. These reference models estimate $\boldsymbol{\beta}$ in $\Omega$ from different "viewing angles". By analogy to the global positioning system, we call the reference models the "satellites".

The *log-satellite-likelihood* is defined as

$$\mathcal{G}_3(\boldsymbol{\beta}) = \mathcal{G}_3^{(1)}(\boldsymbol{\beta}; \boldsymbol{\beta}_1) + \mathcal{G}_3^{(2)}(\boldsymbol{\beta}; \boldsymbol{\beta}_2) + \cdots + \mathcal{G}_3^{(s)}(\boldsymbol{\beta}; \boldsymbol{\beta}_s), \tag{7}$$

where each satellite contributes one log-likelihood approximation,

$$\mathcal{G}_3^{(j)}(\boldsymbol{\beta}; \boldsymbol{\beta}_j) = \sum_{i=1}^{M} \log \frac{1}{\hat{Z}_i^{(j)}} \exp\{- < \boldsymbol{\beta}, \mathbf{h}(\mathbf{I}^{\text{obs}}_{\Lambda_i} | \mathbf{I}^{\text{obs}}_{\partial\Lambda_i}) >\}. \tag{8}$$

Following the importance sampling method in equation (3), we estimate $Z(\mathbf{I}^{\text{obs}}_{\partial\Lambda_i}, \boldsymbol{\beta})$ by Monte Carlo integration.

$$\hat{Z}_i^{(j)} = \frac{Z(\mathbf{I}^{\text{obs}}_{\partial\Lambda_i}, \boldsymbol{\beta}_j)}{L} \sum_{\ell=1}^{L} \exp\{- < \boldsymbol{\beta} - \boldsymbol{\beta}_j, \mathbf{h}(\mathbf{I}^{\text{syn}}_{ij\ell} | \mathbf{I}^{\text{obs}}_{\partial\Lambda_i}) >\}. \tag{9}$$

Notice that for every hole $\Lambda_i$ and for every reference model $p(\mathbf{I}; \boldsymbol{\beta}_j)$, we have a set of $L$ synthesized patches $\mathbf{I}^{\text{syn}}_{ij\ell}$ to fill the hole:

$$H^{\text{syn}}_{ij} = \{\mathbf{I}^{\text{syn}}_{ij\ell}; \ell = 1, 2, ..., L, \forall i, j\}.$$

There are two ways for generating $H^{\text{syn}}_{ij}$.

1. Sampling $\mathbf{I}^{\text{syn}}_{ij\ell} \sim p(\mathbf{I}_{\Lambda_i} | \mathbf{I}^{\text{obs}}_{\partial\Lambda_i}; \boldsymbol{\beta}_j)$ – using the conditional distribution. This is expensive and has to be computed on-line.

2. Sampling $\mathbf{I}^{\rm syn}_{ij\ell} \sim p(\mathbf{I}_{\Lambda_i}; \boldsymbol{\beta}_j)$ – using the marginal distribution. In practice, this is just to fill the holes with randomly selected patches from the synthesized image $\mathbf{I}^{\rm syn}_j$ computed off-line.

In our experiments, we tried both cases and we found that differences are very little for middle sizes $m \times m$ lattices, say $4 \leq m \leq 13$.

Maximizing $\mathcal{G}_3(\boldsymbol{\beta})$ by gradient ascent, we have,

$$\frac{d\boldsymbol{\beta}}{dt} = \sum_{j=1}^{s} \{ \sum_{i=1}^{M} [\sum_{\ell=1}^{L} \omega_{ij} \mathbf{h}(\mathbf{I}^{\rm syn}_{ij\ell} | \mathbf{I}^{\rm obs}_{\partial\Lambda_i}) \ - \ \mathbf{h}(\mathbf{I}^{\rm obs}_{\Lambda_i} | \mathbf{I}^{\rm obs}_{\partial\Lambda_i})] \} \tag{10}$$

$\omega_{ij}$ is the weight for sample $\mathbf{I}^{\rm syn}_{ij\ell}$,

$$\omega_{ij\ell} = \frac{\exp\{- < \boldsymbol{\beta} - \boldsymbol{\beta}_j, \mathbf{h}(\mathbf{I}^{\rm syn}_{ij\ell} | \mathbf{I}^{\rm obs}_{\partial\Lambda_i}) >\}}{\sum_{\ell'=1}^{L} \exp\{- < \boldsymbol{\beta} - \boldsymbol{\beta}_j, \mathbf{h}(\mathbf{I}^{\rm syn}_{ij\ell'} | \mathbf{I}^{\rm obs}_{\partial\Lambda_i}) >\}}$$

Equation (10) converges in the speed of seconds for an average texture model.

