Approximate Parameter Learning in Discriminative Fields

Sanjiv Kumar and Martial Hebert
The Robotics Institute
Carnegie Mellon University
Pittsburgh, PA 15213
{skumar,hebert}@ri.cmu.edu

Abstract

In this paper, we present an approach for approximate maximum likelihood parameter learning in discriminative field models, which is based on approximating true expectations with simple piecewise constant functions constructed using inference techniques. Gradient ascent with these updates shows interesting weak-convergence behavior which is tied closely to the number of errors made during inference. The performance of various approximations was evaluated with different inference techniques showing that the learned parameters lead to good classification performance so long as the method used for approximating the gradient is consistent with the inference mechanism. The proposed approach is general enough to be used for conditional training of conventional MRFs.

1 Introduction

In language processing, natural image analysis etc., the input data shows significant dependencies, which should be modeled appropriately to achieve good classification. In earlier work [1], we presented the Discriminative Random Field (DRF) model for image analysis, which is a type of Conditional Random Field (CRF) proposed by Lafferty et al. [2]. These fields discriminatively model the conditional distribution of the labels given the observed data directly as a Markov Random Field (MRF) and were shown to give better results than the conventional MRFs [1]. The CRFs were developed in the context of analyzing 1D sequence data for which exact maximum likelihood parameter learning is feasible using efficient techniques, e.g. iterative scaling [2], quasi-Newton methods [3] etc. However, when the graphs contain loops, it is not feasible to exactly maximize the likelihood with respect to the parameters. Therefore, a critical issue for the discriminative fields to be practical is the design of effective parameter learning techniques that can operate on arbitrary graphs without needing any hand-tuned control parameters [1]. The objective of this paper is to address this central question.

Through our analysis and experiments, we hope to draw three conclusions: First, *parameter learning* can be achieved by approximating the likelihood gradient using the label estimates obtained through methods such as Maximum A Posteriori (MAP) or Maximum Posterior Marginal (MPM) given the conditional probability model. Second, good classification performance can be achieved by any of the choices of approximation, so long as the method used for inference matches the method used for approximating the gradient in

the parameter learning. We further show that this *learning/inference duality* is rooted in the fact that, in any classification problem, the goal is to minimize the number of errors, which is what our gradient approximation does, which may not necessarily maximize the likelihood. Finally, we present an *explicit comparison* between several choices for which currently there exists no formal comparison on what type of learning approximation should be adopted for a given choice of inference method.

1.1 Discriminative fields

In this section, we review the formulation of discriminative fields. Although the formulation is general to arbitrary graphs with multiple class labels [4], we will discuss the problem of learning in the context of binary classification on 2D image lattices. Let $\boldsymbol{y}$ be the observed data from an input image, where $\boldsymbol{y} = \{\boldsymbol{y}_i\}_{i \in S}$, $\boldsymbol{y}_i$ is the data from i^{th} site, and S is the set of sites. Let the corresponding labels be given by $\boldsymbol{x} = \{x_i\}_{i \in S}$ where $x_i \in \{-1, 1\}$. The DRF formulation combines local discriminative models to capture the class associations at individual sites with the interactions in the neighboring sites as:

$$P(\boldsymbol{x}|\boldsymbol{y}) = \frac{1}{Z} \exp\left(\sum_{i \in S} \log P'(x_i|\boldsymbol{y}) + \sum_{i \in S} \sum_{j \in \mathcal{N}_i} x_i x_j \boldsymbol{v}^T \boldsymbol{\mu}_{ij}(\boldsymbol{y}) \right) \tag{1}$$

