Advanced Structured Prediction

Editors:

Sebastian Nowozin Sebastian.Nowozin@microsoft.com
Microsoft Research
Cambridge, CB1 2FB, United Kingdom

Peter V. Gehler pgehler@tuebingen.mpg.de
Max Planck Insitute for Intelligent Systems
72076 Tübingen, Germany

Jeremy Jancsary jermyj@microsoft.com
Microsoft Research
Cambridge, CB1 2FB, United Kingdom

Christoph H. Lampert chl@ist.ac.at
IST Austria
A-3400 Klosterneuburg, Austria

This is a draft version of the author chapter.

The MIT Press
Cambridge, Massachusetts
London, England

1 Perturb-and-MAP Random Fields: Reducing Random Sampling to Optimization, with Applications in Computer Vision

George Papandreou gpapan@ttic.edu
Toyota Technological Institute at Chicago
Chicago, USA

Alan Yuille yuille@stat.ucla.edu
University of California, Los Angeles
Los Angeles, USA

Probabilistic Bayesian methods such as Markov random fields are well suited for modeling structured data, providing a natural conceptual framework for capturing the uncertainty in interpreting them and automatically learning model parameters from training examples. However, Bayesian methods are often computationally too expensive for large-scale applications compared to deterministic energy minimization techniques.
This chapter presents an overview of a recently introduced "Perturb-and-MAP" generative probabilistic random field model, which produces in a single shot a random sample from the whole field by first injecting noise into the energy function, then solving an optimization problem to find the least energy configuration of the perturbed system. Perturb-and-MAP random fields thus turn fast deterministic energy minimization methods into computationally efficient probabilistic inference machines and make Bayesian inference practically tractable for large-scale problems, as illustrated in challenging computer vision applications such as image inpainting and deblurring, image segmentation, and scene labeling.

Keywords*: MRF, energy minimization, Perturb-and-MAP, extreme value statistics, graph cuts, random sampling.*

1.1 Introduction

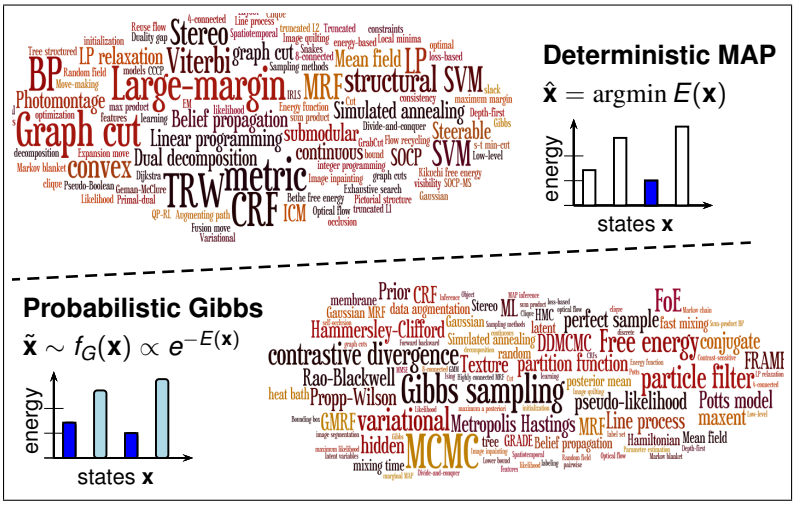


Figure 1.1: Deterministic energy minimization vs. probabilistic Gibbs modeling. Perturb-and-MAP attempts to bridge the gap between these two approaches.

Structured prediction models are typically built around an energy function, which assigns to each possible configuration vector $\boldsymbol{x} = (x_1, \ldots, x_N)$ a real-valued energy $E(\boldsymbol{x})$, with more preferable configurations getting lower energies.

As illustrated in Figure 1.1, there are two quite distinct ways to work with energy-based models. The first is entirely deterministic and amounts to finding a single most probable (MAP) configuration of minimum energy, $\hat{\boldsymbol{x}} = \text{argmin}_{\boldsymbol{x}}\, E(\boldsymbol{x})$. Such deterministic methods are computationally appealing thanks to very fast energy minimization algorithms that can efficiently optimize large scale problems for important families of energy functions involving both continuous and discrete-valued variables. Parameter learning in this setting is typically performed with large-margin methods.

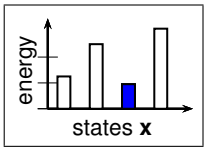
The second class of methods is probabilistic, assigning to each state a Gibbs probability $f_G(\boldsymbol{x}) \propto e^{-E(\boldsymbol{x})}$. Their key advantage over MAP inference is that they also allow uncertainty quantification in interpreting ambiguous data. The probabilistic framework also enables learning model parameters from training examples using maximum likelihood. However, probabilistic inference is in general considerably more difficult than optimization, since it requires capturing multiple solutions plausible under the posterior distribution instead of just a single MAP configuration.

This chapter presents an overview of the recently introduced Perturb-and-MAP method, which attempts to reduce probabilistic inference to an energy minimization problem, thus establishing a link between the optimization and probabilistic inference approaches to energy-based modeling. As illustrated in Figure 1.2, Perturb-and-MAP is a two-step generative process: (1) In a Perturb step, we inject additive random noise $N(\boldsymbol{x})$ into the system's energy function, followed by (2) a MAP step in which we find the minimum energy configuration of the perturbed system. By properly designing the noise injection process we can generate exact Gibbs samples from Gaussian MRFs and good approximate samples from discrete-label MRFs.

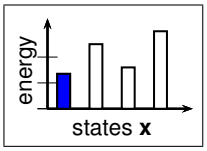
Of course, studying the output sensitivity to input perturbations is omnipresent under many different guises not only in machine learning but also in optimization, signal processing, control, computer science, and theoretical psychology, among others. However, Perturb-and-MAP is unique in using random perturbations as the defining building block of a structured probabilistic model and setting the ambitious goal of replicating the Gibbs distribution using this approach.

function PERTURB-AND-MAP
$\tilde{E}(\boldsymbol{x}) = E(\boldsymbol{x}) + N(\boldsymbol{x})$ ▷ Perturb
$\tilde{\boldsymbol{x}} = \text{argmin}_{\boldsymbol{x}}\, \tilde{E}(\boldsymbol{x})$ ▷ MAP
return $\tilde{\boldsymbol{x}}$ ▷ Random sample
end function

(a)



(b)

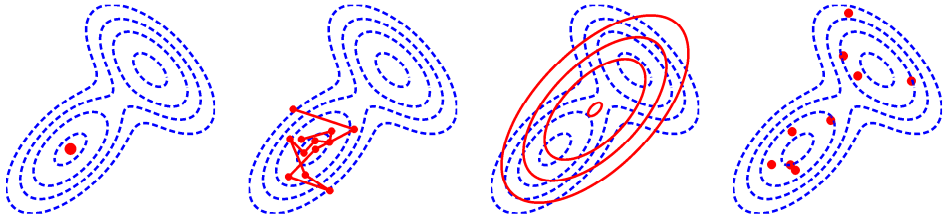


(c)

Figure 1.2: (a) The generic Perturb-and-MAP random sampling algorithm. (b) Original energies $E(\boldsymbol{x})$. (c) Perturbed energies $\tilde{E}(\boldsymbol{x})$. The MAP state $\hat{\boldsymbol{x}}$ and the Perturb-and-MAP sample $\tilde{\boldsymbol{x}}$ are shown shaded in (b) and (c), respectively.

While deterministic MAP inference summarizes the solution space into a single most probable estimate, Perturb-and-MAP gives other low energy states the chance to arise as random samples for some instantiations of the perturbation noise and is thus able to represent the whole probability landscape. Perturb-and-MAP follows a fundamentally different approach compared to other approximate probabilistic inference methods such as Markov Chain Monte-Carlo (MCMC) and Variational Bayes (VB), which are contrasted with Perturb-and-MAP in Figure 1.3. MCMC is broadly applicable and can provide very accurate results but is typically computationally very expensive for large scale problems. When the distribution has multiple modes, MCMC mixes slowly and becomes particularly ineffective because it moves in small steps through the state space. Crucially, Perturb-and-MAP generates samples in a single shot, completely bypassing the Markov Chain

slow mixing problem, and thus has no difficulty in dealing with multimodal distributions. Variational Bayesian methods such as mean field or variational bounding approximate a complicated probability landscape with a simpler parametric distribution. VB is typically faster yet less accurate than MCMC, and also faces difficulties in the presence of multiple modes.



(a) (b) (c) (d)

Figure 1.3: Capturing a complicated probability landscape (in dashed lines) with standard approximate inference methods vs. Perturb-and-MAP. (a) Deterministic MAP. (b) Markov Chain Monte-Carlo. (c) Variational Bayes. (d) Perturb-and-MAP.

Perturb-and-MAP was initially developed for drawing exact random samples from Gaussian MRFs. This efficient Gaussian sampling algorithm can also be used as sub-routine and considerably accellerate both MCMC and VB in applications involving continuous sparse potentials. We discuss these in Section 1.3. This line of research led to the development of Perturb-and-MAP for discrete MRFs, which we discuss in Section 1.4. We present a summary of recent related work in Section 1.5.

