Approximate Maximum Likelihood Hyperparameter Estimation for Gibbs Priors

Zhenyu Zhou, *Member, IEEE,* Richard M. Leahy, *Member, IEEE,* and Jinyi Qi, *Student Member, IEEE*

***Abstract*— The parameters of the prior, the *hyperparameters*, play an important role in Bayesian image estimation. Of particular importance for the case of Gibbs priors is the global hyperparameter, β, which multiplies the Hamiltonian. Here we consider maximum likelihood (ML) estimation of β from *incomplete* data, i.e., problems in which the image, which is drawn from a Gibbs prior, is observed indirectly through some degradation or blurring process. Important applications include image restoration and image reconstruction from projections. Exact ML estimation of β from incomplete data is intractable for most image processing. Here we present an approximate ML estimator that is computed simultaneously with a maximum *a posteriori* (MAP) image estimate. The algorithm is based on a mean field approximation technique through which multidimensional Gibbs distributions are approximated by a separable function equal to a product of one-dimensional (1-D) densities. We show how this approach can be used to simplify the ML estimation problem. We also show how the Gibbs–Bogoliubov–Feynman (GBF) bound can be used to optimize the approximation for a restricted class of problems. We present the results of a Monte Carlo study that examines the bias and variance of this estimator when applied to image restoration.**

I. Introduction

BAYESIAN approaches to inverse problems in image processing typically involve computing a point estimate of an unknown image $\boldsymbol{x} \in \mathcal{X}$ from a set of data $\boldsymbol{y} \in \mathcal{Y}$. We assume that the two quantities are related by a known conditional probability, $P(\boldsymbol{y}|\boldsymbol{x})$. This conditional probability or likelihood function is dependent on the imaging modality and is problem specific. The estimate of $\boldsymbol{x}$ is computed as a function of the posterior density $P(\boldsymbol{x}|\boldsymbol{y})$, which requires the specification of a prior density $P(\boldsymbol{x})$ in addition to the likelihood function. In the context of Bayesian image estimation, the parameters of the prior are referred to as *hyperparameters*. In this paper, we address the problem of estimating these parameters in the case where it is not possible to observe the true image $\boldsymbol{x}$ directly. We describe a practical method for estimating hyperparameters from observations of the data $\boldsymbol{y} \in \mathcal{Y}$. We begin by briefly reviewing two models for image restoration and reconstruction for which this method is applicable.

One of the most widely addressed models in image restoration and reconstruction is the linear Gaussian model $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{n}$, where $\boldsymbol{y}$ is the observed data, $\boldsymbol{x}$ is the underlying image, $\boldsymbol{n}$ is zero-mean Gaussian noise with covariance matrix $\boldsymbol{C}$ and matrix $\boldsymbol{A}$ is a linear degradation operator. Then

$$P(\boldsymbol{y}|\boldsymbol{x}) = (2\pi)^{-N/2}|\boldsymbol{C}|^{-1/2} \cdot \exp\left[-\tfrac{1}{2}(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x})^T\boldsymbol{C}^{-1}(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x})\right]. \quad (1)$$

A second common model is the linear Poisson model, which arises in problems where the data acquisition system is photon limited, e.g., emission tomography, gamma-ray astronomy, and fluorescence microscopy. In this model, the mean of $\boldsymbol{y}$ is related to the image $\boldsymbol{x}$ by a linear operator $\boldsymbol{A}$, i.e., $E[\boldsymbol{y}] = \boldsymbol{A}\boldsymbol{x}$ and $\boldsymbol{y}$ follows a joint Poisson distribution

$$P(\boldsymbol{y}|\boldsymbol{x}) = \prod_i \frac{\left(\sum_j A_{ij}x_j\right)^{y_i}}{(y_i)!} \exp\left(-\sum_j A_{ij}x_j\right). \quad (2)$$

The objective of the inverse problems of interest here, is to obtain a point estimate of $\boldsymbol{x}$ from the observation $\boldsymbol{y}$. Since $\boldsymbol{A}$ is often ill-conditioned, direct inversion based on maximizing the likelihood function does not always provide a unique and stable solution. Bayesian methods solve this type of ill-posed inverse problem by combining information contained in the observed data with prior information concerning the relative probabilities of possible solutions. The unknown image can then be estimated by maximizing over the posterior density $P(\boldsymbol{x}|\boldsymbol{y})$ to form a *maximum a posteriori* (MAP) estimate. The posterior density is proportional to the product of the likelihood function $P(\boldsymbol{y}|\boldsymbol{x})$, and a prior on the image, $P(\boldsymbol{x}|\beta)$. Usually the prior reflects an expectation that images are locally smooth. Markov random fields (MRF's) [3], [10], [24] have been widely used to model local smoothness in images and will be the class of priors considered here. The joint density for the MRF has the form of a Gibbs distribution

$$P(\boldsymbol{x}|\beta) = \frac{1}{Z}\exp\{-\beta U(\boldsymbol{x})\} \quad (3)$$

where $U(\boldsymbol{x})$ is the Gibbs energy function, Z is the partition function, and β is the global hyperparameter.

Image estimation using MRF priors has proven to be a powerful approach to restoration and reconstruction of high-quality images. However, a major problem limiting its utility is the lack of a practical and robust method for selecting the parameters of the prior. Of particular importance for the case of homogeneous isotropic MRF's is the global hyperparameter β, which multiplies the Gibbs energy function. MAP estimates of the image $\boldsymbol{x}$ are clearly functions of β, which controls

Manuscript received June 14, 1995; revised September 18, 1996. This work was supported by the National Cancer Institute under Grant R01-CA59794 and the National Institute of Mental Health under Grant R01-MH53213.

The authors are with the Signal and Image Processing Institute, Department of Electrical Engineering-Systems, University of Southern California, Los Angeles, CA 90089-2564 USA (e-mail: leahy@sipi.usc.edu).

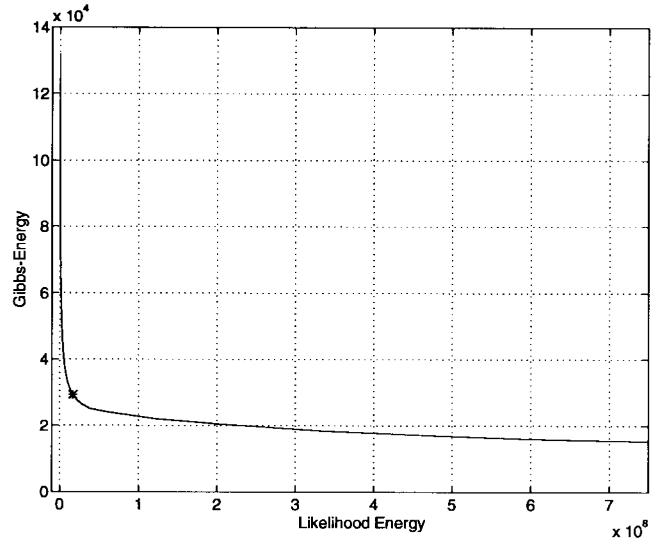
Publisher Item Identifier S 1057-7149(97)03892-X.

the balance of influence of the Gibbs prior and that of the likelihood. If β is too large, the prior will tend to have an over-smoothing effect on the solution. Conversely, if it is too small, the MAP estimate may be unstable, reducing to the ML solution as β goes to zero.

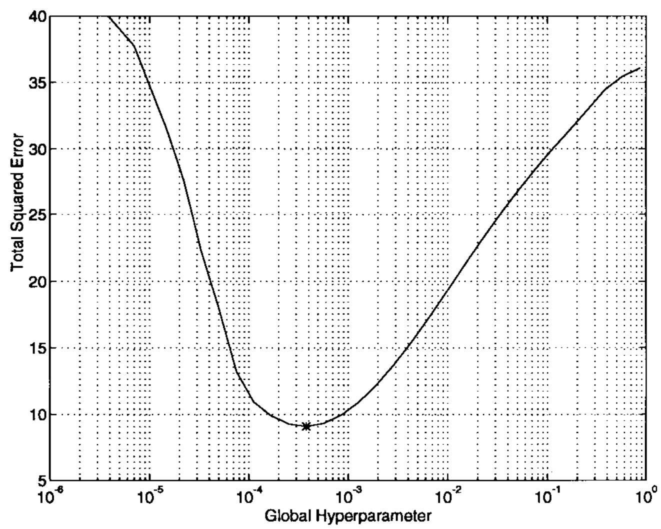
To illustrate the effect of the hyperparameter on the MAP estimate, we show two curves in Fig. 1 computed for a typical application to image restoration. In Fig. 1(a), we have a typical L-curve [15], [40], which is a plot of the value of the Gibbs energy $U(\boldsymbol{x})$ versus the likelihood energy $(\boldsymbol{y}-\boldsymbol{Ax})^T\boldsymbol{C}^{-1}(\boldsymbol{y}-\boldsymbol{Ax})$, computed at the MAP solution for a range of values of β. We observe two characteristic parts on the curve, namely a flat part where the MAP solution is dominated by the prior, and an almost vertical part, where the solution is dominated by the likelihood function. Heuristically, the region between these two characteristic parts, i.e., the "corner," corresponds to a good balance between fidelity to the data and smoothness of the solution. Fig. 1(b) shows a curve of the squared error in the MAP estimate for a range of β values. Here, it is clear that an appropriate choice of β is necessary to achieve a small error. Furthermore, we have observed that the corner of the L-curve corresponds to a value of the hyperparameter β that is close to that which minimizes the squared error for the MAP estimation problem described here (both points are indicated by *). Similar observations were made in [15] concerning the more general regularization problem.

A truly Bayesian formulation requires either that the hyperparameters are known or that we specify a "hyperprior" density. However, in practice the hyperparameters are often unknown because the true images can never be observed directly, and little evidence exists to justify an informative hyperprior density. Even if β is known, problems can arise if there is an unknown gain factor in the transfer function $\boldsymbol{A}$ in (2) or an unknown noise variance in (1). These problems can be avoided if the hyperparameters are estimated directly from the observed data. Data-driven selection of the hyperparameter is often performed in an ad hoc fashion through visual inspection of the resulting images. There are two basic approaches for choosing β in a more principled manner: i) treating β as a regularization parameter and applying techniques such as generalized cross validation, the L-curve, and χ^2 goodness of fit tests; and ii) estimation theoretic approaches such as maximum likelihood (ML).

The generalized cross-validation (GCV) method [9] has been applied in Bayesian image restoration and reconstruction [18]. Several difficulties are associated with this method: The GCV function is often very flat, and its minimum is difficult to locate numerically [34]. Also, the method may fail to select the correct hyperparameter when measurement noise is highly correlated [35]. For problems of large dimensionality, this method may be impractical due to the amount of computation required. Hansen and Leary's L-curve is based on the empirical observation that the corner of the curve, illustrated in Fig. 1(a), corresponds to a good choice of β in terms of other validation measures [15]. The L-curve has similar performance to GCV for uncorrelated measurement errors; however, the L-curve criterion also works, under certain restrictions, for correlated errors [15]. We have used the L-curve to select the



(a)



(b)

Fig. 1. Illustration of the quantitative effect of the global hyperparameter using (a) the L-curve and (b) global mean squared error of restored image versus β. Note that the value of β giving minimum squared error ($*$) corresponds to the corner of the L-curve.

hyperparameter in MAP image reconstruction [40]. The corner of the L-curve is difficult to find without multiple evaluations of the MAP solution for different hyperparameter values. Thus, the computation cost is again very high. χ^2 statistics have been widely used to choose the regularization parameter [33]. For MAP image estimation, Hebert *et al.* [17] developed an adaptive scheme based on a χ^2 statistic to select β. Since the image is estimated from the data, the degrees of freedom of the test should be reduced accordingly. This presents a problem when the data and image are of similar dimension. It has also been observed that χ^2 methods tend to over-smooth the solution [33].

As an alternative to the regularization based methods discussed above, a well-grounded approach to selection of the hyperparameter is to apply ML estimation. The image $\boldsymbol{x}$, which is drawn from the *complete data* sample space $\mathcal{X}$ characterized by the parameter β, is not observed directly. Instead, we observe a second process $\boldsymbol{y}$ which is drawn from the *incomplete data* sample space $\mathcal{Y}$. The ML estimate of the hyperparameter corresponds to the maximizer of the incomplete data likelihood function $P(\boldsymbol{y}|\beta)$, which is found by marginalization of the joint probability density for the complete and incomplete data, $P(\boldsymbol{x}, \boldsymbol{y}|\beta)$, over the complete data sample space. Selection of the hyperparameter can therefore be viewed as an ML estimation problem in an incomplete/complete data framework and is a natural candidate for the expectation maximization (EM) algorithm [8]. However, in most imaging applications, the high dimensionality of the densities involved make the EM approach impractical. Geman and McClure [11] propose using a stochastic relaxation technique, such as a Gibbs sampler, to evaluate the E-step of the EM algorithm. While this approach provides a means of overcoming the intractability of the true EM algorithm, the computational cost remains extremely high. Markov chain Monte Carlo (MCMC) methods [4] have also been used to solve the high dimensional integration involved in ML estimation. Zhang *et al.* [39] and Saquib *et al.* [32] incorporate the MCMC method in EM algorithm to evaluate the expectation function at the E-step. Geyer and Thompson [13] propose a Monte Carlo maximum likelihood method which uses MCMC methods to approximate the likelihood function directly. The major disadvantage of the sampling methods is their high computational cost. Other estimation methods have been studied that do not share the desirable properties of true ML estimation but have much lower computational cost. Several generalized ML approaches have been described [3], [19], [26] that make the simplifying approximation that the ML estimate of β and the MAP estimate of the image $\boldsymbol{x}$ can be found simultaneously as the joint maximizers of the joint density of $\boldsymbol{x}$ and $\boldsymbol{y}$. This approach works well in some situations, but the crudeness of the approximation results in poor performance in general. The method of moments (MOM) [11], [25] defines a statistical moment of the incomplete data that is ideally chosen to reflect the variability in the unobserved image and to establish a one-to-one correspondence between the moment value and the global hyperparameter. Initial computational costs for this method are very large, but the moment versus hyperparameter curve is independent of the observed data and can be computed off line. For each new data set the hyperparameter is determined by simply comparing the computed statistic with the precomputed curve. The major limitation in using this method is in finding a statistic with sufficient slope that the hyperparameter can be reliably determined. In practice, it has been observed that the method performs well only for relatively small values of β [25]. Finally, a variational method is described in [1]. This approach leads to a procedure similar to, but simpler than, the EM algorithm. However, the computational cost remains high, and few validation or experimental results have been published for this method.

