Semi-Supervised Learning via Generalized Maximum Entropy

Ayşe Naz Erkan
Max Planck Institute for
Biological Cybernetics, Tübingen, Germany.
New York University, New York, NY, USA

Yasemin Altun
Max Planck Institute for
Biological Cybernetics, Tübingen, Germany

Abstract

Various supervised inference methods can be analyzed as convex duals of the generalized maximum entropy (MaxEnt) framework. Generalized MaxEnt aims to find a distribution that maximizes an entropy function while respecting prior information represented as potential functions in miscellaneous forms of constraints and/or penalties. We extend this framework to semi-supervised learning by incorporating unlabeled data via modifications to these potential functions reflecting structural assumptions on the data geometry. The proposed approach leads to a family of discriminative semi-supervised algorithms, that are convex, scalable, inherently multi-class, easy to implement, and that can be kernelized naturally. Experimental evaluation of special cases shows the competitiveness of our methodology.

1 Introduction

The scarcity of labeled training samples in many applications ranging from natural language processing to bio-informatics has motivated the research on semi-supervised learning algorithms that exploit unlabeled data. A variety of methods, e.g., (Chapelle et al., 2006) and references therein, have been proposed for semi-supervised learning. The intuition behind many of the semi-supervised learning algorithms is that the outputs should be smooth with respect to the structure of the data, i.e., the labels of two inputs that are similar with respect to the intrinsic geometry of data are likely to be the same. This idea is often further specified via the *cluster assumption*, the *manifold assumption* or the *semi-supervised smoothness assumption*. This paper presents a novel semi-supervised approach by using unlabeled data to impose smoothness criteria in the generalized maximum entropy framework.

Maximum entropy framework has been studied extensively in the supervised setting. Here, the goal is to find a distribution p that maximizes an entropy function and satisfies data constraints that enforce the expected values of some (pre-defined) features with respect to p to match their empirical counterparts approximately. Using different entropy measures, different model spaces for p and different approximation criteria for data constraints, we obtain a family of supervised learning methods (e.g., logistic regression, least squares and boosting) via convex duality techniques (Altun & Smola, 2006; Dudík & Schapire, 2006; Friedlander & Gupta, 2006). This framework is known as the *generalized maximum entropy framework*.

We propose integrating unlabeled data to the entropy maximization problem via additional penalty functions that restrict the model outputs to be consistent within local regions. We investigate two types of penalty functions. *Pairwise penalties* aim to minimize the discrepancy of the conditional class distributions for each sample pair with respect to their proximity. *Expectation penalties* are a relaxed variant of the former, where the conditional output distribution of an instance is enforced to match the weighted average of the conditional distribution over local regions. The proximity of two samples is defined according to a similarity function that reflects our prior knowledge on the geometry of the data. Augmenting the primal maximum entropy problem and applying convex duality techniques yields convex semi-supervised objective functions, which we refer as the dual problems. In this paper we describe two special cases, namely semi-supervised logistic regression and kernel logistic regression, in detail.

Our approach offers a number of advantages over pre-

Appearing in Proceedings of the 13^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2010, Chia Laguna Resort, Sardinia, Italy. Volume 9 of JMLR: W&CP 9. Copyright 2010 by the authors.

vious methods. First, by using different entropy measures, we obtain a family of semi-supervised algorithms. Second, these algorithms can be kernelized allowing the model to exploit unlabeled data in a nonlinear manner as opposed to other information theoretic semi-supervised learning methods such as (Grandvalet & Bengio, 2005; Mann & McCallum, 2007). The resulting objective functions are convex since the unlabeled data is incorporated in the primal MaxEnt problem and the objective functions are then derived using convex duality techniques. Another key advantage is that our method is inherently multi-class. This is often not the case for discriminative semi-supervised classifiers, e.g., Transductive Support Vector Machines (TSVMs), as in multi-class settings they require further elaboration in inference such as the one-vs-rest error assessment scheme. Finally, even though our motivation is similar to other SSL methods that are based on the smoothness criterion, the resulting formulation is substantially different as it enables the algorithm to choose which similarities are salient unlike many SSL algorithms that treat similarities uniformly such as Laplacian SVMs. This is a significant advantage of encoding the similarities in the primal problem via features, as opposed to encoding them within a regularization term in the dual.