1.2 Energy-Based Modeling: Standard Deterministic and Probabilistic Approaches

1.2.1 Energies and Gibbs MRFs for Modeling Inverse Problems

Structured prediction for solving inverse problems is typically formulated in terms of energy functions. Given an input vector of noisy measurements $\boldsymbol{y}$, our goal is to estimate the latent state output vector $\boldsymbol{x} = (x_1, \ldots, x_N)$. The elements of the state vector $x_i \in \mathcal{L}$ can take either continuous or discrete values from the label set $\mathcal{L}$. As shown in Figure 1.4, in image processing applications such as image inpainting or deblurring the state vector $\boldsymbol{x}$ corresponds to a real-valued clean image that we wish to recover from its partial or degraded version $\boldsymbol{y}$. In computer vision applications such as image segmentation or labeling the state vector $\boldsymbol{x}$ corresponds to an assignment of image areas to different image segments or semantic object

classes. Probabilistic Bayesian techniques offer a natural framework for combining the measurements with prior information in tackling such inverse problems.

(a) (b) (c) (d)

Figure 1.4: In inverse modeling we use observations $\boldsymbol{y}$ (top row) to infer a latent interpretation $\boldsymbol{x}$ (bottom row). Image processing examples: (a) Inpainting. (b) Deblurring. Computer vision examples: (c) Figure-ground segmentation. (d) Scene labeling.

Given a specific measurement $\boldsymbol{y}$, we quantify a particular interpretation $\boldsymbol{x}$ by means of a *deterministic* energy function $E(\boldsymbol{x})$, where for notational convenience we are suppressing its dependence on the measurements $\boldsymbol{y}$. We will be working with energy functions of the general form

$$E(\boldsymbol{x};\boldsymbol{\theta}) = \langle \boldsymbol{\theta},\, \boldsymbol{\phi}(\boldsymbol{x}) \rangle = \sum_{j=1}^{M} \theta_j \phi_j(\boldsymbol{x})\,, \tag{1.1}$$

where $\boldsymbol{\theta} \in \mathbb{R}^M$ is a real-valued parameter vector of length M, and $\boldsymbol{\phi}(\boldsymbol{x}) = (\phi_1(\boldsymbol{x}), \ldots, \phi_M(\boldsymbol{x}))^T$ is a vector of potentials or "sufficient statistics". We can interpret θ_j as the weight assigned to the feature $\phi_j(\boldsymbol{x})$: we have many different design goals or sources of information (e.g., smoothness prior, measurements), each giving rise to some features, whose weighted linear combination constitutes the overall energy function. Each potential often depends on a small subset of the latent variables, which is made explicit in a factor graph representation of the energy function shown in Figure 1.5.

The Gibbs distribution is the standard way to induce a *probabilistic* model from the energy function. It defines a Markov random field whose probability density/mass function has the exponential family form

$$f_G(\boldsymbol{x};\boldsymbol{\theta}) = Z^{-1}(\boldsymbol{\theta}) \exp\left(-E(\boldsymbol{x};\boldsymbol{\theta})\right), \tag{1.2}$$

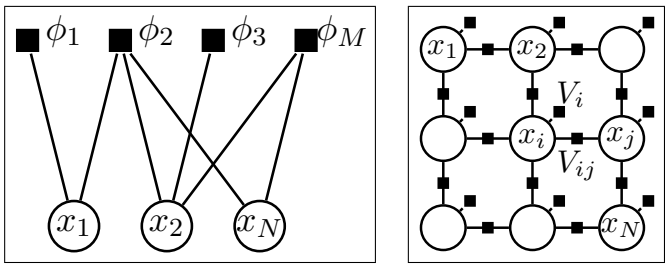


Figure 1.5: (a) The factor graph representation of the energy makes explicit which variables affect each potential. (b) A standard nearest neighbor 2-D grid MRF with unary and pairwise potentials, $\boldsymbol{\phi} = (\{V_i\}, \{V_{ij}\})$.

where $Z(\boldsymbol{\theta}) = \sum_{\boldsymbol{x}} \exp\left(-E(\boldsymbol{x}; \boldsymbol{\theta})\right)$ is the partition function and summation over $\boldsymbol{x}$ should be interpreted as integration in the case of a continuous label space $\mathcal{L}$.

MAP inference in the Gibbs model, i.e., computing the most probable configuration, $\hat{\boldsymbol{x}} = \operatorname{argmax}_{\boldsymbol{x}} f_G(\boldsymbol{x})$, is equivalent to solving the energy minimization problem $\hat{\boldsymbol{x}} = \operatorname{argmin}_{\boldsymbol{x}} E(\boldsymbol{x})$. Thanks to powerful modern energy minimization algorithms, exact or high-quality approximate MAP inference can be performed efficiently for several important energy models. However, other key queries on the Gibbs model such as computing the marginals $f_G(x_i) = \sum_{\boldsymbol{x} \setminus x_i} f_G(\boldsymbol{x})$ or random sampling are computationally hard.

1.2.2 Probabilistic Parameter Learning from Training Examples

While we typically select the feature set $\boldsymbol{\phi}$ by hand, we can exercise much control on the behavior of the energy-based model by setting the parameters $\boldsymbol{\theta}$ to appropriate values. The high-level goal is to select the weight vector $\boldsymbol{\theta}$ in a way that the model assigns low energies to desirable configurations and high energies to "everything else".

When the number of parameters M is small, we can set them to reasonable values by hand. However, a more principled way is to automatically learn the parameters from a training set of K structured labeled examples $\{\boldsymbol{x}_k\}_{k=1}^K$. Discriminative learning criteria such as structured max-margin (Taskar et al., 2003; LeCun et al., 2007; Szummer et al., 2008; Koller and Friedman, 2009) are very powerful and described in detail in other chapters of this volume. Computationally, they are iterative and they typically require modified MAP inference at each parameter update step, which is computationally efficient for many energy models often used in practice.

In the probabilistic setting that is the focus of this chapter, maxi-

mum (penalized) likelihood (ML) is the natural criterion for learning the weights. Given the labeled training set $\{\boldsymbol{x}_k\}_{k=1}^K$, we fit the parameters $\boldsymbol{\theta}$ by maximizing the Gibbs log-likelihood function $L_G(\boldsymbol{\theta}) = -\log Z(\boldsymbol{\theta}) - (1/K)\sum_{k=1}^K E(\boldsymbol{x}_k;\boldsymbol{\theta})$, possibly also including an extra penalty term regularizing the weights. For fully observed models and energies of the form (1.1) the log-likelihood is a concave function of the weights $\boldsymbol{\theta}$ and thus the global maximum can be found by gradient ascent (Hinton and Sejnowski, 1983; Zhu et al., 1998; Koller and Friedman, 2009). The gradient is $\partial L_G/\partial\theta_j = \mathbb{E}^G_{\boldsymbol{\theta}}\{\phi_j(\boldsymbol{x})\} - \mathbb{E}_D\{\phi_j(\boldsymbol{x})\}$. Here $\mathbb{E}^G_{\boldsymbol{\theta}}\{\phi_j(\boldsymbol{x})\} \triangleq \sum_{\boldsymbol{x}} f_G(\boldsymbol{x};\boldsymbol{\theta})\phi_j(\boldsymbol{x}) = -\partial(\log Z)/\partial\theta_j$ and $\mathbb{E}_D\{\phi_j(\boldsymbol{x})\} \triangleq (1/K)\sum_{k=1}^K \phi_j(\boldsymbol{x}_k)$ are, respectively, the expected sufficient statistics under the Gibbs model and the data distribution. Upon convergence, $\mathbb{E}^G_{\boldsymbol{\theta}}\{\phi_j(\boldsymbol{x})\} = \mathbb{E}_D\{\phi_j(\boldsymbol{x})\}$. Thus, ML estimation of the Gibbs model can be thought of as moment matching: random samples drawn from the trained model reproduce the sufficient statistics observed in the training data.

The chief computational challenge in ML parameter learning of the Gibbs model lies in estimating the model sufficient statistics $\mathbb{E}^G_{\boldsymbol{\theta}}\{\phi_j(\boldsymbol{x})\}$. Note that this inference step needs to be repeated at each parameter update step. The model sufficient statistics can be computed exactly in tree-structured (and low tree-width) graphs, but in general graphs one needs to resort to MCMC techniques for approximating them (Hinton and Sejnowski, 1983; Zhu et al., 1998; Hinton, 2002), an avenue considered too costly for many computer vision applications. Deterministic approximations such as variational techniques or loopy sum-product belief propagation do exist, but often are not accurate enough. Simplified criteria such as pseudo-likelihood (Besag, 1975) have been applied as substitutes to ML, but they can sometimes give results grossly different to ML.

Beyond model training, random sampling is very useful in itself, to reveal what are typical instances of the model – what the model has in its "mind" – and in applications such as texture synthesis (Zhu et al., 1998). Further, we might be interested not only in the global minimum energy configuration, but in the marginal densities or posterior means as well (Schmidt et al., 2010). In loopy graphs these quantities are typically intractable to compute, the only viable way being through sampling. Our Perturb-and-MAP random field model is designed specifically so as to be amenable to rapid sampling.

1.3 Perturb-and-MAP for Gaussian and Sparse Continuous MRFs

Gaussian Markov random fields (GMRFs) are an important MRF class describing continuous variables linked by quadratic potentials (Besag, 1974;

Szeliski, 1990; Weiss and Freeman, 2001; Rue and Held, 2005). They are very useful both for modeling inherently Gaussian data and as building blocks for constructing more complex models.