However, we should be aware of the risk that the log-satellite-likelihood $\mathcal{G}_3(\boldsymbol{\beta})$ may not be upper bounded. It is almost surely not upper bounded for the MCMCMLE method. This case occurs when $\mathbf{h}(\mathbf{I}^{\rm obs}_{\Lambda_i} | \mathbf{I}^{\rm obs}_{\partial\Lambda_i})$ cannot be described by a linear combination of the statistics of the sampled patches $\sum_{\ell=1}^{L} \omega_{ij} \mathbf{h}(\mathbf{I}^{\rm syn}_{ij\ell} | \mathbf{I}^{\rm obs}_{\partial\Lambda_i})$. When this occurs, $\boldsymbol{\beta}$ does not converge. We handle this problem by including the observed patch $\mathbf{I}^{\rm obs}_{\Lambda_i}$ in $H^{\rm syn}_{ij}$, therefore the satellite likelihood is always upper bounded. Intuitively, let $\mathbf{I}^{\rm syn}_{ij1} = \mathbf{I}^{\rm obs}_{\Lambda_i}$, $\boldsymbol{\beta}$ is learned so that the conditional probabilities $\omega_{ij1} \to 1$ and $\omega_{ij\ell} \to 0$, $\forall \ell \neq 1$. Since $\ell$ is often very large, say $\ell = 20$, adding one extra sample will not screw the sample set.

To summarize, we compare existing algorithms and the newly proposed algorithms from the perspective of estimating $\boldsymbol{\beta}^*$ in $\Omega$, and divide them into three groups. Figure 3 illustrates the comparison where the ellipse stands for the space $\Omega$ and each Gibbs model is represented by a single point.

Group 1: As Figure 3.a illustrates, the maximum likelihood estimators (including stochastic gradient and MCMCMLE) and the maximum partial/patch likelihood estimators generate and sample a sequence of "satellites" $\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_k$ on-line. These satellites get closer and closer to $\boldsymbol{\beta}^*$ (supposed truth). The shadow area around $\boldsymbol{\beta}^*$ represents the uncertainty in computing $\boldsymbol{\beta}$, whose size can be measured by the Fisher information.

Group 2: As Figure 3.c shows, the maximum pseudo likelihood estimator uses a uniform model $\boldsymbol{\beta}_o = 0$ as a "satellite" to estimate any model, and thus has large variance.

Group 3: The maximum satellite likelihood estimators in Figure 3.b use a general set of satellites which are pre-computed and sampled off-line. To save time, one may select a small subset of satellites for computing a given model. One can choose satellites based on

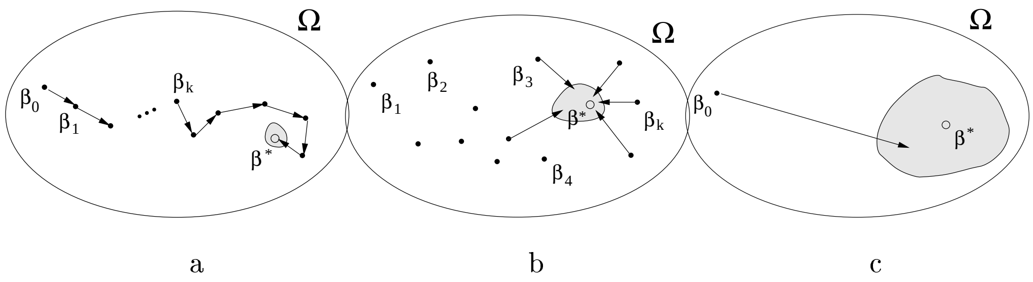



Figure 3: The shadow areas around $\boldsymbol{\beta}^*$illustrates the variance of the estimated $\boldsymbol{\beta}$ or efficiency of the log-likelihood functions. a). Stochastic gradient, and algorithm I and II generate a sequence of satellites on-line to approach $\boldsymbol{\beta}$ closely, $m$ can be small or large. b). The maximum satellite likelihood estimator uses a general set of satellites computed off-line, and can be updated incrementally. This can be used for small size $m$. c). MPLE uses a single satellite: $\boldsymbol{\beta}_o = 0$.

the differences $\mathbf{h}(\mathbf{I}_j^{\text{syn}})$ and $\mathbf{h}(\mathbf{I}^{\text{obs}})$. The smaller the differences are, the closer the satellite is to the estimated model, and thus better approximation. Another criterion is that these satellite should be distributed evenly around $\boldsymbol{\beta}^*$ to obtain good estimation.