The rest of the paper is organized as follows: In Section 2, we overview the generalized Maximum Entropy framework. Section 3 provides the details of our approach. In Section 4, we give a summary of related work. An experimental evaluation of these algorithms on benchmark data sets is presented in Section 5. Comparison to a large number of semi-supervised learning methods shows that our method performs competitively.

2 Duality of Maximum Entropy and Supervised Learning Methods

In this section, we outline a brief summary of the duality relation between generalized Maximum Entropy and various supervised learning methods[1]. We focus on conditional distributions $\mathcal{P} = \{p \,|\, p(y|x) \geq 0, \quad \sum_{y \in \mathcal{Y}} p(y|x) = 1, \quad \forall x \in \mathcal{X}, y \in \mathcal{Y}\}$, where $\mathcal{Y}$ and $\mathcal{X}$ are output and input spaces respectively. The goal in generalized MaxEnt is to minimize the divergence of the target distribution p from a reference distribution while penalizing the discrepancy between observed values $\tilde{\psi}$ of some pre-defined *model* feature functions $\psi : \mathcal{X} \times \mathcal{Y} \to \mathcal{B}$ and their expected values with respect to the target distribution. Here, $\tilde{\psi}$ can be derived from a sample, e.g., $\tilde{\psi} = 1/n \sum_{i=1}^{n} \psi(x_i, y_i)$. The conditional expectation is defined as

$$\mathbb{E}_p[\psi] = \sum_x \tilde{p}(x) \mathbb{E}_{p_x}[\psi] = \sum_x \tilde{p}(x) \mathbb{E}_{y \sim p(.|x)}[\psi(x, y)], \quad (1)$$

where $\tilde{p}$ denotes the empirical marginal distribution over the input space.

When the target distribution is defined on a finite dimensional space with differentiable discrepancy functions over finite dimensional spaces, the maximum entropy problem can be solved using Lagrangian techniques. However, in the generalized MaxEnt framework with non-differentiable penalty functions as outlined in (Dudík & Schapire, 2006) or with infinite dimensional spaces as (Altun & Smola, 2006) points out, a more general duality technique such as Fenchel's duality is required for a proper analysis of the primal-dual space relations. Here we briefly introduce key concepts required for the rest of this paper. For a detailed reference on convex analysis the reader may refer to (Borwein & Zhu, 2005).

Let $\mathcal{B}$ be a Banach space and $\mathcal{B}^*$ be its dual. The **convex conjugate** of a function $h : \mathcal{B} \to \Re$ is $h^* : \mathcal{B}^* \to \Re$ where h^* is defined as

$$h^*(b^*) = \sup_{b \in \mathcal{B}} \{\langle b, b^* \rangle - h(b)\}.$$

Examples of convex conjugacy used in this paper are KL divergence, approximate norm constraints and norm-square penalty functions: $h_1(b; a) = \int_t b(t) \ln b(t)/a(t)$, $h_1^*(b^*; a) = \int_t a(t) \exp(b^*(t) - 1)$; $h_2(b; a, \epsilon) = \mathbb{I}(\|b - a\|_{\mathcal{B}} \leq \epsilon)$, $h_2^*(b^*; a, \epsilon) = \epsilon \|b^*\|_{\mathcal{B}^*} + \langle b^*, a \rangle$; $h_3(b; a, \epsilon) = \|b - a\|_{\mathcal{B}}^2/(2\epsilon)$, $h_3^*(b^*; a, \epsilon) = \epsilon \|b^*\|_{\mathcal{B}^*}^2/2 + \langle b^*, a \rangle$. Here, $\mathbb{I}(a) = 0$ if a holds; and ∞ otherwise.

The following lemma shows the duality of generalized MaxEnt for conditional distributions and various discriminative supervised learning methods. The proof is given in Appendix A.

