Putting MAP back on the map

Patrick Pletscher[1], Sebastian Nowozin[2], Pushmeet Kohli[2], and Carsten Rother[2]

[1] ETH Zurich, Switzerland
[2] Microsoft Research Cambridge, UK

Abstract. Conditional Random Fields (CRFs) are popular models in computer vision for solving labeling problems such as image denoising. This paper tackles the rarely addressed but important problem of learning the full form of the potential functions of pairwise CRFs. We examine two popular learning techniques, maximum likelihood estimation and maximum margin training. The main focus of the paper is on models such as pairwise CRFs, that are simplistic (misspecified) and do not fit the data well. We empirically demonstrate that for misspecified models maximum-margin training with MAP prediction is superior to maximum likelihood estimation with any other prediction method. Additionally we examine the common belief that MLE is better at producing predictions matching image statistics.

1 Introduction

Many computer vision tasks can be cast as an image labeling problem. Applications include semantic image segmentation [8], background-foreground segmentation [14] or image denoising [18,16]. Structured models such as Markov Random Fields and Conditional Random Fields have been successfully applied in this context and shown in practice to outperform other methods. These models combine local evidence, dependencies between neighboring pixels and possibly global cues for specifying the probability of a labeling. The usage of a structured model requires *learning* and *inference* (prediction). Learning consists of estimating the parameters of the model (i.e., the potentials) from labeled training data, while inference is the task of predicting a labeling for a given image.

Inference has received a lot of attention in recent years, the dominant approach being maximum-a-posteriori (MAP) inference, for which several efficient and accurate approximate algorithms have been developed [2,6]. On the other hand, learning is still predominantly done by either hand-tuning the parameters or performing a grid-search over a number of settings. This work considers the relation between learning and inference for image labeling using a structured model. Two approaches are discussed here. The classical approach estimates the parameters $\boldsymbol{w}$ of the posterior distribution $P(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w})$ of a label $\boldsymbol{y}$ given an image $\boldsymbol{x}$ using maximum likelihood. It predicts the label according to Bayesian decision theory, which requires the specification of a suitable loss function Δ. Depending on the loss function, different prediction functions are obtained, such as MAP, maximum marginal or minimum mean squared error (MMSE). Expected risk minimization is the second approach, it directly trains a prediction

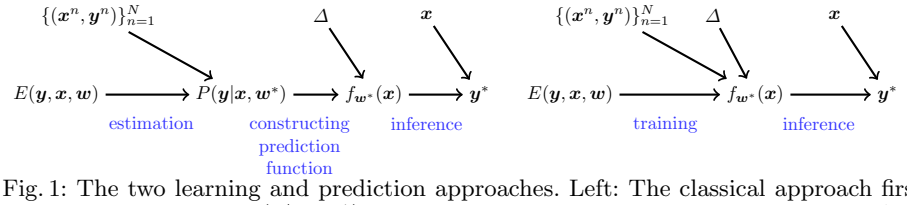

Fig. 1: The two learning and prediction approaches. Left: The classical approach first estimates a posterior $P(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w}^*)$ from training data and incorporates the loss Δ at test-time to infer the optimal label. Right: The alternative approach directly trains a classifier for a specific loss and skips the distributional estimation step.

function which already incorporates the loss. Training and inference in these two paradigms is visualized in Fig. 1. The first approach is known to be superior in the ideal case where the model accurately describes the underlying image acquisition process and sufficient amount of training data is given such that the parameters can be correctly estimated.

The primary contribution of this paper is to show that there exist practical situations in which the direct learning of a prediction function yields better performance. We show that such settings arise when the assumed model does not fully capture the dependencies in the data, a situation referred to as *misspecification* [20]. This is an important insight for computer vision applications since the data generating process is rather complicated and thus often inaccurately modeled. As a second contribution we show that it is possible to learn the full potentials of pairwise structured models even for relatively large state spaces such as in image denoising applications. We conclude that through appropriate training, efficient MAP inference can perform on par with more complex prediction functions such as MMSE. In particular, we also demonstrate that MAP is as good at reproducing image statistics as MMSE. Another goal of this work is to review important facts about prediction and learning for image labeling problems, which we feel are not well-known in the computer vision community.