# 4 Experiments

In this section, we evaluate the performance of various algorithms in the context of learning Gibbs models for textures. We use 12 filters including an intensity filter, two gradient filters, three Laplacian of Gaussian filters and six Gabor filters at a fixed scale and different orientations. Thus $\mathbf{h}(\mathbf{I})$ includes 12 histograms of filter responses and each histogram has 12 bins. So $\boldsymbol{\beta}$ has $12 \times 11$ free parameters.

We choose 15 natural texture images. For each texture, we use the stochastic gradient algorithm[22] to learn $\boldsymbol{\beta}$ which is treated as ground truth $\boldsymbol{\beta}^*$ for comparison. In this way, we also obtained 15 satellites with 15 synthesized images $\mathbf{I}_j^{\text{syn}}$ computed off-line.

**Experiment I: Comparison of five algorithms.**

In the first experiment, we compare the performance of five algorithms in texture synthesis. Figure 4 demonstrate 6 texture patterns of $128 \times 128$ pixels. For each row, the first column is the synthesized image (ground truth) using a stochastic gradient method used in the FRAME model[22], the other four images are respectively synthesized images using maximum pseudo-likelihood, maximum satellite likelihood, maximum patch likelihood and maximum partial likelihood. For the last three algorithms we fixed the total number of foreground pixels to $5,000$. The patch size is fixed to $5 \times 5$ pixels for patch likelihoods and satellite likelihoods. We select 5 satellites out of the rest 14 pre-computed models for each texture.

Since for different textures the model $p(\mathbf{I}; \boldsymbol{\beta})$ may be more sensitive to some elements of $\boldsymbol{\beta}$ (such as tail bins) than to the rest parameters, and the $\boldsymbol{\beta}$ vectors are highly correlated between its components, it is not very meaningful to compare the accuracy of the learned $\boldsymbol{\beta}$ using an error measure $|\boldsymbol{\beta} - \boldsymbol{\beta}^*|$. Instead we sample each learned model $\mathbf{I}^{\text{syn}} \sim p(\mathbf{I}; \boldsymbol{\beta})$ and compare the histogram errors of the synthesized image against the observed, i.e. $|\mathbf{h}(\mathbf{I}^{\text{syn}}) - \mathbf{h}(\mathbf{I}^{\text{obs}})|$, summed over 12 pairs of histograms each being normalized to 1. The table below shows the errors for each algorithms for the synthesized images in Figure 4. The numbers are subject to some computational fluctuations including the sampling process.

| Fig.4. | FRAME | Pseudo | Satellite | Patch | Partial |
|---|---|---|---|---|---|
| Fig.4.a | (0.816) | (1.308) | (1.845) | (1.279) | (1.844) |
| Fig.4.b | (0.449) | (2.078) | (1.704) | (1.219) | (1.559) |
| Fig.4.c | (0.639) | (2.555) | (1.075) | (1.470) | (1.790) |
| Fig.4.d | (0.243) | (0.784) | (1.286) | (0.364) | (1.069) |
| Fig.4.e | (0.225) | (0.283) | (0.325) | (0.291) | (0.378) |
| Fig.4.f | (0.220) | (1.237) | (0.923) | (1.264) | (0.919) |

The experimental results show that the four algorithms work reasonably well. In comparison the satellite method is often close to the patch and partial likelihood methods. Though it sometimes may yield slightly better results than other methods depending on the similarity between the reference models and the model to be learned. The pseudo-likelihood method can also capture some large image features. In particular, it works well for textures of stochastic nature. For example, on the three textures in Figure 4.d, e. and f.