Lemma 1 (MaxEnt Duality for conditionals) *Let $p, q \in \mathcal{P}$ be conditional distributions and $\mathbb{D}$ be a divergence function that measures the discrepancy between two distributions,*

$$\mathbb{D}(p|q) = \sum_x \tilde{p}(x) \mathbb{D}_x \left(p_x | q_x\right). \quad (2)$$

Moreover, let $\psi : \mathcal{X} \times \mathcal{Y} \to \mathcal{B}$ be a feature map to a Banach space $\mathcal{B}$, g be a lower semi-continuous (lsc) convex function and $\mathbb{E}_p$ is the conditional expectation

[1] We use entropy maximization and divergence minimization interchangeably since they are equivalent up to a constant for a fixed reference distribution.

operator in (1). *Define*

$$t := \min_{p \in \mathcal{P}} \{\mathbb{D}(p|q) + g\left(\mathbb{E}_p[\psi]; \tilde{\psi}, \epsilon\right)\}, \qquad (3)$$

$$d := \max_{\lambda \in \mathcal{B}^*} \{-\sum_x \tilde{p}(x)\mathbb{D}_x^*(\langle \psi(x,.), \lambda\rangle\, ; q_x) \qquad (4)$$
$$- g^*(\lambda; \tilde{\psi}, \epsilon)\},$$

where q is a reference distribution (reflecting the prior knowledge for target distribution). Then, $d = t$.

Setting divergence function to KL, $\mathbb{D}_x(p_x|q_x) = \sum_y h_1(p(y|x); q(y|x))$, we can get ℓ_1 and ℓ_2^2 regularized logistic regression by defining g as h_2 and h_3 respectively. In the latter case, if $\mathcal{B}$ is a reproducing kernel Hilbert space (RKHS), we get kernel logistic regression. Other special cases can be found in (Altun & Smola, 2006).

3 Semi-Supervised via Generalized Maximum Entropy

In semi-supervised learning, we are given a sample S that consists of labeled data $L = \{(x_i, y_i)\}_{i=1}^l$ drawn i.i.d. from the probability distribution on $\mathcal{X} \times \mathcal{Y}$ and unlabeled data $U = \{x_i\}_{i=l+1}^n$ drawn i.i.d. from the marginal distribution $P_{\mathcal{X}}$. We focus on multi-class problems where $\mathcal{Y} = \{1, \ldots, C\}$. $S_x = \{x_i\}_{i=1}^n$ denotes the (labeled and unlabeled) observations in the sample.

If the optimal classification function is smooth with respect to $P_{\mathcal{X}}$, in the sense that the outputs of two similar input points x,x' are likely to be the same, one can utilize unlabeled data points to impose the predictive function to be smooth. Various approaches to enforce this smoothness assumption have lead to a large collection of semi-supervised learning methods. For example, Sindhwani et al., (Sindhwani et al., 2005) implement this assumption by adding a new regularizer $\sum_{x,x'} s(x,x') \sum_y (f(x,y) - f(x',y))^2$, to various objective functions where $f(x,y)$ is the predictive function and $s(x,x')$ is the similarity between the samples x, x'. With the same motivation, we extend the primal generalized MaxEnt problem to minimize the discrepancy between conditional probability distributions of similar instances. This yields new optimization methods favoring model outputs that are smooth with respect to the underlying marginal distribution.

Theoretically, the discrepancy function can be any convex proper lsc function. However, one should consider efficiency, feasibility and compatibility with the divergence function $\mathbb{D}$ when choosing the discrepancy function. For example, defining discrepancy as $\mathbb{I}\left(s(x,x')\|p_x - p_{x'}\| \leq \epsilon, \forall x, x'\right)$ may lead to infeasible solutions for small ϵ values or may render unlabeled data ineffective for large ϵ values. Adjusting ϵ for each x, x' pair, on the other hand, leads to a very large number of hyper-parameters rendering optimization intractable. Examining various combinations of the across-sample discrepancy functions and the divergence functions, we observed that ℓ_p, ℓ_p^2 norms of the differences are compatible with many divergence functions. We leave a more thorough analysis as future work.