1.3.1 Exact Gaussian MRF Sampling by Local Perturbations

We will be working with a GMRF defined by the energy function

$$E(\boldsymbol{x};\boldsymbol{\theta}) = \frac{1}{2}(\boldsymbol{F}\boldsymbol{x}-\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{F}\boldsymbol{x}-\boldsymbol{\mu}_0) = \frac{1}{2}\boldsymbol{x}^T\boldsymbol{J}\boldsymbol{x}-\boldsymbol{k}^T\boldsymbol{x}+(\text{const}) \quad (1.3)$$

where $\boldsymbol{J} = \boldsymbol{F}^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{F}$, $\boldsymbol{k} = \boldsymbol{F}^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0$. The energy can be cast in the generic inner product form of Equation (1.1) by defining the parameters $\boldsymbol{\theta} = (\boldsymbol{k}, \text{vec}(\boldsymbol{J}))$ and features $\boldsymbol{\phi}(\boldsymbol{x}) = (-\boldsymbol{x}, \frac{1}{2}\text{vec}(\boldsymbol{x}\boldsymbol{x}^T)$. We assume a diagonal matrix $\boldsymbol{\Sigma}_0 = \text{Diag}(\Sigma_1, \ldots, \Sigma_M)$, implying that the energy can be decomposed as a sum of M independent terms $E(\boldsymbol{x};\boldsymbol{\theta}) = \sum_{j=1}^{M}\frac{1}{2\Sigma_j}(\boldsymbol{f}_j^T\boldsymbol{x}-\mu_j)^2$, where $\boldsymbol{f}_j^T$ is the j-th row of the measurement matrix $\boldsymbol{F}$ and μ_j is the j-th entry of the vector $\boldsymbol{\mu}_0$.

The corresponding Gibbs distribution $f_G(\boldsymbol{x})$ is a multivariate Gaussian $\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$ with covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{J}^{-1}$ and mean vector $\boldsymbol{\mu} = \boldsymbol{J}^{-1}\boldsymbol{k}$. The MAP estimate $\hat{\boldsymbol{x}} = \text{argmin}_{\boldsymbol{x}}\frac{1}{2}\boldsymbol{x}^T\boldsymbol{J}\boldsymbol{x}-\boldsymbol{k}^T\boldsymbol{x}$ under this Gaussian model coincides with the mean and amounts to solving the $N\times N$ linear system $\boldsymbol{J}\boldsymbol{\mu} = \boldsymbol{k}$. Solving this linear system with direct exact methods requires a Cholesky factorization of $\boldsymbol{J}$, whose complexity is $\mathcal{O}(N^2)$ for banded system matrices with tree-width $\mathcal{O}(\sqrt{N})$ arising in typical image analysis problems on 2-D grids. We can perform approximate MAP inference much faster using iterative techniques such as preconditioned conjugate gradients (Golub and Van Loan, 1996) or multigrid (Terzopoulos, 1988), whose complexity for many computer vision models is $\mathcal{O}(N^{3/2})$ or even $\mathcal{O}(N)$.

Standard algorithms for sampling from the Gaussian MRF also require a Cholesky factorization of $\boldsymbol{J}$ and thus have the same large time and memory complexity of direct system solvers. The following result though shows that we can draw *exact* GMRF samples by Perturb-and-MAP:

Proposition 1.1. *Assume that we replace the quadratic potential mean $\boldsymbol{\mu}_0$ by its perturbed version $\tilde{\boldsymbol{\mu}}_0 \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$, followed by finding the MAP of the perturbed model $\tilde{\boldsymbol{x}} = \boldsymbol{F}^T\boldsymbol{\Sigma}_0^{-1}\tilde{\boldsymbol{\mu}}_0$. Then $\tilde{\boldsymbol{x}}$ is an exact sample from the original GMRF $\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$.*

Proof. Since $\tilde{\boldsymbol{\mu}}_0$ is Gaussian, $\tilde{\boldsymbol{x}} = \boldsymbol{J}^{-1}\boldsymbol{F}^T\boldsymbol{\Sigma}_0^{-1}\tilde{\boldsymbol{\mu}}_0$ also follows a multivariate Gaussian distribution. It has mean $\mathbb{E}\{\tilde{\boldsymbol{x}}\} = \boldsymbol{\mu}$ and covariance matrix $\mathbb{E}\{(\tilde{\boldsymbol{x}}-\boldsymbol{\mu})(\tilde{\boldsymbol{x}}-\boldsymbol{\mu})^T\} = \boldsymbol{J}^{-1}\boldsymbol{F}^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{F}\boldsymbol{J}^{-1} = \boldsymbol{\Sigma}$. □

It is noteworthy that the algorithm only involves locally perturbing each

potential separately, $\tilde{\mu}_j \sim \mathcal{N}(\mu_j, \Sigma_j)$, and turns any existing GMRF MAP algorithm into an effective random sampler.

As an example, we show in Figure 1.6 an image inpainting application in which we fill in the flat areas of an image given the values at its edges under a 2-D thin-membrane prior GMRF model (Terzopoulos, 1988; Szeliski, 1990; Malioutov et al., 2008), which involves pairwise quadratic potentials $V_{ij} = \frac{1}{2\Sigma}(x_i - x_j)^2$ between nearest neighbors connected as in Figure 1.5(b). We show both the posterior mean/MAP estimate and a random sample under the model, both computed in a fraction of a second by solving a Poisson equation by a $\mathcal{O}(N)$ multigrid solver originally developed for solving PDE problems (Terzopoulos, 1988).

Figure 1.6: Reconstructing an image from its value on edges under a nearest-neighbor Gaussian MRF model. (a) Masked image. (b) Posterior mean/MAP estimate $\hat{\boldsymbol{x}}$. (c) Random sample $\tilde{\boldsymbol{x}}$.

1.3.2 Efficient MCMC Inference in Conditionally Gaussian Models

Gaussian models have proven inadequate for image modeling as they fail to capture important aspects of natural image statistics such as the heavy tails in marginal histograms of linear filter responses. Nevertheless, much richer statistical image tools can be built if we also incorporate into our models latent variables or allow nonlinear interactions between multiple Gaussian fields and thus the GMRF sampling technique we describe here is very useful within this wider setting (Weiss and Freeman, 2007; Roth and Black, 2009; Papandreou et al., 2008).

In (Papandreou and Yuille, 2010) we discuss the integration of our GMRF sampling algorithm in a block-Gibbs sampling context, where the conditionally Gaussian continuous variables and the conditionally independent latent variables are sampled alternately. The most straightforward way to capture the heavy tailed histograms of natural images is to model each filter response with a Gaussian mixture expert, thus using a single discrete assignment vari-

able at each factor (Papandreou et al., 2008; Schmidt et al., 2010). We show in Figure 1.7 an image inpainting example following this approach in which a wavelet domain hidden Markov tree model is used (Papandreou et al., 2008).

Figure 1.7: Filling in missing image parts from the ancient wall-paintings of Thera (Papandreou, 2009). Image inpainting with a wavelet domain model and block Gibbs sampling inference (Papandreou et al., 2008).

Efficient GMRF Perturb-and-MAP sampling can also be used in conjunction with Gaussian scale mixture (GSM) models for which the latent scale variable is continuous (Andrews and Mallows, 1974). We demonstrate this in the context of Bayesian signal restoration by sampling from the posterior distribution under a total variation (TV) prior, employing the GSM characterization of the Laplacian density. We show in Figure 1.8 an example of 1-D signal restoration under a TV signal model. The standard MAP estimator features characteristic staircasing artifacts (Nikolova, 2007). Block Gibbs sampling from the posterior distribution allows us to efficiently approximate the posterior mean estimator, which outperforms the MAP estimator in terms of mean square error/PSNR. Although individual posterior random samples are worse in terms of PSNR, they accurately capture the micro-texture of the original clean signal.

1.3.3 Variational Inference for Bayesian Compressed Sensing

Variational inference is increasingly popular for probabilistic inference in sparse models, providing the basis for many modern Bayesian compressed sensing methods. At a high level, variational techniques in this setting typically approximate the true posterior distribution with a parameterized Gaussian which allows closed-form computations. Inference amounts to adjusting the variational parameters to make the fit as tight as possible (Wainwright and Jordan, 2008). Mostly related to our work are (Attias, 1999;

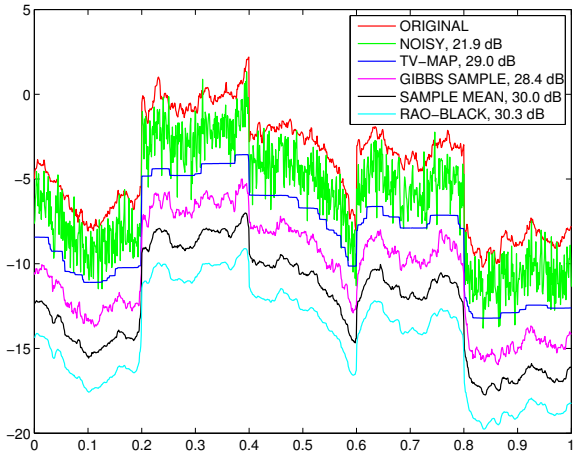


Figure 1.8: Signal denoising under a total variation prior model and alternative estimation criteria. From top to bottom, the graphs show: (a) Original latent clean signal, synthesized by adding Laplacian noise increments to a piece-wise constant signal. (b) Noisy version of the signal, corrupted by Gaussian i.i.d. noise. (c) MAP estimator under a TV prior model. (d) A single sample from the TV posterior Gibbs distribution. (e) Posterior mean estimator obtained by averaging multiple samples. (f) Rao-Blackwellized posterior mean estimator (Papandreou and Yuille, 2010).

Lewicki and Sejnowski, 2000; Girolami, 2001; Chantas et al., 2010; Seeger and Nickisch, 2011a). There exist multiple alternative criteria to quantify the fit quality, giving rise to approximations such as variational bounding (Jordan et al., 1999), mean field or ensemble learning, and, expectation propagation (EP) (Minka, 2001), as well as different iterative algorithms for optimizing each specific criterion. See (Bishop, 2006; Palmer et al., 2005) for further discussions about the relations among these variational approaches.