In terms of computational complexity, the satellite algorithm is the fastest, and it computes $\boldsymbol{\beta}$ in the order of 10 seconds in a HP-workstation. The second fastest is the pseudo-likelihood. It takes a few minutes. However, the pseudo likelihood method consumes a large memory, as it needs to remember all the $k$ histograms for the $g$ grey levels in $N \times N$ pixels. The space complexity is $O(N^2 \times g \times k \times B)$ with $B$ the number of bins. It often needs more than one Gigabyte of memory. The partial likelihood and patch likelihood are very similar to the stochastic gradient algorithm[22]. Since the initial boundary condition is typical, these two estimators take only in general 1/10th of the number of sweeps to convergence. In addition, only a portion of pixels need to be synthesized, which can

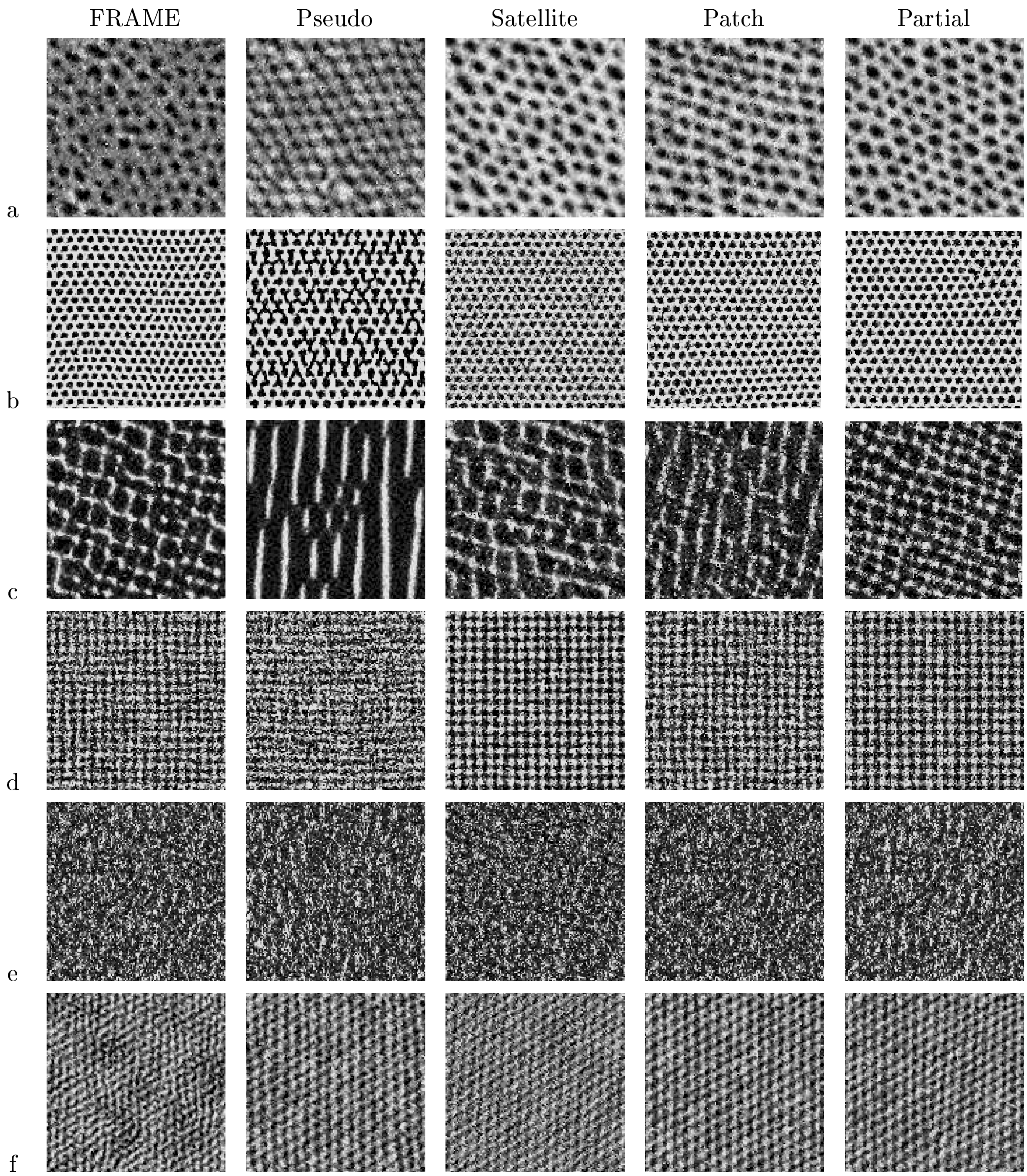



Figure 4: Synthesized texture images using $\boldsymbol{\beta}$ learned from various algorithms. For each column from left to right. 1: Stochastic gradient algorithm as the ground truth, 2: pseudo-likelihood, 3: satellite likelihood, 4: patch likelihood, 5: partial likelihood.

save computation further. The computation time is only about 1/20th of the stochastic gradient algorithm.