3.1 Pairwise Penalties

One way of encoding the smoothness criteria is by augmenting the supervised MaxEnt problem (3) with a discrepancy for all similar x, x' pairs.

$$t_s := \min_{p \in \mathcal{P}} \{\mathbb{D}(p|q) + g\left(\mathbb{E}_p[\psi]; \tilde{\psi}, \epsilon\right) + \bar{g}(p)\}, \qquad (5)$$

where $\bar{g}(p) = \hat{h}(\sum_{x,x'} h(p_x, p_{x'}))$ for $h, \hat{h}$ such that $\bar{g}$ is lsc convex.

Corollary 2 *The dual of semi-supervised MaxEnt with pairwise similarities,* (5)*, is given by*

$$d_s := \max_{\lambda \in \mathcal{B}^*} \{-g^*(\lambda; \tilde{\psi}, \epsilon)$$
$$- \sum_x \tilde{p}(x)(\mathbb{D} + \bar{g})_x^*(\langle \psi(x,.), \lambda\rangle; q_x)\}. \qquad (6)$$

The equality of (5) *and* (6) *follows from Fenchel's duality and Lemma 1 by defining* $f_q(p) = \mathbb{D}(p|q) + \tilde{q}(p)$. *Note* $(\mathbb{D} + \bar{g})^* = \mathbb{D}^* \square \bar{g}^*$*, where* $\square$ *denotes the infimal convolution function. This term can be solved when* $\mathbb{D}$ *and g functions are specified.*

An interesting setting of t_s is when $g = \bar{g}$ is a norm. In this case, the difference between the conditionals can be written as a linear operator Φ which can then be combined with $\mathbb{E}_p[\psi]$ given in (1). Let $\Phi p = \sum_x \Phi_x p_x$ be the expectation operator over *similarity feature functions* ϕ,

$$\phi_{j,k,y}(x_i, y') = \begin{cases} s(x_i, x_k) & \text{if } i = j,\ j \neq k \text{ and } y = y', \\ -s(x_i, x_j) & \text{if } i = k,\ j \neq k \text{ and } y = y', \\ 0 & \text{otherwise,} \end{cases} \qquad (7)$$

for $j, k \in \{1, \ldots, n\}$. Then,

$$(\Phi p)_{i,j,y} = s(x_i, x_j)(p(y|x_i) - p(y|x_j)).$$

Concatenating Φp to $\mathbb{E}_p[\psi]$ and $\mathbf{0}$ vector (of size n^2C) to $\tilde{\psi}$, we get the dual of the semi-supervised MaxEnt as

$$d_s := \max_{\lambda, \gamma} \{-g^*((\lambda, \gamma); (\tilde{\psi}, \mathbf{0}), \epsilon) \qquad (8)$$
$$- \sum_x \tilde{p}(x)\mathbb{D}_x^*(\langle \psi(x,.), \lambda\rangle + \langle \phi(x,.), \gamma\rangle\, ; q_x)\}.$$

by Lemma 1.

The augmented MaxEnt (5) promotes target distributions that are smooth with respect to the similarity measure s in (7) and remains indifferent to distant instance pairs. s can be defined with respect to the geodesic distances in order to impose the manifold assumption, with respect to the ambient distances on high density regions in order to impose the smoothness assumption or with respect to data clusters in order to impose cluster assumption. We assume that $s(x, \tilde{x}) \geq 0, \forall x, \tilde{x} \in \mathcal{X}$.

Investigating the difference between the dual supervised and semi-supervised formulations, (4) and (8), we observe that $\mathbb{D}_x^*$ term is evaluated on both labeled and unlabeled data in the semi-supervised case, since the marginal distribution $\tilde{p}$ is now with respect to S_x. Furthermore, the expectation term $\mathbb{E}_{p_x}$ is evaluated on the similarity features ϕ as well as the original model features ψ. This results adding $n^2 C$ parameters to the optimization problem, where n is the total size of the data and C is the number of classes.

The increase in the number of parameters may be prohibitively expensive for very large data sets. One solution to this problem is to define a sparse similarity function. Then, the parameters for $\hat{x}$ and $\bar{x}$ becomes redundant if $s(\hat{x}, \bar{x}) = 0$. Hence, the number of parameters can be reduced significantly. In Section 3.2, we propose an alternative solution. We now present two special cases of (8), namely Pairwise Semi-Supervised Logistic Regression and Pairwise Semi-Supervised Kernel Logistic Regression.