where Z is the partition function, $\boldsymbol{v}$ are the model parameters, and $\boldsymbol{\mu}_{ij}(\boldsymbol{y})$ are the pairwise relational feature vectors. Note that both terms depend explicitly on all the observations $\boldsymbol{y}$. Here, $P'(x_i|\boldsymbol{y})$ is the local class posterior returned by an arbitrary discriminative classifier. This gives the flexibility to choose domain-specific discriminative classifiers suitable for specific task domains. In this paper, as in our previous work [1], we use a logistic link to give the local class posterior, i.e. $P'(x_i|\boldsymbol{y}) = \sigma(x_i \boldsymbol{w}^T \boldsymbol{h}_i(\boldsymbol{y}))$ where $\sigma(t) = 1/1+e^{-t}$. Here, $\boldsymbol{w}$ are the model parameters, and $\boldsymbol{h}_i(\boldsymbol{y})$ are the sitewise feature vectors. Note that this choice leads to a form of unary potential which is linear in parameters similar to the CRFs given in [2]. So, this particular DRF form can be seen as a 2D extension of 1D CRFs.

2 Parameter learning approaches

2.1 Maximum likelihood parameter learning

Let θ be the set of DRF parameters where $\theta = \{\boldsymbol{w}, \boldsymbol{v}\}$. Given M i.i.d. labeled training images, the maximum likelihood estimates of the parameters are given by maximizing the log-likelihood $l(\theta) = \sum_{m=1}^{M} \log P(\boldsymbol{x}^m|\boldsymbol{y}^m, \theta)$ i.e.,

$$\widehat{\theta} = \underset{\theta}{\operatorname{argmax}} \sum_{m=1}^{M} \left\{ \sum_{i \in S^m} \log \sigma(x_i^m \boldsymbol{w}^T \boldsymbol{h}_i(\boldsymbol{y}^m) + \sum_{i \in S^m} \sum_{j \in \mathcal{N}_i} x_i^m x_j^m \boldsymbol{v}^T \boldsymbol{\mu}_{ij}(\boldsymbol{y}^m) - \log Z^m \right\} \tag{2}$$

Where Z^m is the partition function for the m^{th} image, which is given as $Z^m = \sum_{\boldsymbol{x}} \exp\left\{ \sum_{i \in S^m} \log \sigma(x_i \boldsymbol{w}^T \boldsymbol{h}_i(\boldsymbol{y}^m) + \sum_{i \in S^m} \sum_{j \in \mathcal{N}_i} x_i x_j \boldsymbol{v}^T \boldsymbol{\mu}_{ij}(\boldsymbol{y}^m) \right\}$. Note that Z^m is a function of the parameters θ and the observed data $\boldsymbol{y}^m$. To learn the parameters using gradient ascent, the derivatives of the log-likelihood can be written as,

$$\frac{\partial l(\theta)}{\partial \boldsymbol{w}} = \frac{1}{2} \sum_{m} \sum_{i \in S^m} (x_i^m - \langle x_i \rangle) \boldsymbol{h}_i(\boldsymbol{y}^m) \tag{3}$$

$$\frac{\partial l(\theta)}{\partial \boldsymbol{v}} = \sum_{m} \sum_{i \in S^m} \sum_{j \in \mathcal{N}_i} (x_i^m x_j^m - \langle x_i x_j \rangle) \boldsymbol{\mu}_{ij}(\boldsymbol{y}^m) \tag{4}$$

Here $\langle . \rangle$ denotes expectation with $P(\boldsymbol{x}|\boldsymbol{y}^m, \theta)$. Ignoring μ_{ij}, gradient ascent with (4) resembles the learning problem in Boltzmann machines with all the nodes being observed at the training stage and computing the expectations can be seen as the 'free' phase.