Here we return to the ML approach, but develop an approximation that results in a reasonable computational cost. The major difficulty in computing a true ML estimate of β is in evaluating the multidimensional integrals over the complete data sample space $\mathcal{X}$, which occur in either evaluation of the prior and posterior partition functions or the prior and posterior expectation of Gibbs prior energy functions. We thus approximate these Gibbs distributions with simple and separable densities so that the multidimensional integrals become functions of one dimensional integrals. This approximation renders the ML approach tractable. The approximation is closely related to the mean field approximation methods of statistical mechanics. In the mean field approach, the separable approximation is achieved by replacing the statistical influence of the neighbors of each pixel with their estimated means. In our work, we use a mode-field rather than a mean-field approximation, where the mode of the posterior density is computed using a MAP image estimation algorithm. We use a sequential updating scheme to estimate both the image and the hyperparameter. Successive iterates of a MAP image estimation algorithm are substituted in the mode-field approximation, which in turn is used to update the hyperparameter estimate.

We present a brief summary of the problem formulation for ML estimation of the hyperparameter from incomplete data in Section II. We then describe our mean/mode field approach to parameter estimation in Section III. We also describe how an optimal approximation can be found in special cases using the Gibbs–Bogoliubov–Feynman (GBF) bound [7]. We then describe the application of this method to the problem of image restoration in Section IV. Finally, in Section V, we present the results of extensive Monte Carlo studies that examine the bias and variance of this estimator for cases where the true value of β is known.

II. BACKGROUND

A. Gibbs Priors for Image Estimation

We will assume a homogeneous isotropic MRF model for the image, $\boldsymbol{x}$, characterized by the Gibbs distribution

$$P(\boldsymbol{x}|\beta) = \frac{1}{Z(\beta)} \exp\{-\beta U(\boldsymbol{x})\} \tag{4}$$

with Gibbs energy $U(\boldsymbol{x})$, partition function $Z(\beta)$, and global hyperparameter β. The partition function is the scaling constant

$$Z(\beta) = \int_{\mathcal{X}} \exp\{-\beta U(\boldsymbol{x})\}\, d\boldsymbol{x}. \tag{5}$$

Here, we indicate explicitly that the partition function is dependent on the hyperparameter β.

We restrict the Gibbs energy to pairwise interactions on a second-order (eight nearest-neighbor) system as follows:

$$U(\boldsymbol{x}) = \sum_i \sum_{j>i,\, j\in N_i} \kappa_{ij} V(x_i, x_j) \tag{6}$$

where N_i denotes the set of eight nearest neighbors of pixel i with κ_{ij} unity for horizontal and vertical neighbors and $1/\sqrt{2}$ for diagonal neighbors. The image sample space is $\mathcal{X} = [0, x_{\max}]^N$ where N is the total number of pixels in the image.

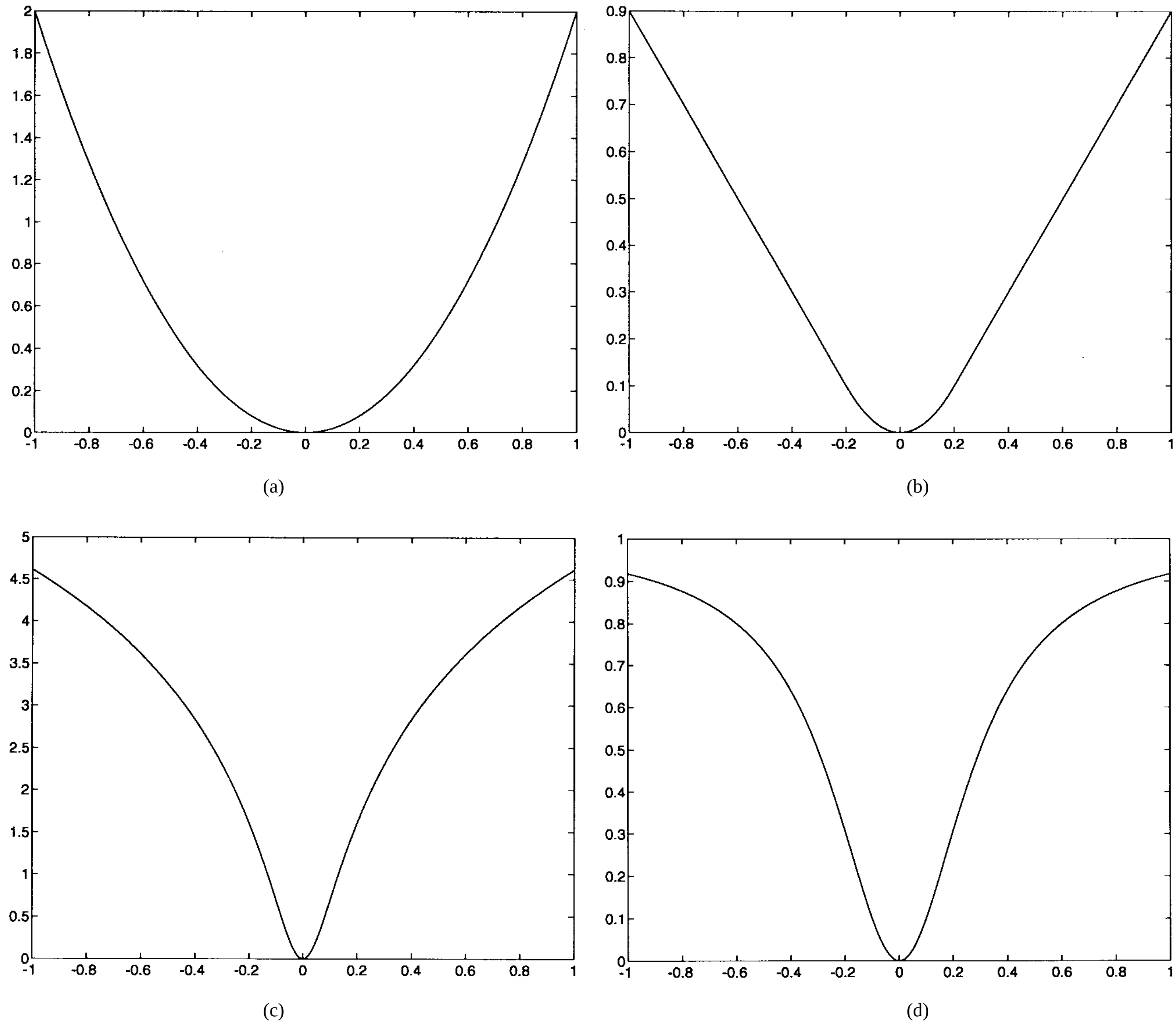


Fig. 2. Plot of the potential functions $V(x_i, x_j)$ versus $(x_i - x_j)$ for the four Gibbs priors in (8): (a) quadratic, (b) Huber, (c) log-quadratic, and (d) saturated quadratic.

The defining feature of a MRF is $P(x_i|x_j, \; \forall j \neq i) = P(x_i|x_j, j \in N_i)$. For the pairwise interaction model above, the conditional density has the special form [24]

$$P(x_i|x_j, j \in N_i) = \frac{\exp\left\{-\beta \sum_{j \in N_i} \kappa_{ij} V(x_i, x_j)\right\}}{\int_{x_i} \exp\left\{-\beta \sum_{j \in N_i} \kappa_{ij} V(x_i, x_j)\right\} dx_i}. \tag{7}$$

The specific potential functions, $V(\cdot, \cdot)$, used in this work are as follows.

Quadratic function:

$$V_1(x_i, x_j) = (x_i - x_j)^2.$$

Huber function:

$$V_2(x_i, x_j) = \begin{cases} \frac{1}{2\delta}(x_i - x_j)^2, & \text{if } |x_i - x_j| \le \delta \\ |x_i - x_j| - \frac{\delta}{2}, & \text{otherwise.} \end{cases}$$

Log-quadratic:

$$V_3(x_i, x_j) = \ln\left[1 + \frac{(x_i - x_j)^2}{\delta^2}\right].$$

Saturated-quadratic:

$$V_4(x_i, x_j) = \frac{(x_i - x_j)^2}{\delta^2 + (x_i - x_j)^2}. \tag{8}$$

These four functions are representatives of three major categories for image priors: strictly convex (V_1), semi-convex (V_2), and nonconvex (V_3, V_4) potential functions. They are illustrated in Fig. 2. All four can be used to model locally smooth images. However, the quadratic function $V_1(\cdot)$ penalizes the differences of neighboring pixels at an increasing rate, which tends to force the image to be smooth everywhere. The Huber function $V_2(\cdot)$ behaves as a quadratic function when the difference of the neighboring pairs are small, but applies a linear penalty when the differences are large; i.e., the rate of the penalty applied to intensity differences does not change beyond the threshold δ. Therefore, this prior does not differentiate substantially between slow monotonic changes and abrupt changes and consequently does not penalize the presence of edges or boundaries in the image. The function $V_3(\cdot)$ was introduced in [17] and $V_4(\cdot)$ in [11]. Both have saturating properties that actually decrease the rate of penalty applied to intensity differences beyond a threshold determined by δ. Consequently, they positively favor the presence of edges in the image. However, $V_3(\cdot)$ and $V_4(\cdot)$ are nonconvex, which presents difficulties in computing global MAP estimates.

B. Maximum Likelihood Hyperparameter Estimation from Incomplete Data

Given the observed incomplete data, $\boldsymbol{y}$, an ML estimate of β can be found from the maximizer of the marginalized likelihood function [8]

$$\begin{aligned} P(\boldsymbol{y}|\beta) &= \int_{\mathcal{X}} P(\boldsymbol{y}, \boldsymbol{x}|\beta)\, d\boldsymbol{x} \\ &= \int_{\mathcal{X}} P(\boldsymbol{y}|\boldsymbol{x}) P(\boldsymbol{x}|\beta)\, d\boldsymbol{x} \\ &= \frac{\int_{\mathcal{X}} \exp\{\ln P(\boldsymbol{y}|\boldsymbol{x}) - \beta U(\boldsymbol{x})\}\, d\boldsymbol{x}}{\int_{\mathcal{X}} \exp\{-\beta U(\boldsymbol{x})\}\, d\boldsymbol{x}} \\ &= \frac{Z(\boldsymbol{y}, \beta)}{Z(\beta)} \end{aligned} \tag{9}$$

where

$$Z(\boldsymbol{y}, \beta) = \int_{\mathcal{X}} \exp\{\ln P(\boldsymbol{y}|\boldsymbol{x}) - \beta U(\boldsymbol{x})\}\, d\boldsymbol{x} \tag{10}$$

is the partition function of the posterior density, $P(\boldsymbol{x}|\boldsymbol{y}, \beta)$. Therefore

$$\ln P(\boldsymbol{y}|\beta) = \ln Z(\boldsymbol{y}, \beta) - \ln Z(\beta) \tag{11}$$

and the ML estimator of the hyperparameter is a root of the equation

$$0 = \frac{\partial \ln P(\boldsymbol{y}|\beta)}{\partial \beta} \leftrightarrow \frac{\partial \ln Z(\boldsymbol{y}, \beta)}{\partial \beta} = \frac{\partial \ln Z(\beta)}{\partial \beta}. \tag{12}$$

It is straightforward to verify that

$$\begin{aligned} \frac{\partial \ln Z(\beta)}{\partial \beta} &= -\frac{\int_{\mathcal{X}} U(\boldsymbol{x}) \exp\{-\beta U(\boldsymbol{x})\}\, d\boldsymbol{x}}{\int_{\mathcal{X}} \exp\{-\beta U(\boldsymbol{x})\}\, d\boldsymbol{x}} \\ &= -\mathcal{E}[U(\boldsymbol{x})|\beta] \end{aligned} \tag{13}$$

$$\begin{aligned} \frac{\partial \ln Z(\boldsymbol{y}, \beta)}{\partial \beta} &= -\frac{\int_{\mathcal{X}} U(\boldsymbol{x}) \exp\{\ln P(\boldsymbol{y}|\boldsymbol{x}) - \beta U(\boldsymbol{x})\}\, d\boldsymbol{x}}{\int_{\mathcal{X}} \exp\{\ln P(\boldsymbol{y}|\boldsymbol{x}) - \beta U(\boldsymbol{x})\}\, d\boldsymbol{x}} \\ &= -\mathcal{E}[U(\boldsymbol{x})|\boldsymbol{y}, \beta] \end{aligned} \tag{14}$$

where $\mathcal{E}[\cdot|\beta]$ and $\mathcal{E}[\cdot|\boldsymbol{y}, \beta]$denote expectation with respect to the prior and posterior densities, respectively. It follows from (12)–(14) that the ML estimate of β from $\boldsymbol{y}$ is a root of the likelihood equation

$$\mathcal{E}[U(\boldsymbol{x})|\boldsymbol{y}, \beta] = \mathcal{E}[U(\boldsymbol{x})|\beta]. \tag{15}$$

This equation can in principle be solved using an EM algorithm [8], [11] as follows.