2 The image labeling problem

Most structured models for image labeling problems can be expressed as an *energy function* of a labeling $\boldsymbol{y}$, an image $\boldsymbol{x}$ and parameters $\boldsymbol{w}$ of the form

$$E(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{w}) = \sum_{t \in \mathcal{T}} \sum_{c \in \mathcal{C}_t} \psi_\alpha(\boldsymbol{y}_c, \boldsymbol{x}, \boldsymbol{w}^t). \tag{1}$$

The model factorizes into cliques which are assumed to be grouped into sets (templates) t that share the same parameter $\boldsymbol{w}^t$. y_i is assumed to be in the set $\{0, \dots, K-1\}$, leading to a total of K^M possible labelings. Here M denotes the number of sites for which a label is predicted. The CRF assumes that the posterior of a labeling for an observed image is given by the Gibbs distribution

$$P(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w}) = \frac{1}{Z(\boldsymbol{x}, \boldsymbol{w})} \exp(-E(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{w})), \tag{2}$$

with partition sum $Z(\boldsymbol{x}, \boldsymbol{w}) = \sum_{\boldsymbol{y}'} \exp(-E(\boldsymbol{y}', \boldsymbol{x}, \boldsymbol{w}))$. In the context of structured models it is usually assumed that the model depends linearly on the parameters [5, Section 4.4.1.2], which we also do here. To make this linear dependence explicit, the energy in (1) is rewritten as $E(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{w}) = -\langle \boldsymbol{w}, \boldsymbol{s}(\boldsymbol{x}, \boldsymbol{y}) \rangle$. Here

$\boldsymbol{s}(\boldsymbol{x}, \boldsymbol{y})$ denotes the sufficient statistics which counts using indicator functions the different configurations of the cliques in (1). We will discuss an example of such a sufficient statistics in more detail in the next section.

2.1 Image Denoising

In this work we discuss as a running example the problem of image denoising. Given an observed noisy image the goal is to reconstruct the original noise-free image. For this task we consider a simple pairwise CRF to illustrate all the concepts. The labeling $\boldsymbol{y}$ in this context is the reconstruction of the original image and $\boldsymbol{x}$ denotes the noisy observation. The energy is assumed to be

$$E(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{w}) = -\sum_{i \in \mathcal{V}} w^u_{|y_i - x_i|} - \sum_{(i,j) \in \mathcal{E}} w^p_{|y_i - y_j|}, \tag{3}$$

where the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is the standard 4-neighborhood grid commonly used in computer vision. The potentials have one parameter for each possible outcome of the unary and pairwise term, respectively. This results in a total of $2K$ parameters. We denote by w^u_j the j-th component of the unary parameter $\boldsymbol{w}^u$ and similarly for $\boldsymbol{w}^p$ the pairwise parameter. For this simple image denoising model the sufficient statistics $\boldsymbol{s}(\boldsymbol{x}, \boldsymbol{y}) = [\boldsymbol{s}^u(\boldsymbol{x}, \boldsymbol{y})^\mathsf{T}, \boldsymbol{s}^p(\boldsymbol{y})^\mathsf{T}]^\mathsf{T}$ are thus given by

$$s^u_k(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i \in \mathcal{V}} \delta_k(|x_i - y_i|), \;\; s^p_k(\boldsymbol{y}) = \sum_{(i,j) \in \mathcal{E}} \delta_k(|y_i - y_j|).$$

Here $\delta_k(z)$ denotes the Kronecker delta function which evaluates to one if $z = k$ and to zero otherwise. For image denoising the state space of the variables y_i is typically quite large, for example $K = 256$ for a grayscale image.

2.2 Learning and Prediction

Most image labeling applications come with some form of labeled training data on which a parameter $\boldsymbol{w}^*$ is learned according to some objective. We will discuss maximum margin learning and maximum likelihood estimation. Having determined $\boldsymbol{w}^*$, the inference task considers predicting the optimal labeling $\boldsymbol{y}^*$ for an observed image. There exist several approaches for this, which we will discuss in § 4. The most popular prediction function is the MAP inference which can be understood as maximizing the posterior distribution in a CRF

$$\boldsymbol{y}^* = \operatorname*{argmin}_{\boldsymbol{y}} E(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{w}^*) = \operatorname*{argmax}_{\boldsymbol{y}} \langle \boldsymbol{w}^*, \boldsymbol{s}(\boldsymbol{x}, \boldsymbol{y}) \rangle.$$

Its popularity stems from the fact that efficient MAP inference algorithms such as graph-cut or TRW-S exist. Strictly speaking, the MAP interpretation of a labeling having minimal energy is only valid if the associated Gibbs distribution leads to reasonable posterior estimates, i.e., the parameter is estimated with the distributional aspect in mind. Here we use MAP to refer to finding the minimum energy labeling regardless of whether (2) accurately describes the posterior.