**Experiment II: Analysis of the maximum satellite likelihood estimator.**

In the second experiment, we study how the performance of the satellite algorithm is influenced by 1). The boundary condition, and 2). The size of patch $m \times m$.

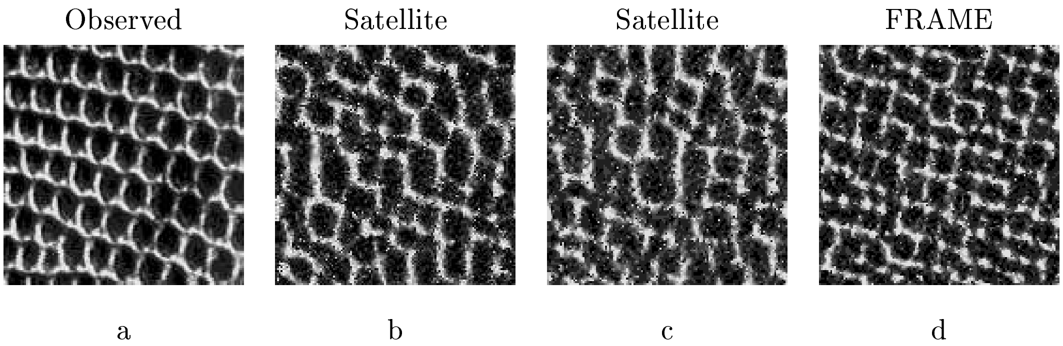



Figure 5: Performance evaluation of the satellite algorithm. (a) Observed texture image. (b) Synthesized image using $\boldsymbol{\beta}$ learned without boundary conditions. (c) Synthesized image using $\boldsymbol{\beta}$ learned with boundary conditions. (d) Synthesized image using $\boldsymbol{\beta}$ learned by stochastic gradient.

I). Influence of boundary conditions.

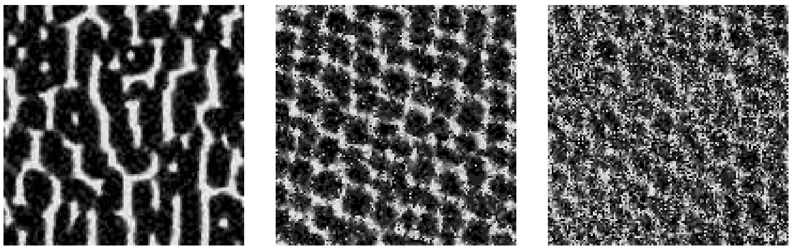
Figure 5.a displays a texture image as $\mathbf{I}^{\text{obs}}$. We run three algorithms for comparison. Figure 5.d is a result from the FRAME (stochastic gradient method). Figure 5.b and c are results using the satellite algorithms. The difference is that Figure 5.c uses observed boundary condition for each patch and does on-line sampling, while Figure 5.b ignores the boundary condition. For all the following results of satellite likelihood method (algorithm III), $H_{ij}^{\text{syn}}$ are generated from the marginal probabilities without on-line sampling.

II). Influences of the hole size $m \times m$.

We fix the total number of foreground pixels $\sum_i |\Lambda_i|$ and study the performance of satellite algorithm with difference hole sizes $m$. Figures 6.a-c show three synthesized images using $\boldsymbol{\beta}$ learned by satellite algorithm with different hole sizes $m = 2, 6, 9$ respectively. It is clear from Figures 6.a-c that the hole size with $6 \times 6$ pixels gives better result.

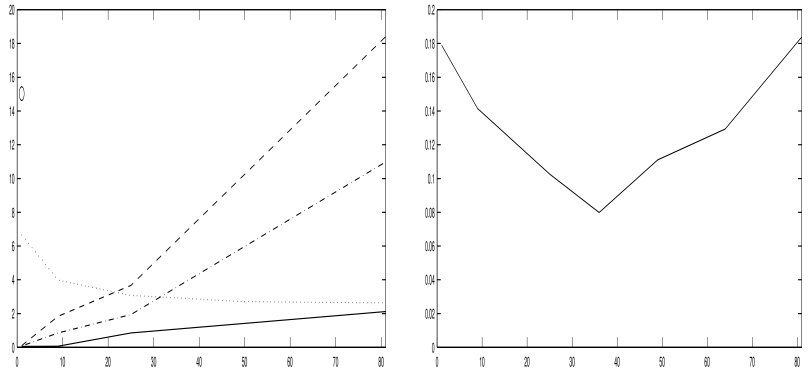
To explain why the hole size of $m = 6$ gives better satellite approximation, we compute the two key factors that determine performance. Figure 7.a shows the numeric results in correspondence to the theoretical analysis displayed in Figure 2.