E-Step: Estimate the complete-data sufficient statistic $U[\boldsymbol{y}, \beta^{(k)}]$ by finding $U^{(k)}[\boldsymbol{y}, \beta^{(k)}] = \mathcal{E}[U(\boldsymbol{x})|\boldsymbol{y}, \beta^{(k)}]$.

M-Step: Determine $\beta^{(k+1)}$ as the solution of the equation $\mathcal{E}[U(\boldsymbol{x})|\beta] = U^{(k)}[\boldsymbol{y}, \beta^{(k)}]$.

Exact solution of this EM problem is impractical. Geman and McClure [11] note that a solution can be found using stochastic sampling from the posterior and prior densities to approximate the expectations. Due to the complexity of sampling from the posterior, the computation cost remains unacceptable. Therefore, in [11], a second estimation method is also described. This method of moments (MOM) simply requires the computation of a statistic $M(\boldsymbol{y})$ of the data. The parameter β is then chosen as a root of the equation

$$M(\boldsymbol{y}) = \mathcal{E}[M(\boldsymbol{y})|\beta]. \tag{16}$$

where the moment curve $\mathcal{E}[M(\boldsymbol{y})|\beta]$ is precomputed for a large range of β and should be monotonic with respect to β to ensure identifiability of the hyperparameter from the moment curve. Unfortunately, this method tends to perform poorly, at least for statistics that we have considered, due to small gradients in the moment curve, which result in high variance estimates of β.

In [3], [19], and [26], an alternative simplified approach is taken whereby, instead of maximizing with respect to β over the marginalized density (9), β is computed with $\boldsymbol{x}$ as the pair $\{\hat{\beta}, \hat{\boldsymbol{x}}\}$ that jointly maximize $P(\boldsymbol{y}, \boldsymbol{x}|\beta) = P(\boldsymbol{y}|\boldsymbol{x}) P(\boldsymbol{x}|\beta)$, as follows:

$$\begin{aligned} \{\hat{\boldsymbol{x}}, \hat{\beta}\} &= \arg\max_{\boldsymbol{x}, \beta} \ln P(\boldsymbol{y}, \boldsymbol{x}|\beta) \\ &= \arg\max_{\boldsymbol{x}} \{\ln P(\boldsymbol{y}|\boldsymbol{x}) + \max_{\beta} \ln P(\boldsymbol{x}|\beta)\}. \end{aligned} \tag{17}$$

Some authors term this the generalized ML (GML) method [26]. The optimization can be performed in a two-step algorithm

$$\hat{\boldsymbol{x}}^{(k)} = \arg\max_{\boldsymbol{x}} P[\boldsymbol{y}, \boldsymbol{x}|\hat{\beta}^{(k)}] \tag{18}$$

$$\hat{\beta}^{(k+1)} = \arg\max_{\beta} P[\hat{\boldsymbol{x}}^{(k)}|\beta]. \tag{19}$$

Note that the first step is actually the MAP estimate of $\boldsymbol{x}$ given the current choice of β, and the second step is the maximum likelihood estimate of β using the current MAP estimate of $\boldsymbol{x}$ as a direct (complete data) observation of $\boldsymbol{x}$. It is straightforward to show that the second step is equivalent to solving the equation $U[\hat{\boldsymbol{x}}^{(k)}] = \mathcal{E}[U(\boldsymbol{x})|\beta]$. From the viewpoint of statistical mechanics, GML gives an approximate solution to the likelihood in (15), which neglects *all* statistical fluctuations in the field $\boldsymbol{x}$ and considers only the contribution of the maximum term to integrals with respect to a Gibbs distribution [12]. As we shall see below, the mean field approach is far less restrictive than the GML approximation, which translates into significantly improved estimates of β when compared to GML.

In practice, direct computation of the GML estimate is still difficult as the second step requires evaluation of the partition function of the prior. This step is usually approximated using maximum pseudolikelihood (MPL) [3], [19], i.e., we replace the second step with

$$\hat{\beta}^{(k+1)} = \arg\max_{\beta} \prod_i P(\hat{x}_i^k | \hat{x}_j^k, j \in N_i, \beta). \tag{20}$$

We refer to this as the generalized maximum pseudolikelihood (GMPL) method in the following.

III. Maximum Likelihood Hyperparameter Estimation Using Mean and Mode Field Approximation

True ML estimation of β is difficult because of the complexity and dimensionality of the joint density $P(\boldsymbol{y}, \boldsymbol{x}|\beta)$. It is essentially impossible to compute the marginal density (partition functions) in (9) or expectations in (15) for each new data set $\boldsymbol{y}$. One approach to simplifying this problem is to approximate the multidimensional densities with separable joint densities equal to a product of one-dimensional (1-D) probability densities. The multidimensional integrals involved in computing marginal densities, partition functions, or moments, can then be approximated with a product or sum of 1-D integrals with respect to these 1-D pixel-wise densities.

Approximating Gibbs distributions using separable joint density functions is the basis for the mean field theory in classical statistical mechanics [7]. The mean field theory was originally developed as a statistical mechanics tool for the analysis of many body systems through approximation as a set of single body systems. The basic idea is to focus on one particular particle (in our case a pixel site) in the system and assume that the role of the neighboring particles (pixels) can be approximated by an average field that acts on the tagged particle. This approach, therefore, neglects the effects of statistical fluctuations in all pixels other than the current tagged one. The corresponding joint description is simply the product of that for each individual particle or pixel. Mean field approximation has previously been applied in the image processing field to surface reconstruction [12], image segmentation [36], image restoration [38], and motion estimation [37]. However, we believe this is the first application of this approach to parameter estimation in image processing.

In this section, we focus first on a restricted class of Gibbs distribution for which we develop an optimal mean field approximation. We use the GBF bound to select the mean field approximation, which leads to an optimal approximation of the partition function. Using this result we describe an "optimal" ML hyperparameter estimator focusing on the problem of image restoration from Gaussian data with a quadratic Gibbs prior. Unfortunately, this optimized approximation is not applicable to the general problem. For the general case, we provide a heuristic development of an alternative approximation that can be applied to problems with Gibbs priors for any of the four potential functions in (8) with either the Gaussian or Poisson likelihood functions.

A. Optimal Approximation of the Partition Function

We can see from (11) that the true ML estimate of β is completely determined by the prior and posterior partition functions. Therefore, for the purposes of computing an accurate ML estimate of β, the mean field approximations of the prior and posterior Gibbs distributions should be chosen to give the best approximations of their respective partition functions. We begin by describing our partition function optimization procedure for a restricted class of Gibbs distributions. We then apply this to approximation of the prior and posterior distributions to develop the mean field approximated ML estimator of β. The development below is based on that in [7] in several places. We emphasize that it is the application of this approximation to parameter estimation, rather than the approximation itself, that is novel.

The approximation involves replacing the true Gibbs distribution, $P(\boldsymbol{x})$, with a *mean field reference* distribution, $P_{\mathrm{MF}}(\boldsymbol{x})$, which is a separable function in $\boldsymbol{x}$

$$P(\boldsymbol{x}) \approx P_{\mathrm{MF}}(\boldsymbol{x}) = \prod_i P_i^{\mathrm{mf}}(x_i). \tag{21}$$

i.e., the pixels are modeled as independent random variables. The choice of the mean field reference distribution is based on the following result.

Theorem 1—Gibbs–Bogoliubov–Feynman Bound [7]: For a Gibbs distribution with partition function Z and Gibbs energy E, and any other Gibbs distribution with partition function Z_{MF} and Gibbs energy E_{MF}, we have the following inequality:

$$Z \geq Z_{\mathrm{MF}} \exp\{-\langle E - E_{\mathrm{MF}}\rangle_{\mathrm{MF}}\} \tag{22}$$

where

$$\langle \cdots \rangle_{\mathrm{MF}} \stackrel{\mathrm{def}}{=} Z_{\mathrm{MF}}^{-1} \int_{\mathcal{X}} [\cdots] \exp(-E_{\mathrm{MF}})\, d\mathbf{x}. \tag{23}$$

Theorem 1 states that if we use *any* Gibbs distribution to approximate the original Gibbs distribution, the quantity $Z_{\mathrm{MF}} \exp\{-\langle E - E_{\mathrm{MF}}\rangle_{\mathrm{MF}}\}$ will never exceed the original Z. Consequently, the mean field reference distribution that leads to the best approximation of the original partition function, can be found by maximizing the quantity on the right side of the GBF bound.

Proposition 1: The partition function Z can be best approximated through a mean field reference distribution with partition function Z_{MF} and Gibbs energy E_{MF} as

$$Z \approx Z_{\mathrm{MF}} \exp\{-\langle E - E_{\mathrm{MF}}\rangle_{\mathrm{MF}}\} \tag{24}$$

where E_{MF} maximizes $Z_{\mathrm{MF}} \exp\{-\langle E - E_{\mathrm{MF}}\rangle_{\mathrm{MF}}\}$.

Unfortunately, a closed-form solution to this optimization problem exists only for a restricted class of Gibbs distributions. This includes the class of continuous state *auto-models* [2], to which we now apply Proposition 1. The auto-models have the form $P(\boldsymbol{x}) = Z^{-1} \exp\{-E(\boldsymbol{x})\}$ where

$$E(\boldsymbol{x}) = \sum_i \left[x_i G_i(x_i) + \tfrac{1}{2} \sum_{j \in N_i} b_{ij} x_i x_j \right] \tag{25}$$

with $b_{ij} = b_{ji}$ and the single pixels sample space $x_i \in [0, x_{\max}]$. The mean field reference distribution $P_{\mathrm{MF}}(\boldsymbol{x})$ is chosen in this case as a separable Gibbs distribution with mean field energy $E_{\mathrm{MF}}(\boldsymbol{x})$ of the form

$$E_{\mathrm{MF}}(\boldsymbol{x}) = \sum_i x_i G_i(x_i) + \sum_i \Delta H_i x_i. \tag{26}$$

This reference distribution approximates the influence of neighboring pixels $\{x_j, j \in N_i\}$ by a constant ΔH_i. We now develop an optimal reference in the sense of choosing ΔH_i to maximize the right side of the GBF bound.

Since the reference field is separable, i.e., $P_{\mathrm{MF}}(\boldsymbol{x}) = \prod_i P_i^{\mathrm{mf}}(x_i)$, we consider first the local mean field reference density

$$P_i^{\mathrm{mf}}(x_i) = \frac{1}{Z_i^{\mathrm{mf}}} \exp\{-[x_i G_i(x_i) + \Delta H_i x_i]\} \tag{27}$$

with

$$Z_i^{\mathrm{mf}} = \int_{x_i} \exp\{-[x_i G_i(x_i) + \Delta H_i x_i]\}\, dx_i \tag{28}$$

the corresponding local mean field partition function. As a direct result of (27) and (28), the mean of the reference field, i.e., the mean field value, is

$$\begin{aligned}\langle x_i\rangle^{\mathrm{mf}} &= \frac{1}{Z_i^{\mathrm{mf}}} \int_{x_i} x_i \exp\{-[x_i G_i(x_i) + \Delta H_i x_i]\}\, dx_i \\ &= -\frac{\partial \ln Z_{\mathrm{MF}}}{\partial \Delta H_i}, \qquad \forall i.\end{aligned} \tag{29}$$

The value of ΔH_i that maximizes the right side of the GBF bound must satisfy

$$\begin{aligned}0 &= \frac{\partial}{\partial \Delta H_i} \ln (Z_{\mathrm{MF}} \exp\{-\langle E - E_{\mathrm{MF}}\rangle_{\mathrm{MF}}\}) \\ &= \frac{\partial \ln Z_{\mathrm{MF}}}{\partial \Delta H_i} - \frac{\partial}{\partial \Delta H_i}\{\langle E - E_{\mathrm{MF}}\rangle_{\mathrm{MF}}\}, \quad \forall i.\end{aligned} \tag{30}$$

We proceed with

$$\begin{aligned}\langle E - E_{\mathrm{MF}}\rangle_{\mathrm{MF}} = \tfrac{1}{2} \sum_i \sum_{j \in N_i} b_{ij} \langle x_i\rangle^{\mathrm{mf}} \langle x_j\rangle^{\mathrm{mf}} \\ - \sum_i \Delta H_i \langle x_i\rangle^{\mathrm{mf}}\end{aligned} \tag{31}$$

where we use the independence of pixels in the reference field to simplify $\langle x_i x_j\rangle^{\mathrm{mf}} = \langle x_i\rangle^{\mathrm{mf}} \langle x_j\rangle^{\mathrm{mf}}$ for $i \neq j$. By combining (29) and (31) in (30), noting that each pair $\langle x_i\rangle^{\mathrm{mf}} \langle x_j\rangle^{\mathrm{mf}}$ appears twice in the summation, and that $b_{ij} = b_{ji}$, we get

$$\begin{aligned}0 = -\langle x_i\rangle^{\mathrm{mf}} - \frac{\partial \langle x_i\rangle^{\mathrm{mf}}}{\partial \Delta H_i} \sum_{j \in N_i} b_{ij} \langle x_j\rangle^{\mathrm{mf}} \\ + \langle x_i\rangle^{\mathrm{mf}} + \Delta H_i \frac{\partial \langle x_i\rangle^{\mathrm{mf}}}{\partial \Delta H_i}.\end{aligned} \tag{32}$$

Solving this gives

$$\Delta H_i = \sum_{j \in N_i} b_{ij} \langle x_j\rangle^{\mathrm{mf}}. \tag{33}$$

This is the value of the constant ΔH_i, which maximizes the right side of the GBF bound over the set of separable Gibbs distribution with energies of the form of (26).