Generally the expectations in (3) and (4) cannot be computed analytically due to the combinatorial size of the label space. Sampling procedures, *e.g.* Markov Chain Monte Carlo (MCMC), can be used to approximate the true expectation. But, MCMC techniques have two main problems, i.e. long 'burn-in' period which make them slow, and high variance in estimates [5]. To avoid MCMC drawbacks, Contrastive Divergence (CD) was proposed by Hinton [5]. In CD, only a single MCMC move is made from the current empirical distribution of the data (P^0) leading to new distribution (P^1), thus eliminating the need of running the chain beyond burn-in. According to this, $\langle x_i \rangle \approx \langle x_i \rangle_{P^1}$ and $\langle x_i x_j \rangle \approx \langle x_i x_j \rangle_{P^1}$. Even though CD is computationally simple and yields estimates with low variance, the bias in estimates can be a problem [6], which was also verified in our experiments in Section 6. However, this approximation of expectation using single sample forms the basis for different approximations we use in this work, as shown in the next section.

2.2 Coupled parameter learning/inference approaches

The approximations defined in the previous section replace the exact gradient of (3) and (4) by expressions of the form:

$$\frac{\partial l(\theta)}{\partial \boldsymbol{w}} = \frac{1}{2} \sum_m \sum_{i \in S^m} (x_i^m - f_i(\theta)) \boldsymbol{h}_i(\boldsymbol{y}^m), \tag{5}$$

$$\frac{\partial l(\theta)}{\partial \boldsymbol{v}} = \sum_m \sum_{i \in S^m} \sum_{j \in \mathcal{N}_i} (x_i^m x_j^m - g_{ij}(\theta)) \boldsymbol{\mu}_{ij}(\boldsymbol{y}^m). \tag{6}$$

Here, f_i and g_{ij} are functions that approximate the true expectations in the gradient. Several approaches have been proposed that compute f_i and g_{ij} using pseudo-marginals [7][8]. In this work, we propose to directly construct f_i and g_{ij} using label estimates obtained through inference at the current parameter estimates as explained in Section 3.2 and 3.3.

The first question is whether replacing the gradient by such an approximation leads to a convergent parameter learning procedure. The answer is that, while the learning procedure is not strictly convergent in general, it is weakly convergent in that it oscillates within a set of good parameters, or converges to a good parameter with isolated large deviations, as shown experimentally in Section 4 and justified in Section 5. The second question is why the parameters learned using a particular choice of approximating functions should yield good classification performance? Informally, if we use for parameter learning the same approximating function f_i that was obtained from inference (e.g. MAP label estimate), then, given input training labels $\{x_i^m\}$,

$$N_E^\theta = \frac{1}{2} \sum_m \sum_{i \in S^m} |x_i^m - f_i(\theta)| \tag{7}$$

can be interpreted as the number of errors in classification. Comparing (7) with (5) shows that the approximated gradient is directly related to the number of errors, so long as the *same approximation is used in both parameter learning and inference*. We will show empirical observations in Section 4 and the formal analysis in Section 5.

3 Candidate Approximations

We first explore the form of f_i and g_{ij} based on pseudo-marginals, and then using two approximations directly based on two different label estimates: Maximum A Posteriori (MAP) which is optimal for 0-1 loss function, and Maximum Posterior Marginal (MPM) which is optimal for 'sitewise' 0-1 loss function. For the binary DRFs, approximate MAP estimates can be obtained using the min-cut/max-flow algorithms as explained in [1]. We use the sum-product version of loopy Belief Propagation (BP) to obtain the MPM estimates [9]. The approximations described below are designed to match these two classes of inference techniques.

3.1 Pseudo Marginal Approximation (PMA)

It is easy to see that, if we had true marginal distributions at each site i, $P_i(x_i|\boldsymbol{y})$, and at each pair of sites i and $j \in \mathcal{N}_i$, $P_{ij}(x_i, x_j|\boldsymbol{y})$, we could compute exact expectations as:

$$\langle x_i \rangle = \sum_{x_i} x_i P_i(x_i|\boldsymbol{y}) \quad \text{and} \quad \langle x_i x_j \rangle = \sum_{x_i, x_j} x_i x_j P_{ij}(x_i, x_j|\boldsymbol{y})$$

Since computing exact marginal distributions is in general infeasible, the simplest thing will be to replace the actual marginals by pseudo marginals. In this work, we used loopy BP to get these marginals, because we use BP to do inference to get MPM estimates. In addition, these marginals are expected to return better approximation than mean-field as the fixed points of BP correspond to the stationary points of Bethe free energy [9]. McCallum et al. [8] use similar approximation, where TRP based pseudo marginals were used for parameter learning in Factorial CRFs.