3 Related work

Early work on learning the potentials for low-level vision from data dates back to the mid '90s [22]. With the advance of structured models in machine learning, more sophisticated techniques for estimating the parameters have also evolved in computer vision. [7] trains a CRF for the tasks of binary image denoising and the detection of man-made structures. More recently, principled discriminative training has gained popularity in high-level vision applications, such as semantic segmentation [11] and object recognition [4]. In the context of low-level vision problems, learning has been done in stereo vision [17] and image denoising. In denoising, the application considered in our work, the Fields-of-Experts (FoEs) model [13] is a popular continuous, generative model with higher-order factors (e.g., of size 3×3). In the original work, Roth and Black train the model using contrastive divergence, an approximate maximum likelihood learning approach, and finally perform MAP inference at test time. Better results can be obtained [16] by a training approach tailored towards the MAP prediction. Finally, [18] demonstrates improved accuracy when using contrastive divergence learning and MMSE instead of MAP inference. They find that their predictions better match the image statistics observed in natural images.

Our work sheds some light on these findings [18] and shows that MAP, while inferior to MMSE in theory for an ideal setting, in practice can still outperform MMSE. This is attributed to the fact that models are often misspecified and approximate maximum likelihood approaches, such as maximum pseudo-likelihood, lead to inaccurate parameter estimates. Experiments are shown for the pairwise model in § 2.1 which differs in several aspects to the FoE model. First, unlike the FoE model, it is a discrete model. This allows us to learn the full shape of the potential without any prior assumptions on the form. In contrast, such assumptions are needed in the FoE model as it is a continuous model whose potentials are functions parametrized by a small set of shape parameters. Second, maximum likelihood and maximum margin training for our model is convex, this is not the case for the FoE due to modeling assumptions. The convexity has the advantage that our learning approach does not get stuck in local minima.

4 Optimal Prediction

Assuming that one is given the true posterior $P(\boldsymbol{y}|\boldsymbol{x})$ (note in particular that we distinguish this from a model posterior $P(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w}^*)$) we now consider the prediction task. In this context the loss $\Delta(\boldsymbol{y}', \boldsymbol{y})$ specifies the error/loss incurred when predicting the label $\boldsymbol{y}'$ if $\boldsymbol{y}$ would be the true label. The loss is application dependent and can be thought of as the error measure used in many computer vision benchmarks: For semantic image segmentation this might be given by the pixelwise accuracy, whereas for image denoising the pixelwise squared distance of prediction and ground-truth might be used. According to Bayesian decision theory [12, Theorem 2.3.2] the optimal prediction minimizes the expected risk:

$$\boldsymbol{y}^* = \operatorname*{argmin}_{\boldsymbol{y}'} \mathbb{E}_{P(\boldsymbol{y}|\boldsymbol{x})}[\Delta(\boldsymbol{y}', \boldsymbol{y})] = \operatorname*{argmin}_{\boldsymbol{y}'} \sum_{\boldsymbol{y}} \Delta(\boldsymbol{y}', \boldsymbol{y}) P(\boldsymbol{y}|\boldsymbol{x}). \tag{4}$$

Next, we relate several prediction functions to their implied loss function. The loss is assumed to be non-negative and zero for the ground-truth labeling.

Zero-one error The zero-one error is given by $\Delta(\boldsymbol{y}', \boldsymbol{y}) = 1 - \delta_{\boldsymbol{y}}(\boldsymbol{y}')$. Here we extend the Kronecker delta function to several variables. This loss treats all labels $\boldsymbol{y}'$ with $\boldsymbol{y}' \neq \boldsymbol{y}$ in the same way by assigning a loss of one to them. A labeling of an image with only one pixel different from the ground-truth is assigned the same loss as a label that is different in every pixel. If the zero-one loss is used in (4), then one identifies the MAP prediction rule $\boldsymbol{y}^* = \operatorname{argmax}_{\boldsymbol{y}} P(\boldsymbol{y}|\boldsymbol{x})$. As most evaluation metrics are not as aggressive as the zero-one error discussed here, it is clear that this is not the best loss term for most labeling tasks.