Substituting (33) into (27), we obtain the optimal local mean field reference distribution for the auto-models

$$\begin{aligned}&P_i^{\mathrm{mf}}(x_i) = \\ &\frac{1}{Z_i^{\mathrm{mf}}} \exp\left\{-\left[x_i G_i(x_i) + \sum_i \sum_{j \in N_i} b_{ij} x_i \langle x_j\rangle^{\mathrm{mf}}\right]\right\}.\end{aligned} \tag{34}$$

We note the optimal mean field local density is equivalent to fixing the values of the neighboring sites of x_i in the Markov local conditional density at their mean field values, i.e., $P(x_i | \langle x_j\rangle^{\mathrm{mf}}, j \in N_i)$.

Substituting (33) into (31), we obtain

$$\begin{aligned}\langle E - E_{\mathrm{MF}}\rangle_{\mathrm{MF}} &= \tfrac{1}{2} \sum_i \sum_{j \in N_i} b_{ij} \langle x_i\rangle^{\mathrm{mf}} \langle x_j\rangle^{\mathrm{mf}} \\ &\quad - \sum_i \sum_{j \in N_i} b_{ij} \langle x_j\rangle^{\mathrm{mf}} \langle x_i\rangle^{\mathrm{mf}} \\ &= -\tfrac{1}{2} \sum_i \sum_{j \in N_i} b_{ij} \langle x_i\rangle^{\mathrm{mf}} \langle x_j\rangle^{\mathrm{mf}}.\end{aligned} \tag{35}$$

The optimal approximation of the partition function Z is then given by (36), shown at the bottom of the page.

B. Hyperparameter Estimation Using an Optimal Approximation

The optimal mean field approximation mechanism developed in the previous subsection can be directly applied to ML hyperparameter estimation in image restoration and reconstruction problems with the Gaussian likelihood function (1) and the quadratic Gibbs prior, $V_1(\cdot, \cdot)$ in (8). We can write the Gibbs energies of the prior and posterior densities for this specific example as, respectively

$$\begin{aligned}E^{(PR)}(\boldsymbol{x}) &= \beta U(\boldsymbol{x}) \\ &= \beta \sum_i \sum_{j \in N_i^{PR},\, j>i} \kappa_{ij} (x_i - x_j)^2 \\ &= \beta \left(\sum_i \sum_{j \in N_i^{PR}} \kappa_{ij} x_i^2 - \sum_i \sum_{j \in N_i^{PR}} \kappa_{ij} x_i x_j \right)\end{aligned} \tag{37}$$

and

$$\begin{aligned}Z &\approx Z_{\mathrm{MF}} \exp\{-\langle E - E_{\mathrm{MF}}\rangle_{\mathrm{MF}}\} \\ &= \left(\prod_i^N \int_{x_i} \exp\left\{-\left[x_i G_i(x_i) + x_i \sum_{j \in N_i} b_{ij} \langle x_j\rangle^{\mathrm{mf}}\right]\right\} dx_i \right) \left(\exp\left[\tfrac{1}{2} \sum_i \sum_{j \in N_i} b_{ij} \langle x_i\rangle^{\mathrm{mf}} \langle x_j\rangle^{\mathrm{mf}}\right] \right)\end{aligned} \tag{36}$$

$$
\begin{aligned}
E^{(PO)}(\boldsymbol{x}) &= -\ln P(\boldsymbol{y}|x) + \beta U(\boldsymbol{x}) \\
&= \tfrac{1}{2}\,(\boldsymbol{y}-\boldsymbol{Ax})^T \boldsymbol{C}^{-1}(\boldsymbol{y}-\boldsymbol{Ax}) \\
&\quad + \beta \sum_i \sum_{j\in N_i^{PR},\, j>i} \kappa_{ij}(x_i - x_j)^2 + K_1 \\
&= \sum_i F_i(x_i) + \sum_i \sum_{j\in N_i^{PO}} \gamma_{ij} x_i x_j \\
&\quad + \beta\left(\sum_i \sum_{j\in N_i^{PR}} \kappa_{ij} x_i^2 - \sum_i \sum_{j\in N_i^{PR}} \kappa_{ij} x_i x_j\right) \\
&\quad + K_2
\end{aligned} \tag{38}
$$

where

$$\gamma_{ij} = \tfrac{1}{2}\,[\boldsymbol{A}^T \boldsymbol{C}^{-1} \boldsymbol{A}]_{ij} \tag{39}$$

and

$$F_i(x_i) = -[\boldsymbol{A}^T \boldsymbol{C}^{-1} \boldsymbol{y}]_i x_i + \tfrac{1}{2}\,[\boldsymbol{A}^T \boldsymbol{C}^{-1}\boldsymbol{A}]_{ii} x_i^2. \tag{40}$$

The superscripts PR and PO denote prior and posterior, respectively. The constant terms K_1 and K_2 are independent of $\boldsymbol{x}$ and β and do not affect the choice of ΔH_i or estimation of β. N_i^{PR} and N_i^{PO} denote the prior and posterior neighborhoods of pixel i.

Clearly, both prior and posterior distributions belong to the class of auto-models discussed in the previous subsection. Therefore, we use the optimal choice of ΔH_i from (33) in (37) and (38), which gives the following mean field energy functions:

$$E_{\rm MF}^{(PO)} = \sum_i \{L_i^{PO}(x_i) + \beta U_i^{PO}(x_i)\} \tag{41}$$

$$E_{\rm MF}^{(PR)} = \beta \sum_i U_i^{PR}(x_i) \tag{42}$$

where

$$L_i^{PO}(x_i) = F_i(x_i) + 2\sum_{j\in N_i^{PO}} \gamma_{ij} x_i \langle x_j\rangle_{\rm mf}^{PO} \tag{43}$$

$$U_i^{PO}(x_i) = \sum_{j\in N_i^{PR}} \kappa_{ij} x_i^2 - 2\sum_{j\in N_i^{PR}} \kappa_{ij} x_i \langle x_j\rangle_{\rm mf}^{PO} \tag{44}$$

$$U_i^{PR}(x_i) = \sum_{j\in N_i^{PR}} \kappa_{ij} x_i^2 - 2\sum_{j\in N_i^{PR}} \kappa_{ij} x_i \langle x_j\rangle_{\rm mf}^{PR}. \tag{45}$$

Here, $\langle x_j\rangle_{\rm mf}^{PO}$ and $\langle x_j\rangle_{\rm mf}^{PR}$ denote $\langle x_j\rangle^{\rm mf}$ with respect to the posterior and prior densities, respectively.

Having developed the optimal mean field reference densities, there are two alternative approaches to computing the approximate ML hyperparameter estimate. One is to compute the partition function approximation as

$$Z(\boldsymbol{y}, \beta) \approx Z_{\rm MF}^{(PO)} \exp\{-\langle E^{(PO)} - E_{\rm MF}^{(PO)}\rangle_{\rm MF}^{PO}\} \tag{46}$$

$$Z(\beta) \approx Z_{\rm MF}^{(PR)} \exp\{-\langle E^{(PR)} - E_{\rm MF}^{(PR)}\rangle_{\rm MF}^{PR}\} \tag{47}$$

and then to compute the mean field approximated ML estimate of β by finding the maximum of $\log Z(\boldsymbol{y}, \beta) - \log Z(\beta)$ based on the likelihood in (11).

The other approach is to compute the approximate prior and posterior expectations of the Gibbs prior energy, $U(x)$, as follows:

$$\mathcal{E}[U(\boldsymbol{x})|\boldsymbol{y}, \beta] \approx \langle U(\boldsymbol{x})\rangle_{\rm MF}^{PO} \tag{48}$$

$$\mathcal{E}[U(\boldsymbol{x})|\beta] \approx \langle U(\boldsymbol{x})\rangle_{\rm MF}^{PR} \tag{49}$$

and substitute these into the likelihood equation (15)

$$\langle U(\boldsymbol{x})\rangle_{\rm MF}^{PO} = \langle U(\boldsymbol{x})\rangle_{\rm MF}^{PR} \tag{50}$$

or equivalently

$$
\begin{aligned}
&\langle U_{\rm MF}(\boldsymbol{x})\rangle_{\rm MF}^{PO} - \langle U(\boldsymbol{x}) - U_{\rm MF}(\boldsymbol{x})\rangle_{\rm MF}^{PO} \\
&\quad = \langle U_{\rm MF}(\boldsymbol{x})\rangle_{\rm MF}^{PR} - \langle U(\boldsymbol{x}) - U_{\rm MF}(\boldsymbol{x})\rangle_{\rm MF}^{PR}.
\end{aligned} \tag{51}
$$

This can be rewritten, using (35), (41), and (42), as

$$
\begin{aligned}
&\sum_i \frac{\int_{x_i} U_i^{PO}(x_i)\exp\{-L_i^{PO}(x_i) - \beta U_i^{PO}(x_i)\}\,dx_i}{\int_{x_i} \exp\{-L_i^{PO}(x_i) - \beta U_i^{PO}(x_i)\}\,dx_i} \\
&\quad - \sum_i \sum_{j\in N_i^{PR}} \kappa_{ij}\langle x_i\rangle_{\rm mf}^{PO}\langle x_j\rangle_{\rm mf}^{PO} \\
&= \sum_i \frac{\int_{x_i} U_i^{PR}(x_i)\exp\{-\beta U_i^{PR}(x_i)\}\,dx_i}{\int_{x_i} \exp\{-\beta U_i^{PR}(x_i)\}\,dx_i} \\
&\quad - \sum_i \sum_{j\in N_i^{PR}} \kappa_{ij}\langle x_i\rangle_{\rm mf}^{PR}\langle x_j\rangle_{\rm mf}^{PR}
\end{aligned} \tag{52}
$$

or equivalently

$$
\begin{aligned}
&\sum_i \mathcal{E}[U_i^{PO}(x_i)|\langle x_j\rangle_{\rm mf}^{PO},\, j\in N_i^{PO}, \boldsymbol{y}, \beta] \\
&\quad - \sum_i \sum_{j\in N_i^{PR}} \kappa_{ij}\langle x_i\rangle_{\rm mf}^{PO}\langle x_j\rangle_{\rm mf}^{PO} \\
&= \sum_i \mathcal{E}[U_i^{PR}(x_i)|\langle x_j\rangle_{\rm mf}^{PR},\, j\in N_i^{PR}, \beta] \\
&\quad - \sum_i \sum_{j\in N_i^{PR}} \kappa_{ij}\langle x_i\rangle_{\rm mf}^{PR}\langle x_j\rangle_{\rm mf}^{PR}.
\end{aligned} \tag{53}
$$

In this paper, we adopt the latter approach, i.e., to compute the estimate of β by finding a root of (53). For a given mean field $\langle x_i\rangle^{\rm mf}$, β can be computed by finding a root of this equation. Since the mean field $\langle x_i\rangle^{\rm mf}$ is itself dependent on the value β, a recursive procedure that alternates between computation of $\langle x_i\rangle^{\rm mf}$ using the current value of β and vice versa, is required. We return to the problem of computing the solution in Section III-D.

C. Hyperparameter Estimation Using a Generalized Approximation

The preceding development works only for the restricted class of auto-Gibbs distributions of the form (25). We now consider the more general case, and present a heuristic development of a mean field reference that can be applied to both

Poisson and Gaussian likelihoods with any of the four potential functions in (8). Consider the general Gibbs distribution, which is to be approximated

$$P(\boldsymbol{x}) = \frac{1}{Z} \exp\{-E(\boldsymbol{x})\} \tag{54}$$

with conditional density

$$P(x_i | x_j, j \in N_i) = \frac{1}{Z_i} \exp\{-E_i(x_i; x_j, j \in N_i)\} \tag{55}$$

where $E_i(x_i; x_j, j \in N_i)$ is the sum over all potential functions in $E(\boldsymbol{x})$ that include site i and N_i denotes the set of neighboring pixel sites of i. We again use a separable mean field approximation

$$\begin{aligned} P_{\mathrm{MF}}(\boldsymbol{x}) &= \frac{1}{Z_{\mathrm{MF}}} \exp\{-E_{\mathrm{MF}}(\boldsymbol{x})\} \\ &\equiv \prod_i P_i^{\mathrm{mf}}(x_i) \end{aligned} \tag{56}$$

where we define the local mean field densities $P_i^{\mathrm{mf}}(x_i)$ to be equal to the conditional density for each site given the mean field of their neighbors, i.e.,

$$\begin{aligned} P_i^{\mathrm{mf}}(x_i) &= \frac{P(x_i, \langle x_{S\setminus i}\rangle^{\mathrm{mf}})}{\int_{x_i} P(x_i, \langle x_{S\setminus i}\rangle^{\mathrm{mf}})\, dx_i} \\ &= P(x_i | x_j = \langle x_j\rangle^{\mathrm{mf}}, j \in N_i) && (57) \\ &= \frac{1}{Z_i^{\mathrm{mf}}} \exp\{-E_i^{\mathrm{mf}}(x_i; \langle x_j\rangle^{\mathrm{mf}}, j \in N_i)\} && (58) \end{aligned}$$

where $S\setminus i$ denotes all sites except i. The corresponding partition function Z_i^{mf} is then given by

$$Z_i^{\mathrm{mf}} = \int_{x_i} \exp\{-E_i^{\mathrm{mf}}(x_i; \langle x_j\rangle^{\mathrm{mf}}, j \in N_i)\}\, dx_i. \tag{59}$$

Combining the local energy and partition functions gives the overall *mean field energy function* $E_{\mathrm{MF}}(\boldsymbol{x})$, and *mean field partition function*, Z_{MF}

$$\begin{aligned} E_{\mathrm{MF}}(\boldsymbol{x}) &= \sum_i E_i^{\mathrm{mf}}(x_i; \langle x_j\rangle^{\mathrm{mf}}, j \in N_i) \\ Z_{\mathrm{MF}} &= \prod_i Z_i^{\mathrm{mf}}. \end{aligned} \tag{60}$$

This mean field approximation can be applied to either the prior, $P(\boldsymbol{x}|\beta)$, or posterior, $P(\boldsymbol{x}|\boldsymbol{y}, \beta)$, densities in Bayesian inverse problems provided the densities are written in the form of a Gibbs distribution. Clearly, this generalized mean field reference system takes the same form as the optimal one for the auto-models [see (34) and the comments that follow].