3.2 Saddle Point Approximation (SPA)

Here, we propose a very simple approximation inspired by CD [5], using MAP label estimates. It is based on approximating the partition function (Z) using the Saddle Point Approximation (SPA). According to this, Z is approximated such that the summation over all the label configurations $\boldsymbol{x}$ in Z is replaced by the most probable label configuration. In other words, if $\hat{\boldsymbol{x}} = \arg\max_{\boldsymbol{x}} P(\boldsymbol{x}|\boldsymbol{y}, \theta)$, then according to SPA, $Z \approx \exp\left\{\sum_{i \in S} \log \sigma(\hat{x}_i \boldsymbol{w}^T \boldsymbol{h}_i(\boldsymbol{y})) + \sum_{i \in S} \sum_{j \in \mathcal{N}_i} \hat{x}_i \hat{x}_j \boldsymbol{v}^T \boldsymbol{\mu}_{ij}(\boldsymbol{y})\right\}$. This leads to a very simple approximation to the expectations in (3) and (4): $\langle x_i \rangle \approx \hat{x}_i$ and $\langle x_i x_j \rangle \approx \hat{x}_i \hat{x}_j$. It is interesting to note that with the saddle point approximation of Z, the gradient ascent updates are similar to the perceptron-learning type updates used in [10] and [11] in non-probabilistic settings.

3.3 Maximum Marginal Approximation (MMA)

This is the second approximation based on BP inference in which MPM label estimates are used for approximating the expectations. Following the arguments of SPA based parameter learning in the previous section, one can make a similar approximation of Z such that all the mass of Z is assumed to be concentrated on the maximum marginal configuration. If $\tilde{x}_i = \arg\max_{x_i} P_i(x_i|\boldsymbol{y}, \theta)$, the expectations can be written as: $\langle x_i \rangle \approx \tilde{x}_i$ and $\langle x_i x_j \rangle \approx \tilde{x}_i \tilde{x}_j$. Clearly, in the binary case, maximum marginals are just the thresholded sitewise marginals. Thus, MMA can be interpreted as a discrete approximation of PMA. We use this approximation to generalize the understanding of the discrete approximations of actual expectations which was also done in the case of SPA.

4 Experimental observations: parameter learning

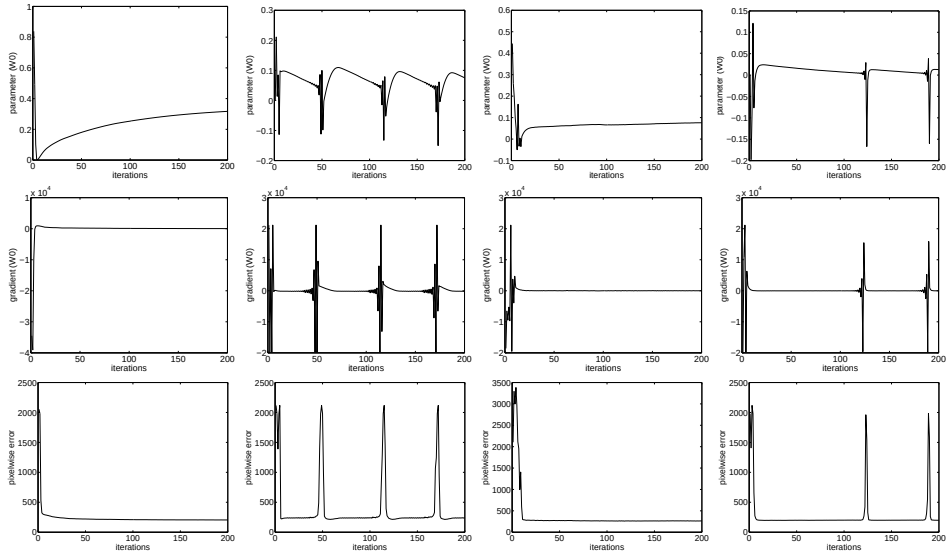
To analyze the convergence performance of various parameter learning procedures described in the previous section, we learned a DRF model for a binary image denoising application. The aim was to obtain true labels from corrupted binary images. A binary image (64×64 pixels) was corrupted by two types of noise: Gaussian noise and Bimodal (mixture of two Gaussians) noise. For each noise model, 10 noisy images were used as the training set for learning the parameters. The details on the features $\boldsymbol{h}_i(\boldsymbol{y})$ and $\boldsymbol{\mu}_{ij}(\boldsymbol{y})$, and the noise parameters of this dataset are given in [1]. Here, the parameter vectors $\boldsymbol{w}$ and $\boldsymbol{v}$ both were two-element vectors, i.e. $\boldsymbol{w} = [w_0 \;\; w_1]$, and $\boldsymbol{v} = [v_0 \;\; v_1]$.