Mean pixel-wise error The mean pixelwise error is given by $\Delta(\boldsymbol{y}', \boldsymbol{y}) = \frac{1}{|\mathcal{V}|} \sum_{i \in \mathcal{V}} (1 - \delta_{y_i}(y_i'))$. When inserting this loss into the Bayes predictor we end up with the max-marginal prediction rule $y_i^* = \operatorname{argmax}_{y_i} P(y_i|\boldsymbol{x})\ \forall i \in \mathcal{V}$. Here $P(y_i|\boldsymbol{x})$ denotes the marginal for the i-th pixel.

Mean squared error The mean squared error (MSE) $\Delta(\boldsymbol{y}', \boldsymbol{y}) = \frac{1}{\mathcal{V}} \sum_{i \in \mathcal{V}} (y_i - y_i')^2$ is a sensible choice if there exists an order on the labels, as for example in image denoising. Optimal prediction is achieved by $y_i^* = \mathbb{E}_{P(y_i|\boldsymbol{x})}[y_i]\ \forall i \in \mathcal{V}$. Thus, taking the mean of the individual variable posterior distribution minimizes the mean squared error. This predictor is referred to as minimum mean squared error (MMSE). For discrete variables one can round the expectation.

The underlying assumption in this section was that the true posterior distribution $P(\boldsymbol{y}|\boldsymbol{x})$ is known. In practice this posterior is modeled by the CRF distribution $P(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w}^*)$ which in many scenarios in computer vision does not accurately model the true posterior. There might exist several reasons for this: First, not enough data might be available to estimate all the parameters accurately. Second, an improper estimation technique could be used for $\boldsymbol{w}^*$. Third, the model might not model all the dependencies in the data. As we will show, if a model posterior distribution $P(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w}^*)$ does not match the true $P(\boldsymbol{y}|\boldsymbol{x})$, optimality of the schemes above is no longer guaranteed.

5 Learning

In this section we consider learning the optimal parameters $\boldsymbol{w}^*$ of a structured model for a given training set $\mathcal{D} = \{(\boldsymbol{x}^n, \boldsymbol{y}^n)\}_{n=1}^N$. We focus on maximum likelihood estimation (MLE) and maximum margin (MM) learning. As MLE is generally intractable we also consider the maximum pseudo-likelihood.

5.1 Maximum Likelihood and Maximum Pseudo-likelihood

MLE of the parameters for a given training set corresponds to finding the parameter with the largest likelihood given the observed data. To prevent overfitting, an L_2 regularizer is often included:

$$\boldsymbol{w}^{mle} = \operatorname*{argmin}_{\boldsymbol{w}} -\frac{1}{N} \sum_{n=1}^{N} \log P(\boldsymbol{y}^n|\boldsymbol{x}^n, \boldsymbol{w}) + \frac{\lambda}{2} \|\boldsymbol{w}\|^2. \tag{5}$$

In general, no closed form solution for the convex MLE objective exists and thus iterative methods are employed. To evaluate the function value and the gradient, the partition sum and the marginals need to be computed. For loopy graphs

these computations are generally intractable and one resorts to approximations. A tractable alternative is given by the maximum pseudo-likelihood estimate [1] (MPLE) which replaces $\log P(\boldsymbol{y}^n|\boldsymbol{x}^n, \boldsymbol{w})$ by $\sum_{i\in\mathcal{V}} \log P(y_i^n|\boldsymbol{y}^n_{\mathcal{N}(i)}, \boldsymbol{x}^n, \boldsymbol{w})$. Here $\mathcal{N}(i)$ denotes the Markov blanket of a variable i and $\boldsymbol{y}_{\mathcal{N}(i)}$ all the variables in the Markov blanket. Conditioning on the ground-truth label of the neighboring variables makes the partition sum collapse to a sum over the different states of variable y_i, which has linear complexity. Interestingly, the MPLE has the desirable property that for enough data it converges to the MLE.