The generalized mean field reference system for a prior with any of the potentials in (8) can then be written as

$$P_{\mathrm{MF}}^{PR}(\boldsymbol{x}|\beta) = \prod_i \frac{\exp\left\{-\beta \sum_{j\in N_i^{PR}} \kappa_{ij} V(x_i, \langle x_j\rangle_{\mathrm{mf}}^{PR})\right\}}{\int_{x_i} \exp\left\{-\beta \sum_{j\in N_i^{PR}} \kappa_{ij} V(x_i, \langle x_j\rangle_{\mathrm{mf}}^{PR})\right\} dx_i}. \tag{61}$$

For the posterior distribution with a Poisson or Gaussian likelihood and a prior with any of the potential functions in (8), the mean field reference can be written as (62), shown at the bottom of the page. $L_i^{PO}(x_i)$ is defined in (43) for the Gaussian likelihood. For the Poisson likelihood model (2), we use

$$L_i^{PO}(x_i) = \ln P(y | x_i; x_j = \langle x_j\rangle_{\mathrm{mf}}^{PO}, \forall j \neq i). \tag{63}$$

Having developed the generalized mean field reference system, we can compute the ML estimate of β by finding a root of $\langle U(\boldsymbol{x})\rangle_{\mathrm{MF}}^{PO} = \langle U(\boldsymbol{x})\rangle_{\mathrm{MF}}^{PR}$ as we did in (50)–(53) for the optimal case. It is easy to show that for any of the four potential functions in (8)

$$\langle U_{\mathrm{MF}}(\boldsymbol{x})\rangle_{\mathrm{MF}}^{PO} = \sum_i \mathcal{E}[U_i^{PO}(x_i) | \langle x_j\rangle_{\mathrm{mf}}^{PO}, j \in N_i^{PO}, \boldsymbol{y}, \beta] \tag{64}$$

$$\langle U_{\mathrm{MF}}(\boldsymbol{x})\rangle_{\mathrm{MF}}^{PR} = \sum_i \mathcal{E}[U_i^{PR}(x_i) | \langle x_j\rangle_{\mathrm{mf}}^{PR}, j \in N_i^{PR}, \beta]. \tag{65}$$

where

$$U_i^{PO}(x_i) = \sum_{j\in N_i^{PR}} \kappa_{ij} V(x_i, \langle x_j\rangle_{\mathrm{mf}}^{PR})$$

and

$$U_i^{PR}(x_i) = \sum_{j\in N_i^{PR}} \kappa_{ij} V(x_i, \langle x_j\rangle_{\mathrm{mf}}^{PO}).$$

The terms of $\langle U(\boldsymbol{x}) - U_{\mathrm{MF}}^{PO}(\boldsymbol{x})\rangle_{\mathrm{MF}}^{PO}$ and $\langle U(\boldsymbol{x}) - U_{\mathrm{MF}}^{PR}(\boldsymbol{x})\rangle_{\mathrm{MF}}^{PR}$ in (51) are difficult to evaluate except for the case of the quadratic potential function. However, we note that for the case where the prior is an auto-model, if we use the posterior mean field, $\langle x_j\rangle_{\mathrm{mf}}^{PO}$, in the place of $\langle x_j\rangle_{\mathrm{mf}}^{PR}$ in (51), then these two terms cancel. Applying this idea for the general case by dropping

$$P_{\mathrm{MF}}^{PO}(\boldsymbol{x}|\boldsymbol{y}, \beta) = \prod_i \frac{\exp\left\{-L_i^{PO}(x_i) - \beta \sum_{j\in N_i^{PR}} \kappa_{ij} V(x_i, \langle x_j\rangle_{\mathrm{mf}}^{PO})\right\}}{\int_{x_i} \exp\left\{-L_i^{PO}(x_i) - \beta \sum_{j\in N_i^{PR}} \kappa_{ij} V(x_i, \langle x_j\rangle_{\mathrm{mf}}^{PO})\right\} dx_i} \tag{62}$$

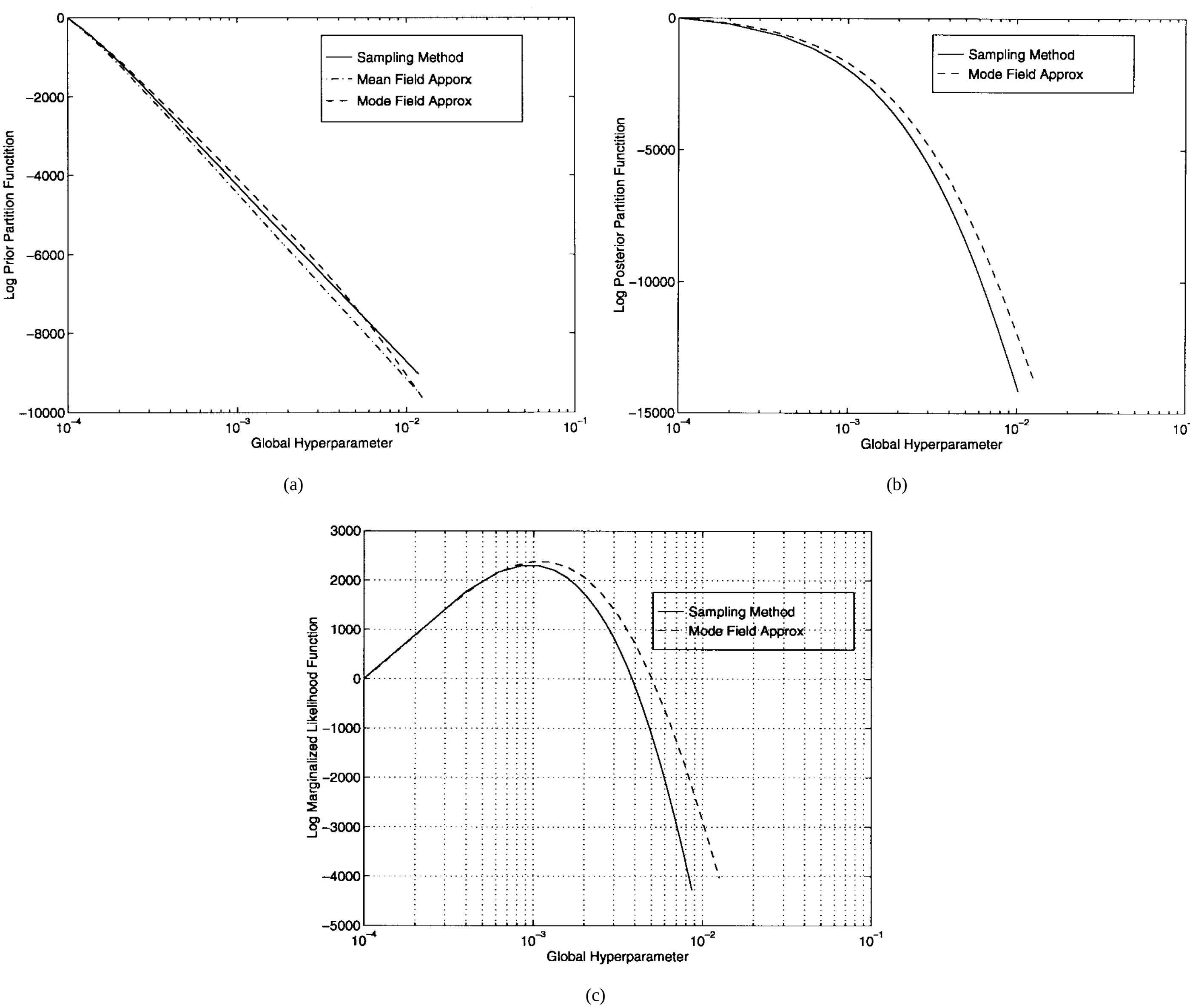


Fig. 3. Comparison of partition function approximation using MCMC sampling and mode field approximation. (a) Plot of $\log Z(\boldsymbol{y}, \beta)$, the prior partition function. (b) Plot of $\log Z(\boldsymbol{y}, \beta)$, the posterior partition function. (c) Plot of $\log p(\boldsymbol{y}|\beta) = \log Z(\boldsymbol{y}, \beta) - \log Z(\beta)$, whose maximum corresponds to ML estimate of β.

the second term on either side of (51), the equation reduces to

$$\sum_i \frac{\int_{x_i} U_i^{PO}(x_i) \exp\{-L_i^{PO}(x_i) - \beta U_i^{PO}(x_i)\}\, dx_i}{\int_{x_i} \exp\{-L_i^{PO}(x_i) - \beta U_i^{PO}(x_i)\}\, dx_i} = \sum_i \frac{\int_{x_i} U_i^{PO}(x_i) \exp\{-\beta U_i^{PO}(x_i)\}\, dx_i}{\int_{x_i} \exp\{-\beta U_i^{PO}(x_i)\}\, dx_i}. \tag{66}$$

We can rewrite (66) as

$$\sum_i \mathcal{E}[U_i^{PO}(x_i)|\langle x_j\rangle_{\mathrm{mf}}^{PO}, j \in N_i^{PO}, \boldsymbol{y}, \beta] = \sum_i \mathcal{E}[U_i^{PO}(x_i)|\langle x_j\rangle_{\mathrm{mf}}^{PO}, j \in N_i^{PR}, \beta] \tag{67}$$

which can be interpreted as a general mean field approximation of the likelihood in (15). Note that this version of the mean field approximated ML estimator is different from that derived using the GBF bound, i.e., (53), even for the auto-models. As we see below, methods that use the GBF bound outperform those based on (67). This is not surprising given the optimal nature of the first and heuristic nature of the second method. However, in cases where the optimal approximation cannot be found, the second method still performs exceptionally well in comparison to other well-known methods.

D. Mean and Mode Field Approximations

In many imaging applications, we are more interested in computing a MAP estimate of the image than a minimum mean squared error estimate. These correspond, respectively, to the mode and the mean of the posterior densities. Therefore, rather than also computing the mean field of the posterior reference

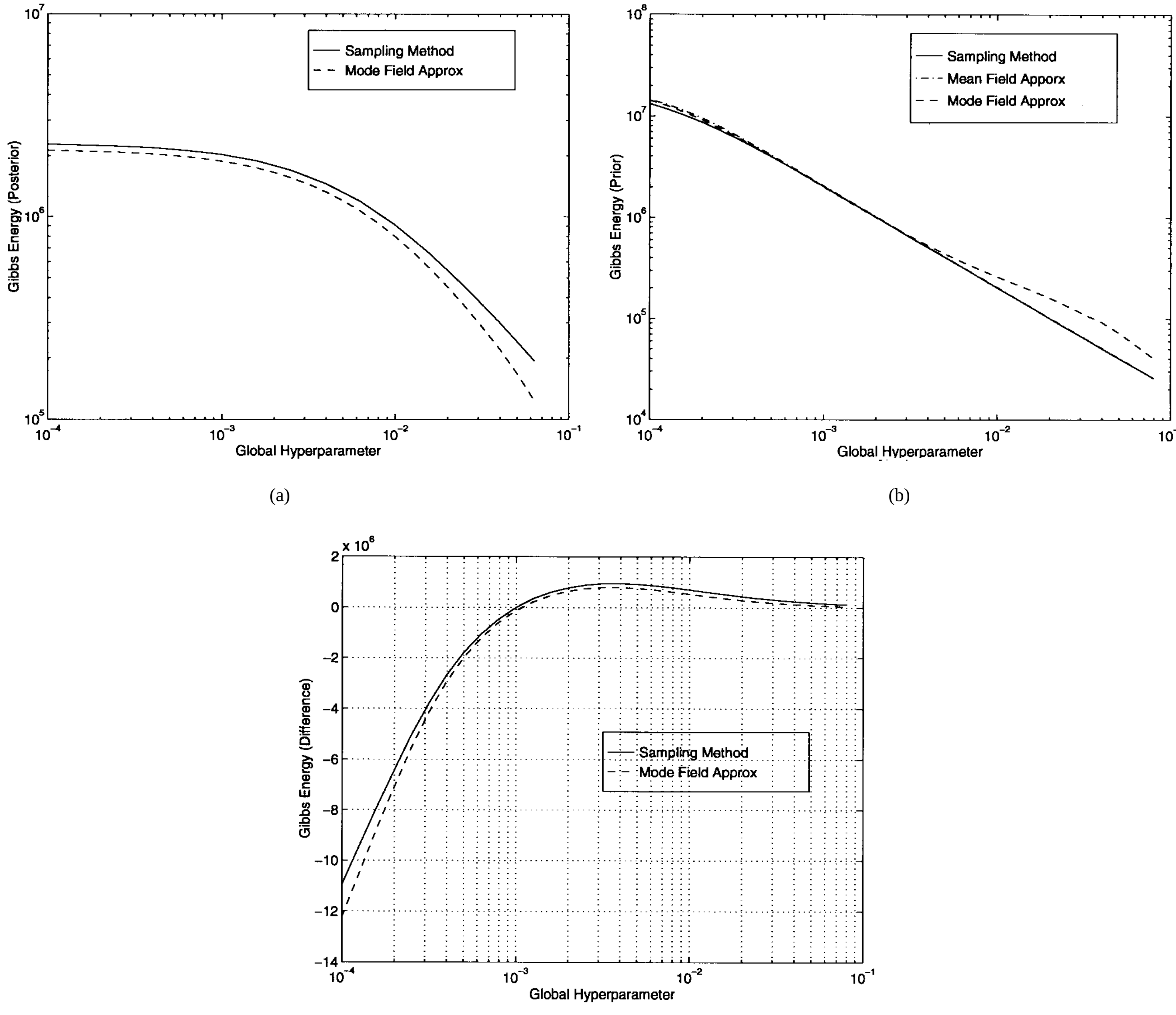


Fig. 4. Comparison of MCMC sampling and mode field approximation methods for solving the likelihood equation (15). (a) $\mathcal{E}[U(\boldsymbol{x})|\boldsymbol{y}, \beta]$. (b) $\mathcal{E}[U(\boldsymbol{x})|\beta]$. (c) $\mathcal{E}[U(\boldsymbol{x})|\boldsymbol{y}, \beta] - \mathcal{E}[U(\boldsymbol{x})|\beta]$. The function in (c) equals zero at the ML value of β.

field, we replace the mean field with a *mode-field*. This mode is computed using an iterative MAP estimation procedure. Note that using the separable approximations described above, the mode of the original and reference fields are identical. We note that this mode-field approximation is referred to as a saddle point approximation in [38]. In cases where the posterior density is unimodal and symmetric, the mean and mode field approximations are identical. This would be the case for Gaussian data with a Gaussian prior on the image. For the case where the single pixel sample space is not the entire real line, or when the MRF prior is nonquadratic, then the mode and mean field approximations will differ. This is also the case for Poisson data, since the Poisson likelihood is asymmetric. We refer to the parameter estimation methods described above as *mode field approximated maximum likelihood* (MFAML). To distinguish the two approximations in Sections III-B and III-C, we refer to them as MFAML-Opt and MFAML-Gen, respectively.