In all the experiments, parameters were initialized from random values and updates were based on gradient ascent. The step size η was fixed to a small value (10^{-5}). Fig. 1 shows for each approximation, plots of the approximated gradients and the parameters at each iteration for a typical run with bimodal noise. For brevity we show plots only for parameter w_0. The rest of the parameters also gave similar plots. The plots in the last row in Fig. 1

show the number of errors (N_E^θ) made at the current estimate of the parameters using the same inference technique on which a particular gradient approximation is based.

For PMA based learning, parameters and gradients were always found to converge (Fig. 1 (a)), and the final parameters values were independent of the initialization as for the true gradient descent, since the log likelihood in (2) is a convex function of parameters. This indicates that the beliefs from loopy BP were converging to reasonable estimates for this dataset.

For SPA and MMA based learning, an interesting behavior emerges since both of them make discrete approximations of the true expectations. It was found that both SPA and MMA show two different scenarios of weak convergence depending on the parameter initialization. For SPA, in the first case (Fig. 1 (b)), the approximated gradients for all the parameters show oscillatory behavior. Initially there are large oscillations in gradients which settle down to low gradient zone. The gradients remain in this zone for a long duration before showing large oscillations with changing sign again. Note that this will not occur for the gradient ascent with true gradients if suitably small η is chosen. To find the best set of parameters, one can use commonly used heuristics to find better parameter values when convergence is not guaranteed, e.g. voted perceptron used by Collins [11]. In this work we used a simpler choice of majority vote parameter setting.



(a) PMA (b) SPA-1 (c) SPA-2 (d) MMA

Figure 1: Plots of DRF parameter (w_0) updates (top row), and the likelihood gradient (second row) for different approximations. PMA shows a converging behavior while the SPA shows oscillatory behavior which may be large (SPA-1) or microscopic (SPA-2). MMA shows similar behavior as SPA. The last row shows number of errors at each parameter update. The errors are low when the gradient magnitudes are small.

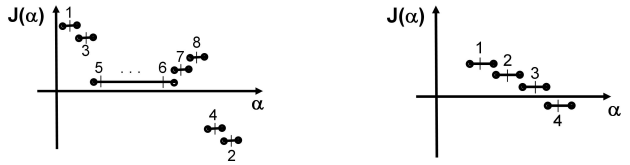
In the second case for SPA, (Fig. 1 (c)), after initial oscillations, the gradients do not show 'periodic' large oscillations again but maintain microscopic oscillations within low gradient zones (not visible in the figure due to the scale of the plots). The MMA based learning showed similar behavior as for the SPA indicating that these behaviors are related to the discrete, piecewise constant approximation of the actual expectations. An oscillating gradients case for MMA is shown in Fig. (1 (c)). In Section 5 we will analyze the weak convergence characteristics of the SPA and MMA based learning procedures.