5.2 Maximum Margin

Instead of taking the detour of first estimating a posterior and subsequently constructing a predictor by incorporating a loss, one can directly train a linear predictor $f_{\boldsymbol{w}}(\boldsymbol{x}) = \operatorname{argmax}_{\boldsymbol{y}} \langle \boldsymbol{w}, \boldsymbol{s}(\boldsymbol{x}, \boldsymbol{y})\rangle$. This predictor can be trained using a particular loss function $\Delta(\boldsymbol{y}', \boldsymbol{y})$. Max-margin training (or equivalently the structured SVM) [19] considers the following training objective

$$\boldsymbol{w}^{mm} = \operatorname*{argmin}_{\boldsymbol{w}} \frac{\lambda}{2}\|\boldsymbol{w}\|^2 + \frac{1}{N}\sum_{n=1}^{N} \max_{\boldsymbol{y}'} \left[\langle \boldsymbol{w}, \boldsymbol{s}(\boldsymbol{x}^n, \boldsymbol{y}') - \boldsymbol{s}(\boldsymbol{x}^n, \boldsymbol{y}^n)\rangle + \Delta(\boldsymbol{y}', \boldsymbol{y}^n)\right]. \tag{6}$$

For computer vision max-margin training has several advantages when compared to MLE. First, inference reduces to a standard MAP problem, and thus neither marginals nor the partition sum need to be computed. Second, it directly incorporates a loss in training and is expected to work well for this particular loss, even if the model is not expressive enough. However, for the ideal setting, the Bayes predictor in (4) is superior to MAP trained using max-margin, as it is more expressive. Most training algorithms for max-margin work by successively generating maximally violated constraints and repetitively solving the quadratic programming problem in (6). Generation of the constraints reduces to the MAP problem for the loss augmented model which incorporates the loss $\Delta(\boldsymbol{y}', \boldsymbol{y}^n)$.

5.3 Insights on statistic matching

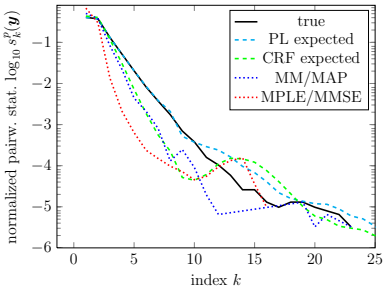

Fig. 2: Pairwise image statistics on logarithmic scale, see text for remarks.

It is widely known that the image statistics of natural images have a heavy tailed distribution [15]. This is conjectured to be an important property that most computer vision systems still fail to model. The image statistic of an image is obtained by applying linear filters to the image and building histograms of the resulting responses. For a pairwise gradient filter the histogram obtained is equivalent to the sufficient statistics $\boldsymbol{s}^p(\boldsymbol{y})$ of our pairwise image denoising model. For the task of image denoising, [18] observes that the MAP prediction of the FoE model trained using maximum likelihood, exhibit poor image statistics. The authors propose MMSE prediction as an alternative resulting in better image statistics. The discussion in § 4 shows that the *superior performance of MMSE can be explained by the loss being more suitable for the image denoising task*. The improved image statistics come only as a byproduct. MMSE in itself is not better at reproducing

natural image statistics. If predictions should explicitly show the heavy tails observed in natural image statistics, then this property has to be either included in the model as in [21], or in the prediction function using an appropriate loss. If no regularization is included in the objective in (5) then MLE can be understood as matching the empirical distribution in training by the expected sufficient statistics under the model distribution $P(\boldsymbol{y}|\boldsymbol{x}^n, \boldsymbol{w}^{mle})$:

$$\frac{1}{N}\sum_{n=1}^{N} \mathbb{E}_{P(\boldsymbol{y}|\boldsymbol{x}^n, \boldsymbol{w}^{mle})}[\boldsymbol{s}(\boldsymbol{x}^n, \boldsymbol{y})] = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{s}(\boldsymbol{x}^n, \boldsymbol{y}^n).$$