All variational algorithms we study in this chapter are of a double-loop nature, requiring Gaussian variance estimation in the outer loop and sparse point estimation in the inner loop (Seeger and Nickisch, 2011a; van Gerven et al., 2010; Seeger and Nickisch, 2011b). The ubiquity of the Gaussian variance computation routine is not coincidental. Variational approximations try to capture uncertainty in the intractable posterior distribution along the directions of sparsity. These are naturally encoded in the covariance matrix of the proxy Gaussian variational approximation. Marginal Gaussian variance computation is also required in automatic relevance determination algorithms for sparse Bayesian learning (MacKay, 1992) and relevance vector machine training (Tipping, 2001); the methods we review here could also be applied in that context.

It turns out that variance computation in large-scale Gaussian models is computationally challenging and a host of sophisticated techniques have

been developed for this purpose, which often only apply to restricted classes of models (Schneider and Willsky, 2001; Sudderth et al., 2004; Malioutov et al., 2008).

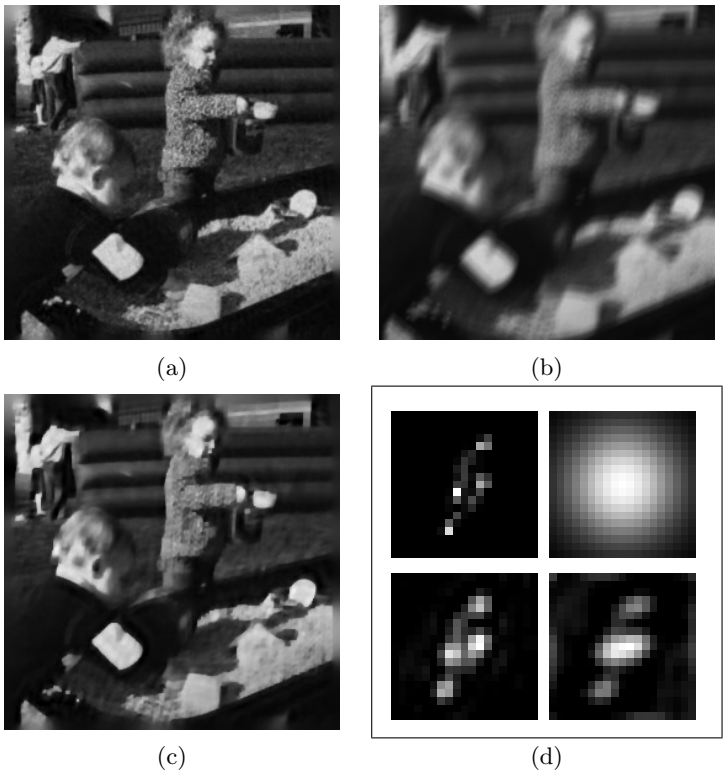


Figure 1.9: Blind image deblurring with variational inference. (a) Ground truth. (b) Blurred input image. (c) Estimated clean image. (d) Ground truth (top-left) and iteratively estimated blur kernel (clock-wise, starting from a diffuse Gaussian profile at top-right).

Perturb-and-MAP allows us to efficiently sample from the GMRF model and thus makes it practical to employ the generic sample-based estimator for computing Gaussian variances. More specifically, we repeatedly draw K independent GMRF samples $\{\tilde{\boldsymbol{x}}_k\}_{k=1}^K$ from which we can estimate the covariance matrix

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{K}\sum_{k=1}^{K}(\tilde{\boldsymbol{x}}_k - \boldsymbol{\mu})(\tilde{\boldsymbol{x}}_k - \boldsymbol{\mu})^T \tag{1.4}$$

This Monte-Carlo estimator, whose accuracy is independent of the problem size, is particularly attractive if only relatively rough variance estimates

suffice, as is often the case in practice. We show in Figure 1.9 an example of applying this variational Bayesian estimation methodology in the problem of blind image deblurring (Papandreou and Yuille, 2011b).

1.4 Perturb-and-MAP for MRFs with Discrete Labels

1.4.1 Introduction

We now turn our attention to Markov random fields on discrete labels, which go back to the classic Ising and Potts models in statistical physics. Discrete-valued MRFs offer a natural and sound probabilistic modeling framework for a host of image analysis and computer vision problems involving discrete labels, such as image segmentation and labeling, texture synthesis, and deep learning (Besag, 1974; Geman and Geman, 1984; Zhu et al., 1998; Hinton, 2002; Koller and Friedman, 2009). Exact probabilistic inference and maximum likelihood model parameter fitting is intractable in general MRFs defined on 2-D domains and one has to employ random sampling schemes to perform these tasks (Geman and Geman, 1984; Hinton, 2002).

Recent powerful discrete energy minimization algorithms such as graph cuts, linear programming relaxations, or loopy belief propagation (Boykov et al., 2001; Kolmogorov and Zabih, 2004; Kolmogorov and Rother, 2007; Koller and Friedman, 2009) can efficiently find or well approximate the most probable (MAP) configuration for certain important classes of MRFs. They have had a particularly big impact on computer vision; for a recent overview, see the volume edited by Blake et al. (2011).

Our work on the Perturb-and-MAP discrete random field model has been motivated by the exact Gaussian MRF sampling algorithm described in Section 1.3. While the underlying mathematics and methods are completely different in the discrete setup, we have shown in (Papandreou and Yuille, 2011a) that the intuition of local perturbations followed by global optimization can also lead to powerful sampling algorithms for discrete label MRFs. Subsequent work by other groups, summarized in 1.5, has extended our results and explored related directions.

A surprising finding of our study has been the identification of a perturbation process from extreme value statistics which turns the Perturb-and-MAP model identical to its Gibbs counterpart even in the discrete setting. Although this perturbation is too expensive to be applicable in large-scale models, it nevertheless suggests low-order perturbations that result in perturbed energies that are effectively as easy to minimize as the original unperturbed one, while producing high-quality random samples.

Perturb-and-MAP endows discrete energy minimization algorithms such as graph cuts with probabilistic capabilities that allow them to support qualitatively new computer vision applications. We illustrate some of them in image segmentation and scene labeling experiments experiments: First, drawing several posterior samples from the model allows us to compute posterior marginal probabilities and quantify our confidence in the MAP solution. Second, efficient random sampling allows learning of MRF or CRF parameters using the moment matching rule, in which the model parameters are updated until the generated samples reproduce the (weighted) sufficient statistics of the observed data.

1.4.2 Model Definition and Weight Space Geometry

We assume a deterministic energy function which takes the inner product form of Equation (1.1), i.e., $E(\boldsymbol{x};\boldsymbol{\theta}) = \langle \boldsymbol{\theta}, \boldsymbol{\phi}(\boldsymbol{x})\rangle$, with x_i taking values in a discrete label set $\mathcal{L}$. A Perturb-and-MAP random sample is obtained by $\tilde{\boldsymbol{x}} = \operatorname{argmin}_{\boldsymbol{x}} E(\boldsymbol{x};\boldsymbol{\theta}+\boldsymbol{\epsilon})$, where $\boldsymbol{\epsilon}$ is a real-valued random additive parameter perturbation vector. By construction, we can efficiently draw exact one-shot samples from the Perturb-and-MAP model by solving an energy minimization problem.

Thanks to the inner product form of the energy function, the Perturb-and-MAP model has a simple geometric interpretation in the parameter space. In particular, a state $\boldsymbol{x} \in \mathcal{L}^N$ will be minimizing the deterministic energy if, and only if, $E(\boldsymbol{x};\boldsymbol{\theta}) \leq E(\boldsymbol{q};\boldsymbol{\theta}), \forall \boldsymbol{q} \in \mathcal{L}^N$. This set of $|\mathcal{L}|^N$ linear inequalities defines a polyhedron $\mathcal{P}_{\boldsymbol{x}}$ in the weight space

$$\mathcal{P}_{\boldsymbol{x}} = \{\boldsymbol{\theta} \in \mathbb{R}^M : \langle \boldsymbol{\theta}, \boldsymbol{\phi}(\boldsymbol{x}) - \boldsymbol{\phi}(\boldsymbol{q})\rangle \leq 0, \forall \boldsymbol{q} \in \mathcal{L}^N\}\,. \tag{1.5}$$

Actually, $\mathcal{P}_{\boldsymbol{x}}$ is a polyhedral cone (Boyd and Vandenberghe, 2004), since $\boldsymbol{\theta} \in \mathcal{P}_{\boldsymbol{x}}$ implies $\alpha\boldsymbol{\theta} \in \mathcal{P}_{\boldsymbol{x}}$, for all $\alpha \geq 0$. These polyhedral cones are dually related to the marginal polytope $\mathcal{M} = \operatorname{conv}(\{\boldsymbol{\phi}(\boldsymbol{x})\}, \boldsymbol{x} \in \mathcal{L}^N)$, as illustrated in Figure 1.10; see (Wainwright and Jordan, 2008) for background on the marginal polytope. The polyhedra $\mathcal{P}_{\boldsymbol{x}}$ partition the weight space $\mathbb{R}^M$ into regions of influence of each discrete state $\boldsymbol{x} \in \mathcal{L}^N$. Under the Perturb-and-MAP model, $\boldsymbol{x}$ will be assigned to a particular state $\boldsymbol{x}$ if, and only if, $\boldsymbol{\theta} + \boldsymbol{\epsilon} \in \mathcal{P}_{\boldsymbol{x}}$ or, equivalently, $\boldsymbol{\epsilon} \in \mathcal{P}_{\boldsymbol{x}} - \boldsymbol{\theta} \triangleq \{\boldsymbol{\epsilon} \in \mathbb{R}^M : \boldsymbol{\theta} + \boldsymbol{\epsilon} \in \mathcal{P}_{\boldsymbol{x}}\}$. In other words, if a specific instantiation of the perturbation $\boldsymbol{\epsilon}$ falls in the shifted polyhedron $\mathcal{P}_{\boldsymbol{x}} - \boldsymbol{\theta}$, then the Perturb-and-MAP model generates $\boldsymbol{x}$ as sample.