To summarize, we have developed an optimal mean field reference distribution for auto-models, of which the Gibbs quadratic prior and the Gibbs posterior formed by a Gaussian likelihood and a quadratic prior are special examples. Based on this mean field reference, we have provided a mechanism for the optimal approximation of partition functions and expectations. To facilitate the generalization of the methodology, we also propose a suboptimal approximation in which both the prior and posterior mean fields are replaced by the posterior mode field.

To examine these approximation strategies, we conducted an experiment using MCMC sampling [4]. A sample image was generated using the Metropolis algorithm for the quadratic prior with $\beta = 0.001$ on a lattice of size 64×64 with a single pixel sample space, [0, 100]. Data $\boldsymbol{y}$ were then generated according to the equation $\boldsymbol{y} = \mathbf{A}\boldsymbol{x}+\boldsymbol{n}$, where $\mathbf{A}$ is the blurring function Opt 1 (defined in Section V-A), $\boldsymbol{n}$ is Gaussian noise with zero mean and variance $\sigma^2 = 4$. Using this data, we then compared the functions computed using our mode field approximations with those obtained using MCMC methods. For each of the curves shown below, the MCMC sampling method used the Metropolis algorithm with a 500 cycle burn-in period. Averages were then computed from the next 1000 cycles. This procedure was repeated for each value of β. We used different initialization points to check for convergence after 1000 averages.

Shown in Fig. 3(a) and (b) are plots of the log-partition function for the prior and posterior densities. In Fig. 3(c), we plot $\log Z(\boldsymbol{y}|\beta) - \log Z(\beta)$ which is equal to the log likelihood $\log P(\boldsymbol{y}|\beta)$ [see (11)]. The value of β at which this function attains its maximum is therefore the ML estimate. We note that there is some displacement between the maxima of the functions using the mode-field approximation and MCMC sampling, but the difference is small. Since our estimation procedure solves an approximated version of the likelihood equation (15), rather than directly maximizing the log-likelihood, we also plot the expected values of the Gibbs energy with respect to the posterior and prior densities in Fig. 4(a) and (b), respectively. Their difference, $\mathcal{E}[U(\boldsymbol{x}|\boldsymbol{y}, \beta] - \mathcal{E}[U(\boldsymbol{x})|\beta]$ is shown in Fig. 4(c). Note from (15) that this function should equal zero at the ML solution. Again, while the mode-field approximation and MCMC sampling curves do not exactly coincide, the ML solutions obtained using both methods are very close. We note that we cannot draw strong conclusions regarding differences in bias and variance between the sampling method and mode-field approximation from these examples, since they are based on a single realization of $\boldsymbol{x}$ and $\boldsymbol{y}$. The results presented for our method in the following section involve averages over 50 realizations for many different cases. Due to the high computational cost involved in the use of the MCMC methods we could not include these methods in our comparisons.

IV. Numerical Methods

A. Combined MAP Image Estimation and ML Hyperparameter Estimation

Using the approximations described above, the MAP estimate of the image and the ML estimate of the hyperparameter can be jointly computed using a two-step iteration, as follows.

1) Initialize the image $\boldsymbol{x}^k = \boldsymbol{x}^0$ and hyperparameter $\beta^k = \beta^0$. Set $k = 0$.
2) Maximize $P(\boldsymbol{x}|\boldsymbol{y}; \beta^k)$ to find $\boldsymbol{x}^{k+1}$.
3) Compute a new hyperparameter value β^{k+1} by solving the approximated likelihood equation (53) or (67) using $\boldsymbol{x}^{k+1}$ as the current mode field.
4) Set $k = k+1$, go to Step 2.

In practice, neither Steps 2 or 3 need be iterated to convergence before moving to the next step. We have no convergence proof for this method. However, in running the method for a wide range conditions, we have never observed a case in which the method does not converge.

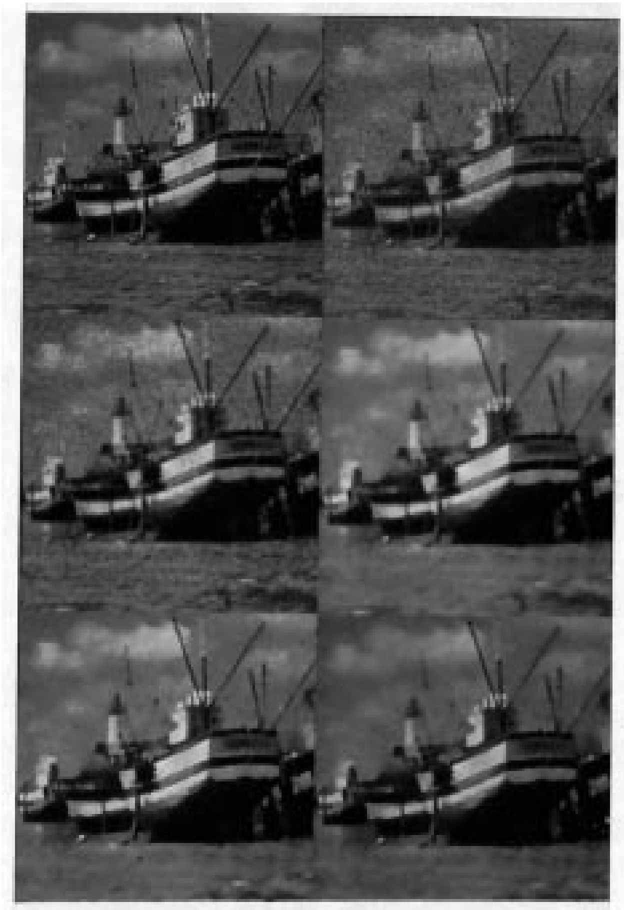


Fig. 5. Experiment for image restoration from Gaussian data, $\sigma^2 = 100$, with blurring kernel Opt-2. Top row: left, original; right, noisy, blurred data. Middle row: left, MAP with quadratic prior; right, MAP with Huber prior. Bottom row: left, MAP with log-quadratic prior; right, MAP with saturated-quadratic prior. All images shown above correspond to the estimated β use MFAML-Gen.

B. Computing the MAP Image Estimate

For a Gibbs prior of the form (5), the MAP estimate is found by maximizing over the log posterior density

$$\overline{\boldsymbol{x}}(\beta) = \arg\max_{\boldsymbol{x}} \left\{ -\tfrac{1}{2}(\boldsymbol{y} - A\boldsymbol{x})^T \mathbf{C}_n^{-1}(\boldsymbol{y} - A\boldsymbol{x}) - \beta \sum_j \sum_{k>j,\, k\in N_j} \kappa_{jk} V(x_j, x_k) \right\} \tag{68}$$

for the Gaussian likelihood, and as shown in (69), at the bottom of the next page, for the Poisson likelihood.

These functions are concave for V_1 and V_2 but not for V_3 and V_4. Gradient-based optimization will therefore lead only to local maxima for the last two potential functions. However, it is widely accepted that for most practical applications a local optimum is acceptable. We therefore restrict attention here to local search methods, although the MFAML method described above can be combined with any numerical procedure for computing a MAP image estimate. Many computational methods for solving large inverse problems in image processing have been studied in recent years. These include Gauss–Siedel procedures (sequential coordinate descent algorithm) [6], con-

TABLE I

MONTE CARLO TEST ($n = 50$) COMPARING PERFORMANCE OF GENERALIZED MAXIMUM PSEUDOLIKELIHOOD (GMPL), THE METHOD OF MOMENTS (MOM), AND THE TWO MODE FIELD APPROXIMATED ML METHODS (MFAML) (A * INDICATES THE ALGORITHM FAILED TO CONVERGE)

True β	0.0004	0.0010	0.0040	0.0100	0.0400	0.100
GMPL Mean	4.134e-4	1.093e-3	7.742e-3	*	*	*
GMPL Bias (%)	3.35%	9.30%	93.6%	*	*	*
GMPL STD (%)	1.74%	1.60%	6.49%	*	*	*
MOM Mean	4.039e-4	1.012e-3	4.175e-3	1.154e-2	0.0775	0.3565
MOM Bias (%)	0.97%	1.20%	4.37%	15.4%	93.7%	257 %
MOM STD (%)	2.13%	3.21%	10.9%	31.0%	340%	380 %
MFAML-Gen Mean	4.010e-4	1.009e-3	4.114-3	1.071e-2	0.0472	0.1241
MFAML-Gen Bias (%)	0.25%	0.89%	2.84%	7.11%	17.7%	24.1%
MFAML-Gen STD (%)	1.64%	1.35%	1.63%	2.23%	3.98%	8.81%
MFAML-Opt Mean	4.153e-4	1.004e-3	3.977e-3	9.526e-3	0.0357	0.07842
MFAML-Opt Bias (%)	3.8%	0.42%	-0.572%	-4.74%	-10.7%	-21.58%
MFAML-Opt STD (%)	1.21%	1.37%	1.27%	2.51%	4.02%	9.02%

jugate gradient methods [27], [30], the method of iterated conditional modes [3], iterated conditional average (ICA) [18], [25], and generalized EM methods. The performance of these algorithms in terms of computation cost and convergence rate is highly problem dependent. We have previously found that preconditioned conjugate gradient methods produce favorable performance for image restoration and reconstruction problems [27] and use this approach in the results presented below. Note that this method includes the use of a penalty function to impose a nonnegativity constraint on the MAP estimate.

C. Computing the Hyperparameter Value

The method that we use to implement Step 3 is an EM-like algorithm. We adopted this approach after finding problems with numerical stability when using a standard Newton–Raphson procedure. For hyperparameter estimation using the mean field approximation based on the GBF bound, we perform Step 3 as follows.

[3a] Compute the mean field approximated statistic $U_{\rm MF}^{(k+1)}(\boldsymbol{x})$ defined as the current left hand side of the mean field likelihood equation (53)

$$U_{\rm MF}^{(k+1)}(\boldsymbol{x}) = \sum_i \mathcal{E}[U_i^{PO}(x_i)|x_j^{k+1}, j \in N_i^{PO}, \boldsymbol{y}, \beta] - \sum_i \sum_{j \in N_i} \kappa_{ij} x_i^{k+1} x_j^{k+1}. \tag{70}$$

[3b] Compute the new hyperparameter value β^{k+1} by solving the equation

$$\sum_i \mathcal{E}[U_i^{PR}(x_i)|\langle x_j\rangle_{\rm mf}^{PR}, j \in N_i; \beta] - \sum_i \sum_{j \in N_i} \kappa_{ij} \langle x_i\rangle_{\rm mf}^{PR} \langle x_j\rangle_{\rm mf}^{PR} = U_{\rm MF}^{(k+1)}(\boldsymbol{x}). \tag{71}$$

For the general approximation, we use the following method to solve Step 3.

[3a] Compute the mean field approximated statistic $U_{\rm MF}^{(k+1)}(\boldsymbol{x})$ defined as the current left hand side of the mean field likelihood equation (67)

$$U_{\rm MF}^{(k+1)}(\boldsymbol{x}) = \sum_i \mathcal{E}[U_i^{PO}(x_i)|x_j^{k+1}, j \in N_i^{PO}, \boldsymbol{y}, \beta]. \tag{72}$$

[3b] Compute the new hyperparameter value β^{k+1} by solving the equation

$$\sum_i \mathcal{E}[U_i^{PO}(x_i)|x_{S\backslash i}^{k+1}; \beta] = U_{\rm MF}^{(k+1)}(\boldsymbol{x}). \tag{73}$$

In Step [3b] of this EM-like algorithm, the new hyperparameter value is computed using one or more iterations of a Newton–Raphson procedure. All integrals encountered were computed numerically using an adaptive quadrature method [29]. We also use a scaling procedure to ensure that the single pixel sample space is approximately [0, 1]. This can be achieved by a corresponding inverse scaling of the elements of the A operator in the likelihood function. This has the effect of avoiding large numerical errors when computing integrals containing integrands of the form $\exp\{-\beta \sum_{j \in N_i} V(x_i, x_j)\}$.