This follows from the derivative of (5). A similar expectation matching is identified for MPLE. However, this *does not guarantee that the sufficient statistics of the predicted labelings also match the observed training image statistics.* This behaviour is demonstrated in Fig. 2 for the image denoising application described in more detail in § 6.2. Here we train on one image (32 gray levels) and predict a labeling for the same image. The expected statistics using the simplified model assumed by pseudo-likelihood $P(\boldsymbol{y}|\boldsymbol{y}^1, \boldsymbol{x}^1, \boldsymbol{w}^{mple}) = \prod_i P(y_i|\boldsymbol{y}^1_{\mathcal{N}i}, \boldsymbol{x}^1, \boldsymbol{w}^{mple})$ (shown as 'PL expected'), are very close to the ground truth statistics (shown as 'true'). Smaller inaccuries are due to sampling. The expected statistics of the CRF model $P(\boldsymbol{y}|\boldsymbol{x}^1, \boldsymbol{w}^{mple})$ (shown as 'CRF expected'), would coincide with the true statistics if exact MLE could be performed. This also illustrates the deficiencies of the pseudo-likelihood approximation for a small dataset. Neither the labeling predicted by MAP trained using max-margin (shown as 'MM/MAP'), nor the labeling predicted by MMSE learned using maximum likelihood (shown as 'MPLE/MMSE'), agree with the true statistics. Expected statistics are obtained using Gibbs sampling of labelings $\boldsymbol{y}$ and averaging $\boldsymbol{s}(\boldsymbol{x}^1, \boldsymbol{y})$ over the sampled $\boldsymbol{y}$.

6 Experiments

In this section we demonstrate the practical implications of the concepts discussed for the simple pairwise CRF model in § 2.1.

6.1 Synthetic data

Here we study the properties of maximum pseudo-likelihood estimation and max-margin learning on synthetic data. The synthetic nature of the dataset allows us to study the consistency property of the MPLE for large datasets. In this experiment we add structured noise to the labels to simulate a case where the model is misspecified, i.e., the relationship between $\boldsymbol{x}$ and $\boldsymbol{y}$ cannot be captured by the assumed posterior $P(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w})$. The dataset $\mathcal{D} = \{(\boldsymbol{x}^n, \boldsymbol{y}^n)\}_{n=1}^N$ is generated as follows: For a given image $\boldsymbol{x}^n$, a label $\boldsymbol{y}^n$ is sampled according to $\boldsymbol{y}^n \sim P(\boldsymbol{y}|\boldsymbol{x}^n, \boldsymbol{w}^{true})$ using a Gibbs sampler.

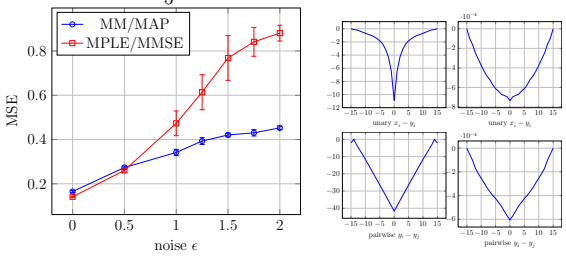

Fig. 3: Results of the synthetic experiment. Left: for increasing levels of misspecification MPLE trained MMSE becomes worse than max-margin trained MAP. Right: Learned potentials for $\epsilon = 2$ (MPLE left, MM right, unary top, pairwise bottom).

The image $\boldsymbol{x}^n$ itself is generated by adding i.i.d. noise to a fixed image $\boldsymbol{x}^0$ and rounding the values to integers within the domain $\{0, \ldots, K-1\}$, here for $K = 16$. For the weights $\boldsymbol{w}^{true}$ we assume $\boldsymbol{w}^{true,u} = -K/[1, 2, \ldots, K]^\mathsf{T}$ and $\boldsymbol{w}^{true,p} = -3 \cdot [0, 2, \ldots, K-1]^\mathsf{T}$. To study the influence of misspecification the labels are perturbed. This is an important scenario to study, as most computer vision models are still far from accurately describing the real world situation. Having a parameter estimation and prediction approach that is robust to misspecification is thus important in practice. To simulate the misspecification, the labels $\boldsymbol{y}^n$ are not sampled from the model in (3), but rather from a model which also includes a 4-neighborhood dependency to the pixel two pixels away (left, right, up, down). The weights of these interactions are chosen to be $\boldsymbol{w}^{true,p,long} = -3\epsilon \cdot [0, 1, \ldots, K-1]^\mathsf{T}$. Parameter estimation is done using the dataset $\mathcal{D}_\epsilon = \{(\boldsymbol{x}^n, \boldsymbol{y}^n)\}_{n=1}^N$ for the model in (3). As ϵ is increased, the model does not match the true data generating posterior anymore. To evaluate the methods we learn a parameter and report the MSE of predictions on held out test data. The results are averaged over five datasets. As we are primarily interested in the sensitivity of the estimation techniques to model misspecification, a relatively large training set of size $N = 500$ is used. In Fig. 3 the MSE of the different methods is shown.