Finally, note that number of errors for all approximations is small whenever gradient magnitudes are small which indicates that all the three techniques are trying to achieve parameter values that minimize the errors for that particular inference. This is especially interesting in the case of SPA and MMA because of the nature of the approximations. We will compare the performance of the parameter learning procedures with different inference techniques on a separate test set in Section 6.

5 Analysis of SPA/MMA based learning

Let us denote the gradients in (5) and (6) by $J(\theta)$. Also, assume a single training image, $M = 1$ for simplicity. Note that, for both SPA and MMA, $f_i(\theta)$ and $g_{ij}(\theta) = f_i(\theta)f_j(\theta)$ are piecewise constant functions, computed using the label estimates. Let Θ be the space of all θ. Thus, if $\Theta = \bigcup_k \Theta_k$ s.t. $f_i(\theta) = f_{ik} \ \forall\, \theta \in \Theta_k, \forall\, i \in S$ then $J(\theta) = J_k \ \forall\, \theta \in \Theta_k$.

Further, we focus on a single component of $J(\theta)$, i.e. $J(\alpha) = H^{\alpha T}(X - F)$ where X, F and H^α are vectors with components, $X_t = x_i$, $F_t = f_i(\theta)$ and $H_t^\alpha = \boldsymbol{h}_i^\alpha(\boldsymbol{y})$ if $\alpha \in \boldsymbol{w}$, and $X_t = x_i x_j$, $F_t = f_i(\theta)f_j(\theta)$ and $H_t^\alpha = \boldsymbol{\mu}_{ij}^\alpha(\boldsymbol{y})$ if $\alpha \in \boldsymbol{v}$. For a given training set, H^α and X are fixed. The gradient ascent will converge when $J(\theta) = \mathbf{0}$, i.e. $J(\alpha) = 0\ \forall\, \alpha \in \theta$. For this, there must exist a θ such that projection of $(X - F)$ on H^α is $0\ \forall\, \alpha \in \theta$. Since, $H_t^\alpha \in \Re$ and $(X_t - F_t) \in \{2, 0, -2\}$, for generic H^α and X, this will happen only if $F = X$ (i.e. zero error case). Thus, convergence requires existence of such θ for which all the sites are correctly labeled while training, which is generally infeasible. However, $J(\alpha)$ will be 'tiled' with many zero-crossings depending on the conditional probability form and the inference criterion. To analyze the weak-convergence of gradient ascent shown in Section 4, we need to consider only those zero-crossings for which $J(\alpha^-) > 0$, and $J(\alpha^+) < 0$. Suppose that θ is initialized such that, in its neighborhood, the tiling of $J(\theta)$ is similar to the one shown in Fig. 2(a). Then, for a fixed η, initially there will be large oscillations with gradients changing signs (1 to 4). Then, the gradients will be low for many iterations (5 to 6) before showing the 'divergence buildup' in steps 6 to 8. Finally, the parameters will jump from 8 to 4 and then to 3, repeating the cycle. This explains the 'periodic' oscillations in gradients and parameters observed for SPA-1 in Fig. 1(b). On the contrary, for the configuration of $J(\theta)$ in Fig. 2(b), we will have only microscopic oscillations in gradients between steps 3 and 4 as observed for SPA-2.



(a) SPA-1 (b) SPA-2

Figure 2: The gradient ascent updates of SPA based parameter learning in two cases. The numbers represent the sequence of parameter updates.