We assume that perturbations are drawn from a density $f_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon})$ which does not depend on the parameters $\boldsymbol{\theta}$. The probability mass of a state $\boldsymbol{x}$ under the Perturb-and-MAP model is then the weighted volume of the corresponding

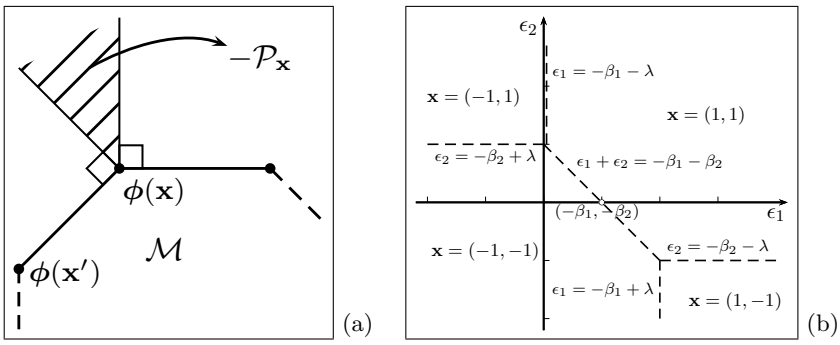


Figure 1.10: Perturb-and-MAP geometry. (a) The polyhedral cones $\mathcal{P}_{\boldsymbol{x}}$ are dual to the corner cones of the marginal polytope $\mathcal{M}$. (b) The Ising P-M model with $N = 2$ nodes and perturbations only in the unary terms, $\tilde{\beta}_i = \beta_i + \epsilon_i$, for parameter values $\beta_1 = -1$, $\beta_2 = 0$, and $\lambda = 1$. The $\boldsymbol{\epsilon}$-space is split into four polyhedra, with $\boldsymbol{x}(\boldsymbol{\epsilon}) = \boldsymbol{x}$ iff $\boldsymbol{\epsilon} \in \mathcal{P}_{\boldsymbol{x}} - \boldsymbol{\theta}$.

shifted polyhedron under the perturbation measure

$$f_{PM}(\boldsymbol{x}; \boldsymbol{\theta}) = \int_{\mathcal{P}_{\boldsymbol{x}} - \boldsymbol{\theta}} f_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon}) d\boldsymbol{\epsilon}\,, \tag{1.6}$$

which is the counterpart of the Gibbs density in Equation (1.2). It is intractable (NP-hard) to compute the volume of general polyhedra in a high-dimensional space; see, e.g., (Ben-Tal et al., 2009, p. 29). However, for the class of perturbed energy functions which can be globally minimized efficiently, we can readily draw exact samples from the Perturb-and-MAP model, without ever explicitly evaluating the integrals in Equation (1.6).

1.4.3 Example: The Perturb-and-MAP Ising model

Let us illustrate these ideas by considering the Perturb-and-MAP version of the classic Ising model. The Ising energy over the discrete "spins" $x_i \in \{-1, 1\}$ is defined as

$$E(\boldsymbol{x}; \boldsymbol{\theta}) = \frac{-1}{2} \sum_{i=1}^{N} \left(\beta_i x_i + \sum_{i'=i+1}^{N} \lambda_{ii'} x_i x_{i'} \right), \tag{1.7}$$

where β_i is the external field strength ($\beta_i > 0$ favors $x_i = 1$) and $\lambda_{ii'}$ is the coupling strength, with attractive coupling $\lambda_{ii'} > 0$ favoring the same spin for x_i and $x_{i'}$. This energy function can be written in the standard inner product form of Equation (1.1) with $\boldsymbol{\theta} = (\{\beta_i\}, \{\lambda_{ii'}\})^T$ and $\boldsymbol{\phi}(\boldsymbol{x}) = \frac{-1}{2}(\{x_i\}, \{x_i x_{i'}\})^T$. The MRF defined by Equation (1.2) is the Ising Gibbs random field.

Defining a Perturb-and-MAP Ising random field requires specifying the

parameter perturbation density. In this example, we leave the binary term parameters $\lambda_{ii'}$ intact and only perturb the unary term parameters β_i. In particular, for each unary factor, we set $\tilde{\beta}_i = \beta_i + \epsilon_i$, with ϵ_i i.i.d. samples from the logistic distribution with density $l(z) = \frac{1}{4}\,\mathrm{sech}^2(\frac{z}{2})$. This corresponds to the order-1 Gumbel perturbation we discuss in Section 1.4.5 and ensures that if a particular node x_i is completely isolated, it will then follow the same Bernoulli distribution $\Pr\{x_i = 1\} = 1/(1 + e^{-\beta_i})$ as in the Gibbs case. The $\boldsymbol{\epsilon}$-space geometry in the case of two labels ($N = 2$) under the Ising energy $E(\boldsymbol{x}; \boldsymbol{\theta}) = -0.5(\beta_1 x_1 + \beta_2 x_2 + \lambda x_1 x_2)$ for a specific value of the parameters $\boldsymbol{\theta}$ and perturbations only to unary terms is depicted in Figure 1.10.

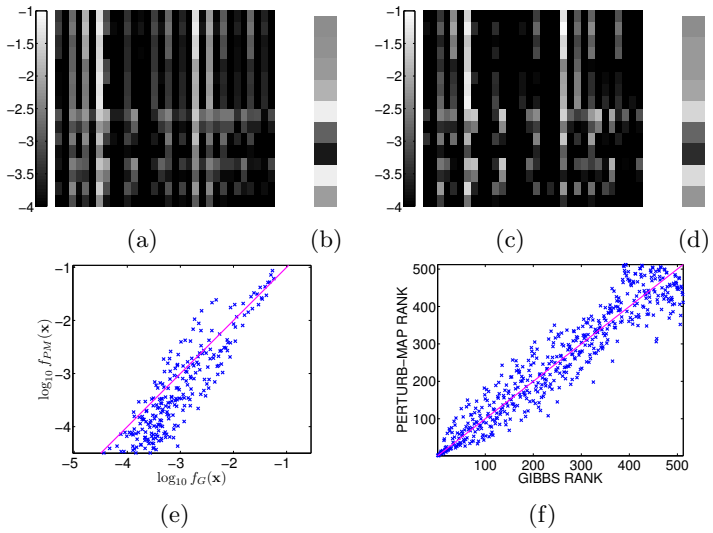


Figure 1.11: We compare the Gibbs (exact computation) and the Perturb-and-MAP (10^6 Monte-Carlo runs) models induced from an Ising energy on 3×3 grid, with β_i and $\lambda_{ii'}$ i.i.d. from $\mathcal{N}(0, 1)$. (a) Gibbs log-probabilities $\log_{10} f_G(\boldsymbol{x})$ for each of the 2^9 states, arranged as a $2^5 \times 2^4$ matrix. (b) Gibbs marginal probabilities $f_G(x_i = 1)$ for each of the 9 nodes. (c) Perturb-and-MAP log-probabilities $\log_{10} f_{PM}(\boldsymbol{x})$. (d) Perturb-and-MAP marginal probabilities $f_{PM}(x_i = 1)$. (e) Scatter-plot of state log probabilities under the two models. (f) Scatter-plot of states ranked by their probabilities under the two models.

We compare in Figure 1.11 the Gibbs and Perturb-and-MAP models for a small-scale Ising energy involving 9 variables on a 3×3 grid with 4-nearest neighbors connectivity and randomly generated parameters. The probability landscape (i.e., the probabilities of each of the 2^9 states) looks quite similar under the two models, see Figure 1.11 (a) and (c). The same holds for the corresponding marginal probabilities, shown in Figure 1.11 (b) and (d). To

further compare the probability landscape under the two models, we show a scatter plot of their log probabilities in Figure 1.11(e), as well as a scatter plot of the states ranked by their probability in Figure 1.11(f). Perturb-and-MAP in this example is particularly close to Gibbs for the leading (most probable) states but tends to under-estimate the least probable states.

1.4.4 Parameter Estimation by Moment Matching

We would like to estimate the parameters $\boldsymbol{\theta}$ of the Perturb-and-MAP model from a labeled training set $\{\boldsymbol{x}_k\}_{k=1}^K$ by maximizing the log-likelihood

$$L_{PM}(\boldsymbol{\theta}) = (1/K) \sum_{k=1}^{K} \log f_{PM}(\boldsymbol{x}_k; \boldsymbol{\theta}) \,. \tag{1.8}$$

We can design the perturbations so as the Perturb-and-MAP log-likelihood L_{PM} is a concave function of $\boldsymbol{\theta}$. This ensures that the likelihood landscape is well-behaved and allows the use of local search techniques for parameter estimation, exactly as in the Gibbs case. Specifically, the following result is shown in (Papandreou and Yuille, 2011a):

Proposition 1.2. *If the perturbations $\boldsymbol{\epsilon}$ are drawn from a log-concave density $f_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon})$, the log-likelihood $L_{PM}(\boldsymbol{\theta})$ is a concave function of the energy parameters $\boldsymbol{\theta}$.*

The family of log-concave distributions (Boyd and Vandenberghe, 2004), i.e., $\log f_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon})$ is a concave function of $\boldsymbol{\epsilon}$, includes the Gaussian, the logistic, and other commonly used distributions.