When the hole size is small, the partition function can be estimated accurately as shown by the small $E_p[(\hat{Z} - Z)^2]$ in solid, dash-dotted, and dashed curves in Figure 7. However,



a b c

Figure 6: Synthesized images using $\boldsymbol{\beta}$ learned by the satellite method with different hole sizes. (a) $m = 2$. (b) $m = 6$. (c) $m = 9$.



a b

Figure 7: The x-axes are the hole size $m^2$. a). Dotted curve is $E_f[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^2]$ plotted against the hole size $m^2$. The solid, dash-dotted, and dashed curves are $E_p[(\hat{Z} - Z)^2]$ for three different reference models. b). Average synthesis error per filter with respect to the hole size $m^2$.

the variance $E_f[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^2]$ is large for small holes, which is shown by the dotted curve in figure 7.a. The optimal choice of the hole size is thus approximately the intersection of the two curves. Since the reference models that we used are close to the dash-dotted line shown in figure 7.a, we predict optimal hole size is between $5 \times 5$ and $6 \times 6$. Figure 7.b shows the average error between the statistics of synthesized image $\mathbf{I}^{\text{syn}} \sim p(\mathbf{I}; \boldsymbol{\beta})$ and the observed statistics err $= \frac{1}{12}|\mathbf{h}(\mathbf{I}^{\text{obs}}) - \mathbf{h}(\mathbf{I}^{\text{syn}})|$, where $\boldsymbol{\beta}$ is learned using the satellite method for $m = 1, 2, ..., 9$. Here the hole size of $6 \times 6$ pixels gives better result.

# 5 Conclusion

To conclude our study, we compare qualitatively ten existing Gibbs learning algorithms in Figure 8 along three factors (or dimensions): I: Accuracy in approximating $\log Z$. II: The diameter of foreground lattices and thus efficiency of the likelihood. Factor III: Computational complexity.

The ten algorithms are: 1). stochastic gradient MLE[22], 2). Maximum pseudo likelihood (MPLE)[3, 4], 3). MCMCMLE[12, 13, 9, 16], 4). Maximum patch likelihood, 5). Maximum partial likelihood, 6). Maximum satellite likelihood, 7). Minuteman minimax[6], 8). Variational method[2, 1], 9). Learning by diffusion[21], 10). Generalized maximum pseudo likelihood (Y.N. Wu, private communication).

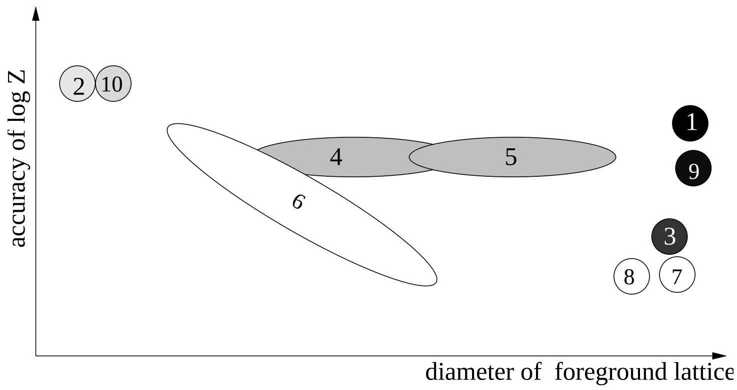



Figure 8: A common framework for learning Gibbs models. The horizontal axis is the size of foreground patches which is proportional to Fisher's information. The vertical axis is the accuracy in estimating $\log Z$. The brightness of the ellipses implies the learning speed, and darker is slower. This figure is intended only for a qualitative comparison.

## Acknowledgments

The project was supported by an NSF grant NSF-9877127 and an NSF Career award IIS-00-92664 to S.C. Zhu. The first author would like to thank Yingnian Wu for valuable discussions.