The max-margin learning combined with MAP prediction leads to smaller MSE values than MPLE based learning with MMSE. This is in agreement with our intuition: max-margin learning directly considers the prediction function and should therefore be more robust to misspecifications. In the non-misspecified setting likelihood based learning combined with MMSE inference performed better. While the experiment was carried out using pseudo-likelihood, we conjecture that the same problem is also present in maximum likelihood estimation as we also performed an experiment that showed that MPLE converged for $\epsilon = 0$.

6.2 Image denoising

We consider the real world task of image denoising, an active field of research. The state of the art methods can broadly be grouped into modifications of the Fields-of-Experts framework [18,16,13] and sparse coding approaches [9,3]. The image denoising experiment was performed on the images from the Berkeley image segmentation dataset [10]. The same train/test set split as in [18] was used. The images are reduced to grayscale values and i.i.d. Gaussian noise with $\sigma = 25$ is added. The resulting pixel values are rounded to integers in $\{0, \ldots, 255\}$. Furthermore, the image and the noisy version thereof are further discretized to 64 labels to obtain the label $\boldsymbol{y}$ and the input image $\boldsymbol{x}$.

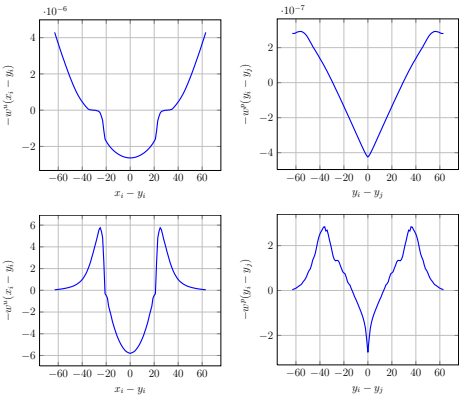

Fig. 4: Learned unary (left) and pairwise (right) potentials. Top: result for max-margin learning. Bottom: weights estimated by maximum pseudo-likelihood.

noisy image,20.29dB original image MM/MAP,26.03dB MPLE/MMSE,25.44dB MPLE/MAP,23.11dB

Fig. 5: MAP can outperform MMSE if trained with maximum margin. In the (cropped) image above we observe that MM/MAP better preserves the fine structure on the rock. MAP prediction with the MPLE estimate leads to substantially worse results.

Maximum margin and MPLE training are performed on the 40 training examples. The resulting learned weights are shown in Fig. 4. We trained on the full images as opposed to only on smaller subpatches, as it is often done for contrastive divergence. We observe that the learned weights for the MM learning are much smaller. The pairwise potential is almost linear and the unary potential has a roughly quadratic shape with truncation areas. The potentials trained by MPLE differ substantially and show a much more varying shape. As it is standard for image denoising problems we use the peak signal-to-noise ratio (PSNR) for comparison of the different methods. The test set consisted of 68 images. Comparing the results of the different approaches in Table 1, we see that max-margin training combined with MAP prediction leads to a lower MSE and PSNR than maximum pseudo-likelihood estimation followed by MMSE prediction. For comparison we also show the results obtained using the BM3D algorithm [3], considered state-of-the-art. For BM3D we used the full 256 level grayscale images and discretized the result to 64 levels. We also trained our pairwise model with BM3D predictions as a secondary unary feature. The MAP labeling obtained using MM training result in a small improvement over BM3D.

	standard model		with BM3D	
method	MSE	PSNR	MSE	PSNR
MM/MAP	**8.65**	**27.05**	**6.86**	**28.23**
MPLE/MAP	17.42	24.30	13.31	25.5
MPLE/MMSE	10.04	26.65	8.47	27.54
BM3D only	-	-	6.95	28.19

Table 1: Image denoising test results of the different methods. For MSE smaller is better, for PSNR higher is better.