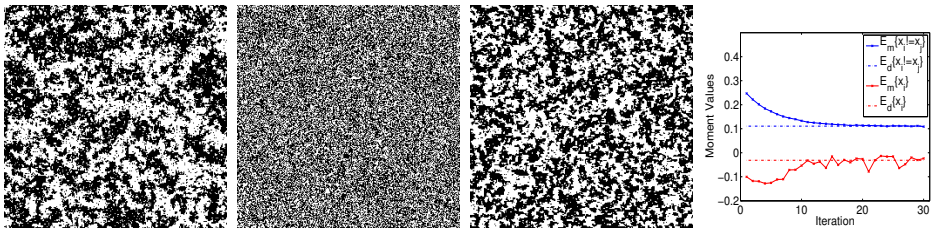
The gradient of $L_{PM}(\boldsymbol{\theta})$ is in general hard to compute. Motivated by the parameter update formula in the Gibbs case from Section 1.2.2, we opt for the moment matching learning rule, $\theta_j(t+1) = \theta_j(t) + r(t)\Delta\theta_j$, where

$$\Delta\theta_j = \mathbb{E}_{\boldsymbol{\theta}}^{PM}\{\phi_j(\boldsymbol{x})\} - \mathbb{E}_D\{\phi_j(\boldsymbol{x})\} \,. \tag{1.9}$$

Here $\mathbb{E}_{\boldsymbol{\theta}}^{PM}\{\phi_j(\boldsymbol{x})\} \triangleq \sum_{\boldsymbol{x}} f_{PM}(\boldsymbol{x}; \boldsymbol{\theta})\phi_j(\boldsymbol{x})$ is the expected sufficient statistic under the Perturb-and-MAP model for the current parameter values $\boldsymbol{\theta}$, which we can efficiently estimate by drawing exact samples from it. We typically adjust the learning rate by a Robbins-Monro type schedule, e.g., $r(t) = r_1/(r_2 + t)$. Figure 1.12 illustrates parameter learning by moment matching in a spatially homogeneous Ising energy model.

While the above moment matching rule was originally motivated by analogy to the Gibbs case (Papandreou and Yuille, 2011a), its fixed points do not need to be exact minima of the Perturb-and-MAP log-likelihood (1.8). Subsequent work has shown that moment matching performs gradient ascent

for an objective function that lower bounds the Gibbs likelihood function (Hazan and Jaakkola, 2012). Moreover, this lower bound turns out to be concave even for perturbation densities $f_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon})$ which are not log-concave.



(a) (b) (c) (d)

Figure 1.12: Perturb-and-MAP Ising random field parameter learning. The two model parameters, the global coupling strength λ and field strength β are fitted by moment matching. (a) Gibbs Ising model sample, used as training image. (b) Perturb-and-MAP Ising sample at initial parameter values. (c) Perturb-and-MAP Ising sample at final parameter values. (d) Model moments as they converge to training data moments.

1.4.5 Perturb-and-MAP Perturbation Design

Although any perturbation density induces a legitimate Perturb-and-MAP model, it is desirable to carefully design it so as the Perturb-and-MAP model approximates as closely as possible the corresponding Gibbs MRF. The Gibbs MRF has important structural properties that are not automatically satisfied by the Perturb-and-MAP model under arbitrary perturbations: (a) Unlike the Gibbs MRF, the Perturb-and-MAP model is not guaranteed to respect the state ranking induced by the energy, i.e., $E(\boldsymbol{x}) \leq E(\boldsymbol{x}')$ does not necessarily imply $f_{PM}(\boldsymbol{x}) \geq f_{PM}(\boldsymbol{x}')$, see Figure 1.11(f). (b) The Markov dependence structure of the Gibbs MRF follows directly from the support of the potentials $\phi_j(\boldsymbol{x})$, while the Perturb-and-MAP might give rise to longer-range probabilistic dependencies. (c) The maximum entropy distribution under moment constraints $\mathbb{E}\{\phi_j(\boldsymbol{x})\} = \bar{\phi}_j$ has the Gibbs form; the Perturb-and-MAP model trained by moment matching can reproduce these moments but will in general have smaller entropy than its Gibbs counterpart.

The *Gumbel* distribution arising in extreme value theory (Steutel and Van Harn, 2004) turns out to play an important role in our effort to design a perturbation mechanism that yields a Perturb-and-MAP model closely resembling the Gibbs MRF. It is a continuous univariate distribution with log-concave density $g(z) = \exp(-(-z + e^z))$. We can efficiently draw independent Gumbel variates by transforming standard uniform samples by

$u \to \log(-\log(u))$. The Gumbel density naturally fits into the Perturb-and-MAP model, thanks to the following key Lemma – also see (Kuzmin and Warmuth, 2005):

Lemma 1.3. *Let $(\theta_1, \ldots, \theta_m)$, with $\theta_n \in \mathbb{R}$, $n = 1, \ldots, m$. We additively perturb them by $\tilde{\theta_n} = \theta_n + \epsilon_n$, with ϵ_n i.i.d. zero-mode Gumbel samples. Then:*

(a) The minimum of the perturbed parameters $\tilde{\theta}_{min} \triangleq \min_{n=1:m}\{\tilde{\theta}_n\}$ follows a Gumbel distribution with mode θ_0, where $e^{-\theta_0} = \sum_{n=1}^{m} e^{-\theta_n}$.

(b) The probability that $\tilde{\theta}_n$ is the minimum value is $\Pr\{\text{argmin}(\tilde{\theta}_1, \ldots, \tilde{\theta}_m) = n\} = e^{-\theta_n}/e^{-\theta_0}$.

Note that $\theta_0 = -\log(\sum_{n=1}^{m} e^{-\theta_n}) = -\log Z$. This connection is pursued in detail by Hazan and Jaakkola (2012), which develops a Perturb-and-MAP based approximation to the partition function.

We can use this Lemma to construct a Perturb-and-MAP model that exactly replicates the Gibbs distribution, as follows. The Gibbs random field on N sites x_i, $i = 1, \ldots, N$, each allowed to take a value from the discrete label set $\mathcal{L}$, can be considered as a discrete distribution with $|\mathcal{L}|^N$ states. This can be made explicit if we enumerate $\{\boldsymbol{x}_j, j = 1, \ldots, \bar{M} = |\mathcal{L}|^N\}$ all the states and consider the maximal equivalent re-parameterization of Equation (1.1)

$$\bar{E}(\boldsymbol{x}; \bar{\boldsymbol{\theta}}) \triangleq \langle \bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\phi}}(\boldsymbol{x}) \rangle = \langle \boldsymbol{\theta}, \boldsymbol{\phi}(\boldsymbol{x}) \rangle \,, \tag{1.10}$$

where $\bar{\theta}_j = E(\boldsymbol{x}_j; \boldsymbol{\theta}) = \langle \boldsymbol{\theta}, \boldsymbol{\phi}(\boldsymbol{x}_j) \rangle$, $j = 1, \ldots, \bar{M}$, is the *fully-expanded* potential table and $\bar{\phi}_j(\boldsymbol{x})$ is the indicator function of the state $\boldsymbol{x}_j$ (i.e., equals 1, if $\boldsymbol{x} = \boldsymbol{x}_j$ and 0 otherwise). Using Lemma 1.3 we can show:

Proposition 1.4. *If we perturb each entry of the fully expanded $\mathcal{L}^N$ potential table with i.i.d. Gumbel noise samples $\epsilon_j, j = 1, \ldots, \bar{M}$, then the Perturb-and-MAP and Gibbs models coincide, i.e., $f_{PM}(\boldsymbol{x}; \boldsymbol{\theta}) = f_G(\boldsymbol{x}; \boldsymbol{\theta})$.*

This order-N perturbation is not practically applicable when N is large since it independently perturbs all $\bar{M} = |\mathcal{L}|^N$ entries of the fully expanded potential table and effectively destroys the local Markov structure of the energy function, rendering it too hard to minimize. Nevertheless, it shows that it is possible to design a Perturb-and-MAP model that exactly replicates the Gibbs MRF.

In practice, we employ low-order Gumbel perturbations. In our simplest order-1 design, we only add Gumbel noise to the unary potential tables. More specifically, for an energy function $E(\boldsymbol{x}) = \sum_{i=1}^{N} V_i(x_i) + \sum_j V_j(\boldsymbol{x}_j)$

which includes potentials $V_i(x_i)$ of order-1 and potentials $V_j(\boldsymbol{x}_j)$ of order-2 or higher, we add i.i.d. Gumbel noise to each of the $|\mathcal{L}|$ entries of each order-1 potential, while leaving the higher order potentials intact. This yields perturbed energies effectively as easy to minimize as the original unperturbed one, while producing random samples closely resembling Gibbs MRF samples. We can improve the Perturb-and-MAP sample quality by Gumbel perturbations of order-2 or higher, as described in (Papandreou and Yuille, 2011a). However, high order perturbations typically make the perturbed energy minimization problem harder to solve.

1.4.6 Applications and Experiments

We present experiments with the Perturb-and-MAP model applied to image segmentation and scene labeling.

Our *interactive image segmentation* experiments have been performed on the Grabcut dataset which includes human annotated ground truth segmentations (Rother et al., 2004). The task is to segment a foreground object, given a relatively tight tri-map imitating user input obtained by a lasso or pen tool.

In our implementation we closely follow the CRF formulation of (Rother et al., 2011), using the same parameters for defining the image-based CRF terms and considering pixel interactions in a 8-neighborhood. We used our Perturb-and-MAP sampling algorithm with order-2 Gumbel perturbation and QPBO optimization (Kolmogorov and Rother, 2007) to learn the weights of the potentials – 5 weights in total, one for the unary and one for each of the 4 pairwise connections of the center pixel with its S, E, NE, SE neighbors. Using these parameters, we obtained a classification error rate of 5.6% with the global MAP decision rule. This is similar to the best results attainable with the particular CRF model and hand-tuned weights.