## References

[1] M. P. Almeida and B. Gidas, “A variational method for estimating the parameters of MRF from complete and incomplete data”, *The Annals of Applied Statistics*, 3:103-136, 1993.

[2] C. H. Anderson and W. D. Langer, “Statistical models of image texture”, *Unpublished preprint*, Washington University, St Louis, MO, 1996.

[3] J. Besag, “Spatial interaction and the statistical analysis of lattice systems (with discussion)”, *J. of the Royal Stat. Soc. B,* vol 36, pp192-236, 1973.

[4] J. Besag, “Efficiency of pseudo-likelihood estimation for simple Gaussian fields, *Biometrika*, **64**: 616-618, 1977.

[5] R. Chellappa and A. K. Jain, *Markov random fields: theory and applications*, Academic Press, 1993.

[6] J. Coughlan, and A. L. Yuille, “Minutemax: a fast approximation for minimax learning”, *Proc. of Neural Information Processing Systems*, Dever, Co. 1998.

[7] G. R. Cross and A. K. Jain, “Markov random field texture models.” *IEEE, PAMI*, **5**, 25-39. 1983.

[8] H. Derin and H. Elliott, “Modeling and segmentation of noisy and textured images using Gibbs random fields”, *IEEE Trans. on PAMI*, vol.9, no.1, pp. 39-55, 1987.

[9] X. Descombes, R. Morris, J. Zerubia, and M. Berthod, “Maximum likelihood estimation of Markov random field parameters using Markov chain Monte Carlo algorithms”, *Int’l Conf, on EMMCVPR*, Venice, Italy, May, 1997.

[10] S. Geman and D. McClure, ”Bayesian images analysis: An application to single photon emission tomography,” *Proc. Statist. Comput. Sect. Amer. Stat. Assoc.*, pp.12-18, Washington. DC, 1985.

[11] S. Geman and D. McClure, ”Statistical methods for tomographic image reconstruction,” *Bull. Int. Stat. Inst.*, vol. LII-4, pp. 5-21, 1987.

[12] C. J. Geyer and E. A. Thompson, “Constrained Monte Carlo maximum likelihood for dependent data”, *J. of Royal Stat. Soc. B*, vol. 54, pp. 657-699, 1992.

[13] C. J. Geyer, "On the convergence of Monte Carlo maximum likelihood calculations", *J. of Royal Stat. Soc. B*, vol. 56, pp. 261-274, 1994.

[14] B. Gidas, "Consistency of maximum likelihood and pseudo-likelihood estimators for Gibbs distributions", in W. Fleming and P.L. Lions (eds.) *Stochastic Differential Systems, Stochastic Control Theory and Applications*, Springer, NY, 1988.

[15] M. Jerrum and A. Sinclair, "Polynomial-time approximation algorithms for the Ising model", *SIAM Journal on Computing*, vol. 22, pp. 1087-1116, 1993.

[16] G. G. Potamianos and J. K. Goutsias, "Partition function estimation of Gibbs random field images using Monte Carlo simulations", *IEEE Trans. on Information Theory*, vol. 39, pp. 1322-1332, 1993.

[17] G. G. Potamianos and J. Goutsias, "Stochastic Approximation Algorithms for Partition Function Estimation of Gibbs Random Fields", *IEEE Transactions on Information Theory*, vol. 43, pp. 1948-1965, 1997.

[18] H. Robbins and S. Monro, "A stochastic approximation method." *Annals of Mathematical Statistics*, 22, 400-407, 1951.

[19] J. Shah, "Minimax entropy and learning by diffusion", *Proc. of CVPR*, Santa Barbara, CA. 1998.

[20] Y. N. Wu, S. C. Zhu, and X. W. Liu, "Equivalence of Gibbs and Julesz ensembles", *Proc. of Int'l Conf. on Computer Vision*, Corfu, Greece, 1999.

[21] L. Younes, "Estimation and annealing for Gibbsian fields", *Annales de l'Institut Henri Poincare, Section B, Calcul des Probabilities et Statistique*, **24**, 269-294, 1988.

[22] S. C. Zhu, Y. Wu, and D. Mumford. "Minimax Entropy Principle and Its Application to Texture Modeling". Neural Computation, Vol. 9, no 8, Nov. 1997.

[23] S. C. Zhu, X. Liu, and Y. Wu, "Exploring Julesz Ensembles by Efficient Markov Chain Monte Carlo", *IEEE Trans. on PAMI*, vol.22, no.6, pp554-569, 2000.