Unlike in the synthetic experiment, we can not give a final conclusion on why MMSE performs worse: it could be either the inaccurate approximation made by pseudo-likelihood or as the model is simplistic, that misspecification becomes a problem as in the synthetic experiment. However, the image denoising experiment shows that in practice if trained appropriately, MAP can lead to accurate predictions on par with MMSE. *Unless the full image zero-one loss is desired as an evaluation criteria, MAP should not be used in combination with maximum likelihood learning.* We visualize the test set image statistics in Fig. 6. One observes that for the pairwise statistics the MM/MAP predictions show a very similar behavior as the MPLE/MMSE solutions. MM/MAP

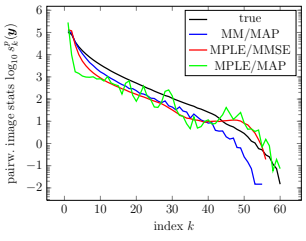

Fig. 6: Aggregated pairwise image statistics on the test images. MPLE/MMSE and MM/MAP result in similar statistics.

seems to be a bit closer to the true image statistics for the more often occurring configurations. An example of the predictions is shown in Fig. 5.

7 Conclusions

This paper gives a general review of learning and inference for structured models. For image denoising we found that if appropriately trained, MAP is competitive with MMSE, the optimal prediction in theory. We explain this by misspecifications of the model and the approximations needed in order for maximum likelihood learning to become tractable. MAP, with many efficient inference algorithms readily available, is therefore back on the road map of computer vision. Our investigations also show that there exist scenarios where MMSE can outperform MAP. As models become more accurate, these differences might get more pronounced in the future. However, we suspect that better approximate maximum likelihood approaches are needed for MMSE to substantially outperform MAP in practice.

Acknowledgments PP was supported in parts by the Swiss National Science Foundation (SNF) under grant number 200021-117946.

References

1. Besag, J.: Statistical analysis of non-lattice data. The Statistician (1975)
2. Boykov, Y.: Fast approximate energy minimization via graph cuts. PAMI (2001)
3. Dabov, K., Foi, A., Katkovnik, V., Egiazarian, K.: Image denoising by sparse 3-D transform-domain collaborative filtering. IEEE TIP 16(8), 2080–95 (2007)
4. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. PAMI 32(9), 1627–45 (2010)
5. Koller, D., Friedman, N.: Probabilistic Graphical Models. MIT Press (2009)
6. Kolmogorov, V.: Convergent tree-reweighted message passing for energy minimization. PAMI 28(10) (2006)
7. Kumar, S., Hebert, M.: Discriminative Random Fields. IJCV 68(2), 179–201 (2006)
8. Ladicky, L., Russell, C., Kohli, P., Torr, P.H.S.: Associative hierarchical CRFs for object class image segmentation. In: ICCV (2009)
9. Mairal, J., Bach, F., Ponce, J., Sapiro, G.: Online Learning for Matrix Factorization and Sparse Coding. JMLR 11, 19–60 (2010)
10. Martin, D., Fowlkes, C., Tal, D., Malik, J.: A database of human segmented natural images. In: ICCV (2001)
11. Nowozin, S., Gehler, P.V., Lampert, C.H.: On Parameter Learning in CRF-based Approaches to Object Class Image Segmentation. In: ECCV (2010)
12. Robert, C.P.: The Bayesian Choice. From decision Theoretic Foundations to Computational Implementation. Springer (2001)
13. Roth, S., Black, M.J.: Fields of Experts. IJCV (2008)
14. Rother, C., Kolomogorov, V., Blake, A.: GrabCut Interactive Foreground Extraction using Iterated Graph Cuts. TOG 23(3), 309–314 (2004)
15. Ruderman, D.: The statistics of natural images. Comp. in Neural Systems (1994)
16. Samuel, K.G.G., Tappen, M.F.: Learning Optimized MAP Estimates in Continuously-Valued MRF Models. In: CVPR (2009)
17. Scharstein, D.: Learning Conditional Random Fields for Stereo. In: CVPR (2007)
18. Schmidt, U., Gao, Q., Roth, S.: A Generative Perspective on MRFs in Low-Level Vision. In: CVPR (2010)
19. Taskar, Guestrin, Koller: Max-Margin Markov Networks. In: NIPS (2003)
20. White, H.: Maximum-likelihood estimation of misspecified models. Econom. (1982)
21. Woodford, O.J., Rother, C., Kolmogorov, V.: A Global Perspective on MAP Inference for Low-Level Vision. In: ICCV (2009)
22. Zhu, S., Mumford, D.: Prior learning and Gibbs reaction-diffusion. PAMI (1997)