In Figure 1.13 we illustrate the ability of the Perturb-and-MAP model to produce soft segmentation maps. The soft segmentation map (average over 20 posterior samples) gives a qualitatively accurate estimate of the segmentation uncertainty, which could potentially be useful in guiding user interaction in an interactive segmentation application.

We next consider an application of Perturb-and-MAP random fields in *scene layout labeling* (Hoiem et al., 2007). We use the tiered layout model of (Felzenszwalb and Veksler, 2010), which allows exact global inference by efficient dynamic programming (Felzenszwalb and Veksler, 2010). The model has a relatively large number of parameters, making it difficult to hand tune. Training them with the proposed techniques illustrates our ability to effectively learn model parameters from labeled data.

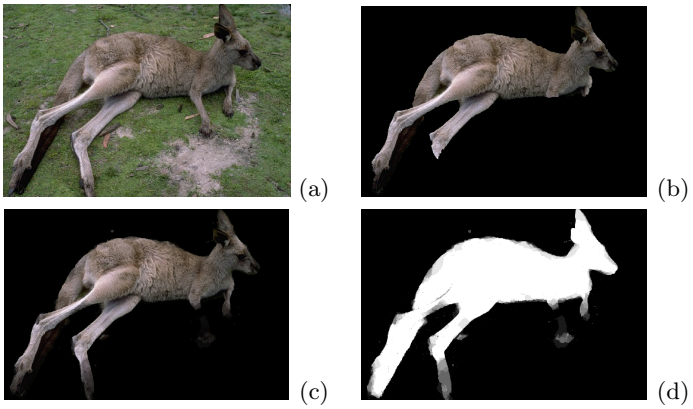


Figure 1.13: Interactive image segmentation results on the Grabcut dataset. Parameters learned by Perturb-and-MAP moment matching. (a) Original image. (b) Least energy MAP solution. (c) Soft Perturb-and-MAP segmentation. (d) The corresponding segmentation mask.

We closely follow the evaluation approach of (Felzenszwalb and Veksler, 2010) in setting up the experiment: We use the dataset of 300 outdoor images (and the standard cross-validation splits into training/test sets) with ground truth from (Hoiem et al., 2007). Similarly to (Felzenszwalb and Veksler, 2010), we use five labels: T (sky), B (ground), and three labels for the middle region, L (facing left), R (facing right), C (front facing), while we exclude the classes "porous" and "solid". The unary scores are produced using classifiers that we trained using the dataset and software provided by Hoiem et al. (2007) following the standard five-fold cross-validation protocol.

We first fit the tiered scene model parameters (pairwise compatibility tables between the different classes) on the training data using Perturb-and-MAP moment matching (order-1 Gumbel perturbation). Weights are initialized as Potts CRF potentials and refined by moment matching rule; we separated the training set in batches of 10 images each and stopped after 50 epochs over the training set. We have measured the performance of the trained model in terms of average accuracy on the test set. We have tried two decision criteria, MAP (least energy configuration) and marginal MODE (i.e., assign each pixel to the label that appears most frequently in 20 random Perturb-And-Map conditional samples from the model), obtaining accuracy 82.7% and 82.6%, respectively. Our results are better than the unary-only baseline mean accuracy of 82.1% (Hoiem et al., 2007), and the MAP and MODE results of 82.1% and 81.8%, respectively, that we obtained with the hand-set weights of (Felzenszwalb and Veksler, 2010).

In Figure 1.14 we show some indicative examples of different scene layout labelings obtained by the unary-only, the tiered MAP, and the Perturb-and-

MAP model. The uncertainty of the solution is indicated by entropy maps. The marginal mode and entropies shown are Monte Carlo estimates using 20 Perturb-and-MAP samples.

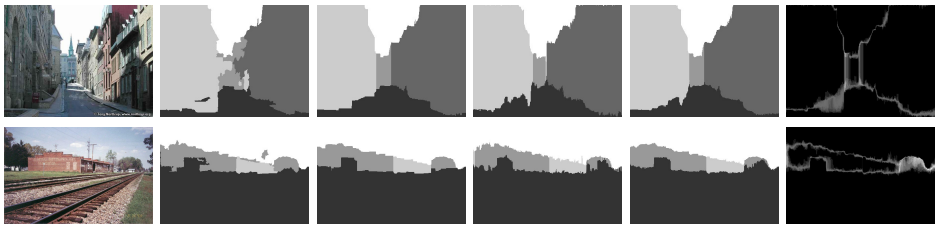


Figure 1.14: Tiered scene labeling results with pairwise potentials learned by our Perturb-and-MAP moment matching algorithm. Left to right: image; unary-only MAP; tiered MAP; one tiered Perturb-and-MAP sample; tiered Perturb-and-MAP marginal mode; tiered Perturb-and-MAP marginal entropy.

1.5 Related Work and Recent Developments

To our knowledge, adding noise to the weighted edges of a graph so as to randomize the minimum energy configuration found by mincuts was first proposed by Blum et al. (2004) in the context of a submodular binary MRF energy arising in semi-supervised learning. Their goal was to break graph symmetries and allow the standard mincut algorithm to produce a different solution at each run. They interpret the relative frequency of each node receiving one or the other label as a confidence score for binary classification. However, beyond randomizing the deterministic mincut algorithm, they do not study the implied probabilistic model as a standalone object nor attempt to design the perturbation mechanism so as to approximate the corresponding Gibbs model. Indeed, the choice of perturbation distribution is not discussed at all in (Blum et al., 2004).

Herding (Welling, 2009) builds a deterministic dynamical system on the model parameters designed so as to reproduce the data sufficient statistics, which is similar in spirit to the moment-matching algorithm we use for learning. However, herding is still not a probabilistic model and cannot summarize the data into a concise set of model parameters.

As pointed out to us by McAllester (2012), Perturb-and-MAP is closely related to PAC-Bayes (McAllester, 1998) and PAC-Bayesian theorems such as those in (Germain et al., 2009) can be adapted to the Perturb-and-MAP setting. Model perturbations through the associated concept of stochastic

Gibbs classifier play a key role to PAC-Bayesian theory, but PAC-Bayes typically aims at producing generalization guarantees for the deterministic classifier instead of capturing the uncertainty in the posterior distribution.

Averaging over multiple samples, Perturb-and-MAP allows efficiently estimating (sum-) marginal densities and thus quantifying the per-node solution uncertainty even in graphs with loops. Max-product belief propagation (Wainwright et al., 2005) and dynamic graph-cuts (Kohli and Torr, 2008) can compute max-marginals, which give some indication of the uncertainty in label assignments (Kohli and Torr, 2008) but cannot directly estimate marginal densities.

A number of different groups have followed up on our work (Papandreou and Yuille, 2011a) and further developed it in different directions. In their randomized optimum models, Tarlow et al. (2012) introduce variants of the Perturb-and-MAP model for discrete problems such as bi-partite matching and pursue maximum-likelihood learning of the model parameters using efficient MCMC algorithms.

The work in (Hazan and Jaakkola, 2012) has offered a better understanding of the Perturb-and-MAP moment matching learning rule, showing that it optimizes a well-defined concave lower bound of the Gibbs likelihood function. Moreover, they have shown how Perturb-and-MAP can be used for computing approximations to the partition function. This connection relates Perturb-and-MAP more directly to the standard MRF inference problem.

Another related partition function estimation algorithm is proposed in (Ermon et al., 2013). Interestingly, their method amounts to progressively introducing more random constraints, followed by energy minimization, in a randomized Constrain-and-MAP scheme.

While probabilistic random sampling allows one to explore alternative plausible solutions, Batra et al. (2012) propose to explicitly enforce diversity in generating a sequence of deterministic solutions.

The work in (Roig et al., 2013) is an excellent demonstration of how uncertainty quantification can yield practical benefits in a semantic image labeling setting. They employ Perturb-and-MAP to identify on the fly image areas with ambiguous labeling and only compute expensive features when their addition is likely to considerably decrease labeling entropy.

1.6 Discussion

This chapter has presented an overview of the Perturb-and-MAP method, which turns established deterministic energy minimization algorithms into efficient probabilistic inference machines. This is a promising new direction

with many important open questions for both theoretical and application-driven research: (1) An in-depth systematic comparison of Perturb-and-MAP and more established approximate inference techniques such as MCMC or Variational Bayes is still lacking. (2) So far, there is no clear characterization of the approximation quality of the Perturb-and-MAP model relative to its Gibbs counterpart and how perturbation design affects it. (3) Unlike MCMC which allows trading off approximation quality with computation time by simply running the Markov chain for longer, there is currently no way to iteratively improve the quality of Perturb-and-MAP samples. (4) The modeling capacity of Perturb-and-MAP needs to be explored in several more computer vision and machine learning applications.

For further information and links to related works in this exciting emerging area we point the reader to the NIPS Workshop on Perturbations, Optimization, and Statistics, organized in 2012 and 2013 by T. Hazan, D. Tarlow, A. Rakhlin, and the first author of this chapter.

Acknowledgements

Our work has been supported by the U.S. Office of Naval Research under MURI grant N000141010933; the NSF under award 0917141; the AFOSR under grant 9550-08-1-0489; and the Korean Ministry of Education, Science, and Technology, under the Korean National Research Foundation WCU program R31-10008. We would like to thank M. Welling, M. Seeger, T. Hazan, D. Tarlow, D. McAllester, A. Montanari, S. Roth, I. Kokkinos, M. Raptis, M. Ranzato, and C. Lampert for their feedback at various stages of this work.

1.7 References

D. Andrews and C. Mallows. Scale mixtures of normal distributions. *J. of Royal Stat. Soc. (Series B)*, 36(1):99–102, 1974.