As shown in Fig. 1, the number of classification errors are closely tied to the approximated gradients. If we define the number of errors at the parameter estimate θ as in (7), i.e. $N_E^\theta = (1/2)\sum_{i \in S} |x_i - f_i(\theta)|$, then using the form of $J(\boldsymbol{w})$ in (5), $\|J(\boldsymbol{w})\| \leq R N_E^\theta$, where $R = \max_i \|\boldsymbol{h}_i(\boldsymbol{y})\|$ and $\|.\|$ is the L_1 norm. Similarly, define the pairwise error N_P^θ as, $N_P^\theta = (1/2)\sum_{i \in S}\sum_{j \in \mathcal{N}_i} |x_i x_j - f_i(\theta)f_j(\theta)|$. Using the form of $J(\boldsymbol{v})$ in (6) with $g_{ij}(\theta) = f_i(\theta)f_j(\theta)$, it is easy to see that $\|J(\boldsymbol{v})\| \leq 2Q N_P^\theta$, where $Q = \max_{ij} \|\boldsymbol{\mu}_{ij}(\boldsymbol{y})\|$. This implies, $\|J(\boldsymbol{v})\| \leq 2QdN_E^\theta$, since $N_P^\theta \leq dN_E^\theta$ where d is the maximum degree of the graph, i.e. $d = \max_i |\mathcal{N}_i|$. Two conclusions can be drawn from this discussion. First, if $\|J(\theta)\|$ is large, then N_E^θ is also large as verified in the plots in Fig. 1. Second, if at

some θ, N_E^θ is small, $\|J(\theta)\|$ will also be small. Thus, for a suitably small step size η, the parameter change will also be small. This would mean that one will stay in a low error zone for a long period as seen in Fig. 1.

Finally, we need to argue that small $\|J(\theta)\|$ implies, in general, small N_E^θ. In other words, if the projection of $(X - F)$ on each plane $H^\alpha \; \forall \; \alpha \in \theta$ is small, it implies $\|X - F\|$ is small. This phenomenon is related to the MAP or MPM inference criteria of selecting F given a probability model. We leave the formal reasoning for future exploration. However, it is worth mentioning that in addition to the observations in this work, we also observed this behavior empirically even on real-world problems with large number of parameters [4].

6 Experimental observations: inference

The aim of these experiments was to compare the performance of different parameter learning procedures for a *fixed* inference procedure. For each noise model introduced in Section 4, a test set of 200 noisy images was generated using 50 noisy images each from four base images. For comparison, we also obtain the local MAP solution using Iterated Conditional Modes (ICM) [12] which has been shown to be robust to incorrect parameter settings. In addition, we also compare results with parameters learned through pseudo-Likelihood (PL), which uses a factored approximation of the partition function for tractability [1].

The overall pixelwise errors on the test set are given in table 1. There are three key observations: First, the MAP inference works the best with SPA parameters (both use min-cut), and MPM works the best with PMA parameters (both use BP), empirically verifying the claim of *learning/inference duality*. Second, for the MAP inference, SPA based learning is the most accurate as well as the most efficient approach. The SPA learning is more than 14 times faster than the next best set of parameters, i.e. from PMA. Last, MMA is able to learn reasonable parameters for MPM inference (both use BP), at almost half the training time for PMA at the cost of slight decrease in performance from PMA. Note that both PMA and MMA use BP at the learning stage and slightly better results of PMA may be because PMA returns a single converged estimates of the parameters while in MMA one has to heuristically pick the best set of parameters. Better performance may be expected if a better heuristic is used instead of picking the majority voted parameters.

Table 1: Pixelwise classification errors ($\%$) on 200 test images (64×64 pixels each). The rows show different parameter learning procedures and the columns show different inference techniques used for two different noise models. See text for more.