D. Computational Cost

The computational cost of the algorithm we describe above is highly problem dependent. We usually run 5–10 iterations of the conjugate gradient algorithm to update the MAP image estimate for a given value of β, and then use one or two Newton–Raphson iterations to update the value of β. We typically repeat this procedure 10–20 times to achieve effective convergence in β. We have observed that the number of iterations required increases with both the degree of blurring and the variance of the additive noise. For image restoration

$$\overline{\boldsymbol{x}}(\beta) = \arg\max_{\boldsymbol{x}} \left\{ \sum_i \left[-\sum_j A_{ij} x_j + y_i \ln \left(\sum_j A_{ij} x_j \right) \right] - \beta \sum_j \sum_{k>j, k \in N_j} \kappa_{jk} V(x_j, x_k) \right\} \tag{69}$$

TABLE II
MONTE CARLO TEST OF MFAML-OPT AND MFAML-GEN PERFORMANCE AS A FUNCTION OF ADDITIVE NOISE VARIANCE

True β	σ^2	MFAML-Gen			MFAML-Opt		
		mean	Bias (%)	Var (%)	mean	Bias (%)	Var (%)
0.001	4	1.009e-3	0.9%	1.35%	1.004e-3	0.42%	1.37%
	16	1.053e-3	52.8%	1.39%	1.005e-3	0.47%	1.68%
	36	1.111e-3	11.1%	1.85%	9.863e-4	-1.37%	1.99%
	100	1.232e-3	23.2%	3.83%	8.844e-4	-11.5%	2.19%
	400	1.348e-3	34.8%	10.4%	7.450e-4	-25.5%	12.1%

TABLE III
ROBUSTNESS OF MFAML-GEN AND MFAML-OPT TO DIFFERENT SMOOTHING OPERATORS

True β	3×3 operator	MFAML-Gen			MFAML-Opt		
		mean	Bias (%)	Var (%)	mean	Bias (%)	Var (%)
0.004	Opt 1	4.114e-3	2.84%	1.63%	3.977e-3	-0.572%	1.27%
	Opt 2	9.821e-3	145%	3.78%	5.867e-3	46.6%	2.07%
	Opt 3	1.079e-2	170%	3.31%	5.840e-3	46.0%	2.14%
0.01	Opt 1	1.071e-2	7.11%	2.23%	9.526e-3	-4.74%	2.51%
	Opt 2	3.401e-2	240%	5.88%	1.060e-2	5.96%	1.97%
	Opt 3	3.628e-2	262%	10.2%	1.085e-3	8.48%	2.01%
0.04	Opt 1	0.0472	17.7%	3.89%	0.0357	-10.7%	4.02%
	Opt 2	0.1030	157%	8.7%	0.0215	-46.2%	1.26%
	Opt 3	0.1055	164%	9.71%	0.0225	44.1%	5.21%

with local blurring only, the dominant computational costs are the Newton–Raphson iterations required for updating the hyperparameter. On a SunSPARC 20/61 workstation, each iteration of the conjugate gradient MAP algorithm for a 256×256 pixel image requires only a few seconds. Each iteration of the Newton–Raphson algorithm can take from several seconds to several minutes because each iteration requires $3 \times 256 \times 256$ 1-D numerical integrations. For problems with Gaussian likelihoods and quadratic priors, we can replace the numerical integrals with an error function look-up table, thus reducing the per iteration cost to a few seconds.

V. PERFORMANCE STUDIES

We have applied the mode field approximated maximum likelihood (MFAML) method to image restoration and reconstruction. We present the results for image restoration below. Application of this method to parameter estimation in positron emission tomography (PET), where the data are Poisson, is described in [28] and [41]. We simply note here that we have observed similar performance for the PET problem to that described below for image restoration.

A. Estimator Bias and Variance Using Stochastic Sampling

We used extensive Monte Carlo simulations to evaluate the performance of the new MFAML hyperparameter estimators in the problem of image restoration from blurred data with additive Gaussian noise. We have compared the performance of the MFAML methods described above with generalized maximum pseudolikelihood (GMPL) and the method of moments (MOM), for which the statistic $M(\boldsymbol{y})$ takes the same form as the Gibbs energy function of the prior, computed over the noisy image $\boldsymbol{y}$ with an eight nearest neighbor interaction.

We performed Monte Carlo studies for image restoration as follows. For each value of the hyperparameter, 50 sample images were drawn from a specific prior using the Metropolis algorithm [24]. Each sampled image was then blurred by one of the following 3×3 kernels:

$$\text{Opt 1:} \quad \begin{pmatrix} 0.001 & 0.028 & 0.001 \\ 0.028 & 0.884 & 0.028 \\ 0.001 & 0.028 & 0.001 \end{pmatrix}$$

$$\text{Opt 2:} \quad \begin{pmatrix} \frac{1}{16} & \frac{1}{8} & \frac{1}{16} \\ \frac{1}{8} & \frac{1}{4} & \frac{1}{8} \\ \frac{1}{16} & \frac{1}{8} & \frac{1}{16} \end{pmatrix}$$

and

$$\text{Opt 3:} \quad \begin{pmatrix} \frac{1}{9} & \frac{1}{9} & \frac{1}{9} \\ \frac{1}{9} & \frac{1}{9} & \frac{1}{9} \\ \frac{1}{9} & \frac{1}{9} & \frac{1}{9} \end{pmatrix}.$$

Note that the degree of smoothing increases from Opt 1–3. Pseudorandom Gaussian noise with known variance σ was generated to contaminate each of the resulting blurred images. The likelihood function for these noisy data take the form of (1). The hyperparameters were estimated for each method of interest for each of the 50 noisy images. Since the original images are sampled from specific priors with known hyperparameter values, we were able to calculate bias and variance across the 50 resulting estimates.

A comparison of the performance of the various methods for a range of values of β is shown in Table I. The original images were generated using the Metropolis algorithm with the

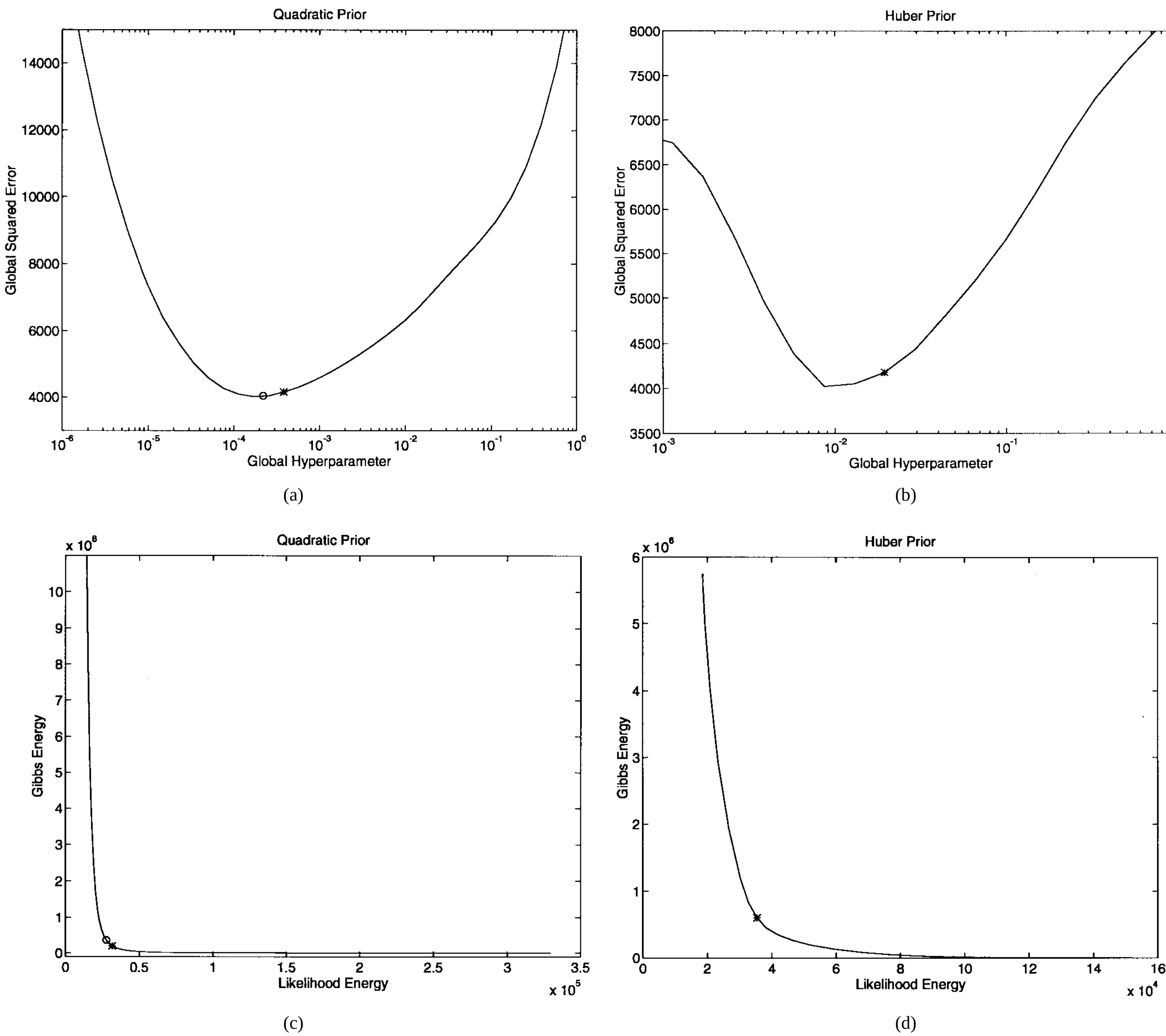


Fig. 6. Total squared error versus β for (a) quadratic and (b) Huber priors. Shown are the L-curves for (c) quadratic and (d) Huber priors. A "$*$" indicates the β value obtained using MFAML-Gen and "o" using MFAML-Opt (quadratic case only). Note the estimated β values correspond approximately to the minimum squared error.

quadratic prior and a single pixel sample space [0, 100]. These were then blurred using Opt 1 and contaminated by zero mean Gaussian noise with variance $\sigma^2 = 4$. All methods perform best when β is small and deteriorate as β increases and the images become smoother. The GMPL method works only for the smaller values of β. As β increases, the two-step method, which iterates between MAP estimation of the image x and estimation of β, fails to converge. The MOM method performs better in general, but as β increases, the slope of the moment curve decreases, leading to increased bias and variance. In all cases, both the general and optimal forms of MFAML outperform both of the other techniques. The differences are very clear for the cases where β is large, which corresponds to the case of increasingly smooth images. For these larger β values, MFAML shows approximately a tenfold reduction in bias and variance relative to the MOM method. The optimal form of MFAML exhibits lower bias than the general form, with slightly larger variance and overall superior performance. However, in practice these differences are small and lead to little noticeable difference in image quality when applied to real images.

To test the robustness of the MFAML methods to noise, we used the same set-up as in the comparative studies above and generated data for a range of additive noise variances. As before, ensemble statistics were computed to determine the effects of different noise levels on the bias and variance of β. We summarize these results in Table II. Although we do observe deterioration in the performance when noise variance increases, both MFAML methods appear to perform well and are stable even for very large additive noise variances.

The conditioning of the likelihood affects the degree of ill-posedness of the inverse problem, i.e., the conditioning of the

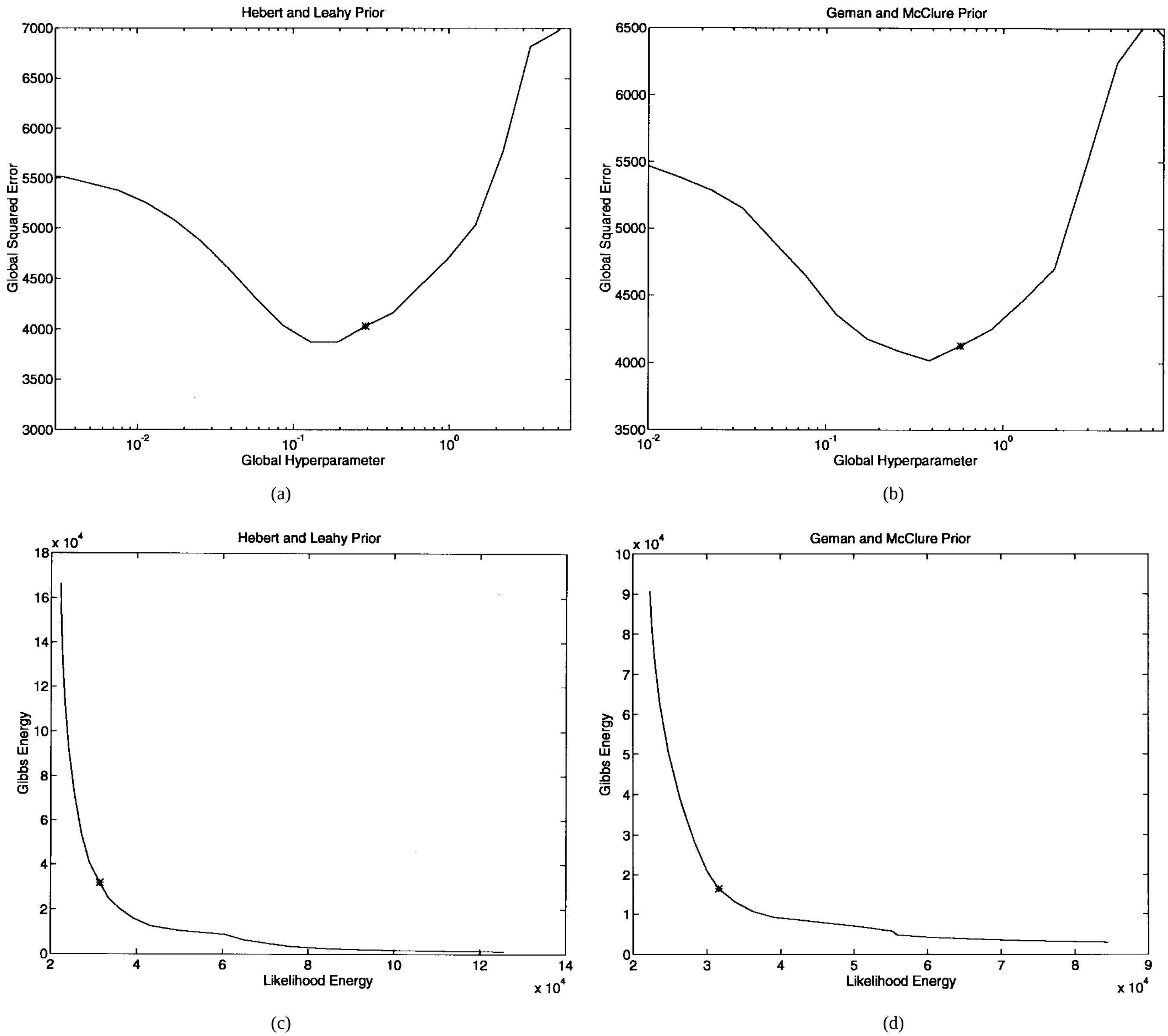


Fig. 7. Total squared error versus β for (a) log-quadratic and (b) saturated-quadratic priors. Shown are the L-curves for (c) log-quadratic and (d) saturated-quadratic priors. A "*" indicates the β value obtained using MFAML-Gen. Note the estimated β values correspond approximately to the minimum squared error.