H. Attias. Independent factor analysis. *Neural Computation*, 11:803–851, 1999.

D. Batra, P. Yadollahpour, A. Guzman-Rivera, and G. Shakhnarovich. Diverse m-best solutions in Markov random fields. In *Proc. European Conf. on Computer Vision*, 2012.

A. Ben-Tal, L. El Ghaoui, and A. Nemirovski. *Robust Optimization*. Princeton Univ. Press, 2009.

J. Besag. Spatial interaction and the statistical analysis of lattice systems. *J. of Royal Stat. Soc. (Series B)*, 36(2):192–236, 1974.

J. Besag. Statistical analysis of non-lattice data. *J. of Royal Stat. Soc. Series D*

(The Statistician), 24(3):179–195, 1975.

C. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

A. Blake, P. Kohli, and C. Rother, editors. *Markov Random Fields for Vision and Image Processing*. MIT Press, 2011.

A. Blum, J. Lafferty, M. Rwebangira, and R. Reddy. Semi-supervised learning using randomized mincuts. In *Proc. Int. Conf. on Machine Learning*, 2004.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge Univ. Press, 2004.

Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(11):1222–1239, 2001.

G. Chantas, N. Galatsanos, R. Molina, and A. Katsaggelos. Variational Bayesian image restoration with a product of spatially weighted total variation image priors. *IEEE Trans. Image Process.*, 19(2):351–362, 2010.

S. Ermon, C. Gomes, A. Sabharwal, and B. Selman. Taming the curse of dimensionality: Discrete integration by hashing and optimization. In *Proc. Int. Conf. on Machine Learning*, 2013.

P. Felzenszwalb and O. Veksler. Tiered scene labeling with dynamic programming. In *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition*, 2010.

S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 6(6): 721–741, 1984.

P. Germain, A. Lacasse, F. Laviolette, and M. Marchand. PAC-Bayesian learning of linear classifiers. In *Proc. Int. Conf. on Machine Learning*, 2009.

M. Girolami. A variational method for learning sparse and overcomplete representations. *Neural Computation*, 13:2517–2532, 2001.

G. Golub and C. Van Loan. *Matrix Computations*. John Hopkins Press, 1996.

T. Hazan and T. Jaakkola. On the partition function and random maximum a-posteriori perturbations. In *Proc. Int. Conf. on Machine Learning*, 2012.

G. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002.

G. Hinton and T. Sejnowski. Optimal perceptual inference. In *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition*, 1983.

D. Hoiem, A. Efros, and M. Hebert. Recovering surface layout from an image. *Int. J. of Comp. Vis.*, 75(1):151–172, 2007.

M. Jordan, J. Ghahramani, T. Jaakkola, and L. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37:183–233, 1999.

P. Kohli and P. Torr. Measuring uncertainty in graph cut solutions. *Computer Vision and Image Understanding*, 112(1):30–38, 2008.

D. Koller and N. Friedman. *Probabilistic Graphical Models*. MIT Press, 2009.

V. Kolmogorov and C. Rother. Minimizing non-submodular functions with graph cuts – a review. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(7):1274–1279, 2007.

V. Kolmogorov and R. Zabih. What energy functions can be minimized via graph cuts? *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(2):147–159, 2004.

D. Kuzmin and M. K. Warmuth. Optimum follow the leader algorithm. In *Proc. Conf. on Learning Theory*, 2005.

Y. LeCun, S. Chopra, R. Hadsell, M. Ranzato, and F.-J. Huang. A tutorial on energy-based learning. In G. Bakir, T. Hofmann, B. Schölkopf, A. Smola,

B. Taskar, and S. Vishwanathan, editors, *Predicting Structured Data*. MIT Press, 2007.

M. Lewicki and T. Sejnowski. Learning overcomplete representations. *Neural Computation*, 12:337–365, 2000.

D. MacKay. Bayesian interpolation. *Neural Computation*, 4(3):415–447, 1992.

D. Malioutov, J. Johnson, M. Choi, and A. Willsky. Low-rank variance approximation in GMRF models: Single and multiscale approaches. *IEEE Trans. Signal Process.*, 56(10):4621–4634, 2008.

D. McAllester. Some PAC-Bayesian theorems. In *Proc. Conf. on Learning Theory*, 1998.

D. McAllester. Connections between Perturb-and-MAP and PAC-Bayes. Personal communication, 2012.

T. Minka. Expectation propagation for approximate Bayesian inference. In *Proc. Int. Conf. on Uncertainty in Artificial Intelligence*, 2001.

M. Nikolova. Model distortions in Bayesian MAP reconstruction. *Inv. Pr. and Imag.*, 1(2):399–422, 2007.

J. Palmer, D. Wipf, K. Kreutz-Delgado, and B. Rao. Variational EM algorithms for non-Gaussian latent variable models. In *Proc. Advances in Neural Information Processing Systems*, 2005.

G. Papandreou. *Image Analysis and Computer Vision: Theory and Applications in the Restoration of Ancient Wall Paintings*. PhD thesis, NTUA, School of ECE, 2009.

G. Papandreou and A. Yuille. Gaussian sampling by local perturbations. In *Proc. Advances in Neural Information Processing Systems*, 2010.

G. Papandreou and A. Yuille. Perturb-and-MAP random fields: Using discrete optimization to learn and sample from energy models. In *Proc. IEEE Int. Conf. on Computer Vision*, 2011a.

G. Papandreou and A. Yuille. Efficient variational inference in large-scale Bayesian compressed sensing. In *Proc. IEEE Workshop on Information Theory in Computer Vision and Pattern Recognition (in conjunction with ICCV)*, 2011b.

G. Papandreou, P. Maragos, and A. Kokaram. Image inpainting with a wavelet domain hidden Markov tree model. In *Proc. IEEE Int. Conf. Acous., Speech, and Signal Processing*, 2008.

G. Roig, X. Boix, S. Ramos, R. de Nijs, and L. Van Gool. Active MAP inference in CRFs for efficient semantic segmentation. In *Proc. IEEE Int. Conf. on Computer Vision*, 2013.

S. Roth and M. Black. Fields of experts. *Int. J. of Comp. Vis.*, 82(2):205–229, 2009.

C. Rother, V. Kolmogorov, and A. Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. In *Proc. ACM Int. Conference on Computer Graphics and Interactive Techniques*, pages 309–314, 2004.

C. Rother, V. Kolmogorov, Y. Boykov, and A. Blake. Interactive foreground extraction using graph cut. In *Advances in Markov Random Fields for Vision and Image Processing*. MIT Press, 2011.

H. Rue and L. Held. *Gaussian Markov random fields. Theory and Applications*. Chapman & Hall, 2005.

U. Schmidt, Q. Gao, and S. Roth. A generative perspective on MRFs in low-level

vision. In *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition*, 2010.

M. Schneider and A. Willsky. Krylov subspace estimation. *SIAM J. Sci. Comp.*, 22(5):1840–1864, 2001.

M. Seeger and H. Nickisch. Large scale Bayesian inference and experimental design for sparse linear models. *SIAM J. Imaging Sci.*, 4(1):166–199, 2011a.

M. Seeger and H. Nickisch. Fast convergent algorithms for expectation propagation approximate Bayesian inference. In *Proc. Int. Conf. on Artificial Intelligence and Statistics*, 2011b.

F. Steutel and K. Van Harn. *Infinite divisibility of probability distributions on the real line*. Dekker, 2004.

E. Sudderth, M. Wainwright, and A. Willsky. Embedded trees: Estimation of Gaussian processes on graphs with cycles. *IEEE Trans. Signal Process.*, 52(11): 3136–3150, 2004.

R. Szeliski. Bayesian modeling of uncertainty in low-level vision. *Int. J. of Comp. Vis.*, 5(3):271–301, 1990.

M. Szummer, P. Kohli, and D. Hoiem. Learning CRFs using graph cuts. In *Proc. European Conf. on Computer Vision*, 2008.

D. Tarlow, R. Adams, and R. Zemel. Randomized optimum models for structured prediction. In *Proc. Int. Conf. on Artificial Intelligence and Statistics*, 2012.

B. Taskar, C. Guestrin, and D. Koller. Max-margin Markov networks. In *Proc. Advances in Neural Information Processing Systems*, 2003.

D. Terzopoulos. The computation of visible-surface representations. *IEEE Trans. Pattern Anal. Mach. Intell.*, 10(4):417–438, 1988.

M. Tipping. Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1:211–244, 2001.

M. van Gerven, B. Cseke, F. de Lange, and T. Heskes. Efficient Bayesian multivariate fMRI analysis using a sparsifying spatio-temporal prior. *NeuroImage*, 50: 150–161, 2010.

M. Wainwright and M. Jordan. Graphical models, exponential families, and variational inference. *Found. and Trends in Machine Learning*, 1(1-2):1–305, 2008.

M. Wainwright, T. Jaakkola, and A. Willsky. MAP estimation via agreement on trees: Message-passing and linear programming. *IEEE Trans. Inf. Theory*, 51 (11):3697–3717, 2005.

Y. Weiss and W. Freeman. Correctness of belief propagation in Gaussian graphical models of arbitrary topology. *Neural Computation*, 13(10):2173–2200, 2001.

Y. Weiss and W. Freeman. What makes a good model of natural images? In *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition*, 2007.

M. Welling. Herding dynamical weights to learn. In *Proc. Int. Conf. on Machine Learning*, 2009.

S. Zhu, Y. Wu, and D. Mumford. Filters, random fields and maximum entropy (FRAME): Towards a unified theory for texture modeling. *Int. J. of Comp. Vis.*, 27(2):107–126, 1998.