	Gaussian noise			Bimodal noise			Learning time
	MAP	MPM	ICM	MAP	MPM	ICM	(Sec)
PMA	2.73	**2.51**	3.91	6.45	**5.48**	17.39	1183.13
SPA	**2.49**	7.64	3.98	**5.82**	19.19	14.88	81.52
MMA	34.34	2.96	4.11	26.53	5.70	16.00	635.78
PL	3.82	3.10	**3.89**	17.69	7.31	22.22	299.75
CD	3.78	2.82	4.09	8.88	6.29	**8.92**	206.93
Inference time (Sec)	5.52	90.04	5.20	5.96	113.84	5.20	

An interesting observation is that the MAP inference is very poor with MMA parameters and the same is true for MPM inference with SPA parameters. This further enforces the idea that learning/inference duality is rooted in minimizing the classification error for a learning/inference pair, rather than maximizing the true likelihood.

As a by-product of this comparison, we find that MPM inference is more robust to the parameters returned by other techniques than MAP which gives significantly worse results

with parameters other than SPA and PMA. In addition, the PL and CD parameters generally give bad estimates while ICM does poor inference due to the problem of label initialization.

Finally, if it is argued that MPM works well with PMA, not because of duality, but because PMA returns true ML parameters (i.e. BP converged to true marginals), the results indicate that MAP inference is not the best with ML parameters and the inference based approximation (SPA in this case) may yield better results enforcing the role of duality even when exact ML parameter learning is possible.

7 Conclusion and future work

We have presented an approach for learning the parameters of discriminative field models, which uses inference to approximate the gradients used in maximum likelihood learning. We showed that proposed approximations lead to a weakly convergent behavior of the learning procedures. Further, the learned parameters lead to good classification performance so long as the method used for approximating the gradient is consistent with the inference mechanism. We also provided an experimental comparison of commonly used learning and inference techniques for discriminative fields. For MAP inference, SPA based learning was found to be most accurate as well as efficient. In fact, although we limited the presentation to the restricted case of binary fields, we have already used maximum marginal approximation to successfully learn more than 3000 parameters for multiclass DRFs applied to object detection [4]. We are currently evaluating the performance of the proposed approximate parameter learning procedures with conventional MRFs.

Acknowledgments

Our thanks to J. August and T. Minka for very helpful discussions on SPA based learning. Thanks to J. Lafferty and R. Mugizi for providing the min-cut code.

References

[1] S. Kumar and M. Hebert. Discriminative fields for modeling spatial dependencies in natural images. *in advances in Neural Information Processing Systems (NIPS)*, December 2003.

[2] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *In Proc. Int. Conf. on Machine Learning*, 2001.

[3] F. Sha and F. Pereira. Shallow parsing with conditional random fields. *In Proc. Human Language Technology-NAACL*, 2003.

[4] S. Kumar and M. Hebert. Multiclass discriminative fields for parts-based object detection. *Snowbird Learning Workshop, Utah*, 2004.

[5] G. E. Hinton. Training product of experts by minimizing contrastive divergence. *Neural Computation*, 14:1771–1800, 2002.

[6] C. K. I. Williams and F. V. Agakov. *An Analysis of Contrastive Divergence Learning in Gaussian Boltzmann Machines*. EDI-INF-RR-0120, Informatics Research Report, May 2002.

[7] M. J. Wainwright, T. Jaakkola, and A. S. Willsky. Tree-reweighted belief propagation and approximate ml estimation by pseudo-moment matching. *9th Workshop on AI Stat*, 2003.

[8] A. McCallum, K. Rohanimanesh, and C. Sutton. Dynamic conditional random fields for jointly labeling multiple sequences. *NIPS'03 workshop on Syntax, Semantics and Statistic*, 2003.

[9] J. S. Yedidia, W. T. Freeman, and Yair Weiss. Generalized belief propagation. *In Advances Neural Information Processing Systems*, 13:689–695, 2001.

[10] Y. LeCun, L. Bottou, Y. Bengio, and P Haffner. Gradient-based learning applied to document recognition. *Proc. of the IEEE*, 86(11):2278–2324, 1998.

[11] M. Collins. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. *In Proc. EMNLP*, 2002.

[12] J. Besag. On the statistical analysis of dirty pictures. *Journal of Royal Statistical Soc.*, B-48:259–302, 1986.