operator $\boldsymbol{A}$ determines our ability to recover the image x from the blurred data, which in turn affects our ability to accurately estimate β. Results in Table III show that as the degree of blurring increases and the inverse problem becomes more ill posed, performance of the MFAML methods deteriorates. The bias in the estimator appears to be more affected than variance by changes in the degree of blurring. Note also that in this example, there are more substantial differences in performance between the general and optimal MFAML methods than was seen in Table I. For the Opt-2 and Opt-3 blurring kernels, GMPL does not converge and MOM is unable to identify the parameter due to the flatness of the moment curve.

B. Applications and Validations with Real Images

In this experiment, we used the 3×3 blurring mask Opt 2 to blur the 256×256 pixel Boat image. The single pixel sample space of the Boat image is [0, 255]. We generated Gaussian noise with a variance of 100 to contaminate the resulting blurred Boat image. The images were then restored using MAP estimation for each of the four potential functions in (8) and the appropriate likelihood function. Images were reconstructed for a range of fixed values of β and the total squared error between the original and restored image calculated. The images were then reconstructed again with simultaneous MFAML estimation of β. For the case of Gaussian noise and the quadratic prior we use both the MFAML-Gen and MFAML-Opt estimators. In all other cases we use only the MFAML-Gen method.

The restored images for the cases where β is estimated are shown in Fig. 5. The corresponding curves showing the restored image error as a function of hyperparameter are shown in Figs. 6 and 7. Note the log-scale on the β axis.

We show the location of the MFAML-Gen and MFAML-Opt estimate of the hyperparameter on the curves using "*" and "O," respectively. Also shown in these figures is the corresponding L-curve, again with the location of the estimated hyperparameter indicated. These results show that the estimated hyperparameter gives close to the minimum squared error in all cases, and is located close to the knee of the L-curve.

VI. Conclusion

We have described a general method for estimating the hyperparameter of Gibbs priors from incomplete data. This method is based on a mean-field-like approximation of the Gibbs distributions involved. The result provides a balance between the oversimplified model implicit in the generalized ML methods and the intractability of a true ML estimator. While computational costs are significant, we anticipate they will be acceptable in practical situations. Convergence of the method by which the solution is computed simultaneously with a MAP image estimate has not been shown; however, we have not encountered any problems with convergence in the many cases we have run.

The results presented indicate that good performance is achieved over a range of conditions when applied to image restoration. We have also observed similar behavior in applications to PET [28], [41]. We do observe that the estimator degrades as the degree of blurring increases. This is inevitable in the sense that the ultimate performance of the method is limited by the slope of the likelihood function $p(\boldsymbol{y}|\beta)$. The method described here is not limited to estimation of the specific problems described. It appears straightforward to modify this approach for estimation of the hyperparameters of discrete spatial processes such as those used for image segmentation and labeling.

References

[1] M. Almeida and G. Gidas, "A variational method for estimating the parameters of MRF from complete or incomplete data," Tech. Rep., Brown Univ., Providence, RI, 1989.

[2] J. Besag, "Spatial interaction and the statistical analysis of lattice systems (with discussion)," *J. Roy. Stat. Soc. B*, vol. 36, pp. 192–326, 1974.

[3] ——, "On the statistical analysis of dirty pictures," *J. Roy. Stat. Soc. B*, vol. 48, no. 3, pp. 259–302, 1986.

[4] J. Besag, P. Green, D. Higdon, and K. Mengersen, "Bayesian computation and stochastic systems," *Stat. Sci.*, vol. 10, pp. 3–66, 1995.

[5] A. Blake and A. Zisserman, "Visual reconstruction," in *Artificial Intelligence.* Cambridge, MA: MIT Press, 1987.

[6] C. Bouman and K. Sauer, "Fast numerical methods for emission and transmission tomographic reconstruction," in *Proc. Conf. Information Science and Systems*, Johns Hopkins Univ., Baltimore, MD, 1993.

[7] D. Chandler, *Introduction to Modern Statistical Mechanics.* Oxford, U.K.: Oxford Univ. Press, 1987.

[8] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Stat. Soc.*, vol. 29, pp. 1–38, 1977.

[9] P. Craven and G. Wahba, "Smoothing noisy data with spline function," *Numerische Math.*, vol. 31, pp. 377–403, 1979.

[10] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-6, pp. 721–741, 1984.

[11] S. Geman and D. McClure, "Statistical methods for tomographic image reconstruction," in *Proc. 46th Sess. ISI, Bulletin ISI,* 1987, vol. 52.

[12] D. Geiger and F. Girosi, "Parallel and deterministic algorithms for MRFs: Surface reconstruction and integration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 12, pp. 401–412, 1991.

[13] Geyer and Thompson, "Constrained Monte Carlo maximum likelihood for dependent data," *J. Roy. Stat. Soc. B*, vol. 54, pp. 657–699, 1992.

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

[15] P. C. Hansen, "Analysis of discrete ill-posed problems by means of the L-curve," *SIAM Rev.*, vol. 34, no. 4, pp. 561–580, 1992.

[16] P. C. Hansen and D. P. O'Leary, "The use of the L-curve in the regularization of discrete ill-posed problems," Rep. UMIACS-TR-91-142, Dept. Comput. Sci., Univ. Maryland, College Park, MD.

[17] T. J. Hebert and R. Leahy, "Statistic-based MAP image reconstruction from Poisson data using Gibbs priors," *IEEE Trans. Signal Processing*, vol. 40, no. 9, pp. 2290–2303, Sept. 1992.

[18] V. E. Johnson, W. H. Wong, X. Hu, and C. Chen, "Image restoration using Gibbs priors: Boundary modeling, treatment of blurring, and selection of hyperparameter," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 13, pp. 413–425, 1991.

[19] S. Lakshmanan and H. Derin, "Simultaneous parameter estimation and segmentation of Gibbs random fields using simulated annealing," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 11, pp. 799–813, 1989.

[20] R. Leahy and X. Yan, "Incorporation of anatomical MR data for improved functional imaging with PET," in *Information Processing in Medical Imaging*, A. C. F. Colchester and D. J. Hawkes, Eds. New York: Springer-Verlag, 1991, pp. 105–120.

[21] P. J. M. van Laarhoven and E. H. L. Arts, *Simulated Annealing: Theory and Applications.* Amsterdam, The Netherlands: Reidel, 1987.

[22] D. Luenberger, *Linear and Nonlinear Programming.* Reading, MA: Addison-Wesley, 1989.

[23] J. Mathews and R. L. Walker, *Mathematical Methods of Physics.* Menlo Park, CA: Benjamin Cummings, 1970.

[24] J. Marroquin, "Probabilistic solution of inverse problems," Ph.D. dissertation, Mass. Inst. Technol., Cambridge, MA, 1985.

[25] K. M. Manbeck, "Bayesian statistical methods applied to emission tomography with physical phantom and patient data," Ph.D. dissertation, Brown Univ., Providence, RI, 1990.

[26] A. Mohammad-Djafari, "On the estimation of hyperparameters in Bayesian approach of solving inverse problems," in *Proc. ICASSP*, 1993, pp. V495–V498.

[27] E. Mumcuoglu, R. Leahy, S. Cherry, and Z. Zhou, "Fast gradient-based methods for Bayesian reconstruction of transmission and emission PET images," *IEEE Trans. Med. Imag.*, vol. 13, no. 4, pp. 687–701, 1994.

[28] E. Mumcuoglu, R. Leahy, and S. Cherry, "Bayesian reconstruction of PET images: Methodology and performance analysis," *Phys. Med. Biol.*, vol. 41, no. 9, pp. 1–31, 1996.

[29] H. R. Schwarz and J. Waldvogel, *Numerical Analysis—A Comprehensive Introduction.* New York: Wiley, pp. 339–342.

[30] A. Rangarajan, "Representation and recovery of discontinuities in some early vision problems," Ph.D. dissertation, Univ. Southern California, Los Angeles, 1990.

[31] G. Gindi, M. Lee, A. Rangarajan *et al.*, "A continuation method for emission tomography," in *Proc. 1992 Nuclear Science Symp. and Medical Imaging Conf.*, 1992, pp. 1204–1206.

[32] S. Saquib, C. Bouman, and K. Sauer, "ML parameter estimation for Markov random fields, with applications to Bayesian tomography," submitted for publication.

[33] A. M. Thompson *et al.*, "A study of methods of choosing the smoothing parameter in image restoration by regularization," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 13, pp. 326–339, 1991.

[34] J. Varah, "Pitfalls in the numerical solution of ill-posed problems," *SIAM J. Sci. Stat. Comput.*, vol. 4, pp. 164–176, 1983.

[35] G. Wahba, "Spline models for observational data," in *Proc. CBMS-NSF Reg. Conf. Ser. Applied Mathematics*, vol. 59, Soc. Ind. Appl. Math., Philadelphia, PA, 1990.

[36] J. Zhang, "The mean field theory in EM procedures for Markov random fields," *IEEE Trans. Signal Processing*, vol. 40, pp. 2570–2583, Oct. 1992.

[37] J. Zhang and J. Hanauer, "The mean field theory for image motion estimation," in *Proc. ICASSP*, 1993, pp. V197–V200.

[38] J. Zhang, "The mean field theory in EM procedures for blind Markov random field image restoration," *IEEE Trans. Image Processing*, vol. 2, pp. 27–40, Jan. 1993.

[39] J. Zhang, J. W. Modestino, and D. A. Langan, "Maximum-likelihood parameter estimation for unsupervised stochastic model-based image segmentation," *IEEE Trans. Image Processing*, vol. 3, pp. 404–420, Apr. 1994.

[40] Z. Zhou, R. Leahy, and E. Mumguoclu, “A comparative study of using anatomical boundary in PET reconstruction,” in *Proc. IEEE Nuclear Science Symp. and Medical Imaging Conf.,* 1993, pp. 1749–1753.

[41] ——, “Maximum likelihood hyperparameter estimation for Gibbs priors from incomplete data with applications to positron emission tomography,” in *Information Processing in Medical Imaging,* Y. Bizais, C. Barillot, and R. Di Paola, Eds. Boston, MA: Kluwer, 1995, pp. 39–51.

[42] Z. Zhou, “Maximum likelihood hyperparameter estimation for Gibbs priors from incomplete data with applications in image processing,” Ph.D. dissertation, Univ. Southern California, Los Angeles, 1994.

Zhenyu Zhou (S’91–M’95) was born in China on June 26, 1964. He received the B.E. and M.E. degrees in information science and electronic engineering from Zhejiang University, Hangzhou, China, in 1987 and 1989, respectively. He received the M.S. and Ph.D. degrees from the Signal and Image Processing Institute, Department of Electrical Engineering-Systems, University of Southern California (USC), Los Angeles, in 1992 and 1995, respectively.

From 1990 to 1995, he was with the Signal and Image Processing Institute, USC, as a Research Assistant or Research Associate. He primarily worked on image processing theory and its medical applications. Since 1995, he has been with the Multimedia Communication Division, Rockwell International Inc., Newport Beach, CA. His current research interests include the application of signal and image processing techniques in various multimedia communication systems, video and speech compression algorithms, and medical imaging.

Richard M. Leahy (M’85) was born in Surrey, England, in 1960. He received the B.Sc. and Ph.D. degrees in electrical engineering from the University of Newcastle upon Tyne, U.K., in 1981 and 1985, respectively.

In 1985, he joined the University of Southern California (USC), Los Angeles, where he is currently an Associate Professor and Director of the Signal and Image Processing Institute, Department of Electrical Engineering-Systems. He holds joint appointments with the Departments of Radiology and Biomedical Engineering at USC. His research interests lie primarily in the application of signal and image processing theory to biomedical inverse problems. His current research includes the development of inverse methods for reconstruction of functional brain images and methods for the analysis of 3-D brain images.

Jinyi Qi (SM’94) was born in China in 1970. He received the B.S. degree (with highest honors) from the Tsinghua University, Beijing, China, in 1993. He is currently pursuing the Ph.D. degree in electrical engineering at the University of Southern California (USC), Los Angeles.

Since January 1996, he has been a Research Assistant in the Department of Electrical Engineering-Systems, USC. His research interests include signal and image processing, wavelets, parameter estimation, and image reconstruction.