3.1.1 Semi-Supervised Logistic Regression with Pairwise Penalty

The semi-supervised logistic regression with ℓ_2^2 regularization and pairwise semi-supervised penalty is given by setting the divergence function to KL, $\mathbb{D}_x = h_1$ with uniform q, and g to norm-square penalty function h_3,

$$\min_{p \in \mathcal{P}} \mathrm{KL}(p||q) + \|\tilde{\psi} - \mathbb{E}_p[\psi]\|_2^2 + \|\Phi p\|_2^2.$$

Note that

$$\|\Phi p\|_2^2 = \sum_{x,x'} \sum_y \left(s(x,x')(p(y|x) - p(y|x'))\right)^2.$$

Plugging the convex conjugates of the corresponding functions to (8) and negating it, we get the minimization problem of

$$\mathbf{Q}(\lambda, \gamma) = \sum_{x \in S_x} \log Z_x(\lambda; \gamma) - \left\langle \lambda, \tilde{\psi} \right\rangle + \epsilon\|\lambda\|_2^2 + \epsilon\|\gamma\|_2^2, \tag{9}$$

$$Z_x(\lambda, \gamma) = \sum_y \exp\left(F(x, y; \lambda, \gamma)\right),$$

$$\begin{aligned} F(x, y; \lambda, \gamma) = &\langle \lambda, \psi(x, y) \rangle + \sum_{\hat{x}} s(\hat{x}, x)\gamma_{\hat{x}xy} \\ &- \sum_{\bar{x}} s(x, \bar{x})\gamma_{x\bar{x}y} , \end{aligned} \tag{10}$$

where the relation between the primal parameter p and the dual parameters λ, γ is given by $p(y|x) = \exp(F(x, y))/Z_x$. Note that $\mathbf{Q}(\lambda, \gamma)$ is no longer the negative log-likelihood term. First, there is no inner product term on similarity parameters. Second, the log-partition function is computed for both labeled and unlabeled data. The similarity terms in F can be seen as a flow problem, where the weighted average of incoming flow from neighbors $s(\hat{x}, x)\gamma_{\hat{x}xy}$ is matched to the outgoing flow $s(x, \bar{x})\gamma_{x\bar{x}y}$.

It is important to note that $p(y|x)$ is well-defined for all x, hence it can be applied to out-of-sample data. From this perspective, this is a proper semi-supervised learning method. However, for out-of-sample data the similarity features are all 0. Hence, the penalty function remains ineffective for these instances. From this perspective, this is a transduction method since the performance is expected to improve from supervised to semi-supervised optimization only on the in-sample unlabeled data. As in other transductive methods, one can use interpolation techniques to improve the performance on the out-of-sample instances.

The gradients of the objective function with respect to the dual variables is given by

$$\begin{aligned} \frac{\partial \mathbf{Q}(\lambda, \gamma)}{\partial \lambda} =& \mathbb{E}_{p_x}[\psi(x, y)] - \tilde{\psi} + 2\epsilon\lambda , \\ \frac{\partial \mathbf{Q}(\lambda, \gamma)}{\partial \gamma_{\hat{x},\bar{x},y}} =& - p(y|\bar{x})s(\hat{x}, \bar{x}) + p(y|\hat{x})s(\hat{x}, \bar{x}) + 2\epsilon\gamma_{\hat{x},\bar{x},y} , \end{aligned}$$

Any gradient based optimization can be applied to find λ, γ that minimize $\mathbf{Q}$. In practice, we use the quasi-Newton method BFGS.

3.1.2 Semi-Supervised Kernel Logistic Regression with Pairwise Penalty

In addition to the assumptions of the previous section, if the domain of ψ is a Reproducing Kernel Hilbert Space $\mathcal{H}$ with kernel k, the dual problem gives the semi-supervised kernel logistic regression with pairwise penalty function. Semiparametric Representer Theorem (Schölkopf & Smola, 2001, Theorem 4.3) states that optimal λ in (9) admits the form

$$\lambda_y = \sum_{i=1}^{n} \alpha_{iy} k(x_i, x).$$

Plugging this λ into the formulas in the previous section gives the semi-supervised kernel logistic regression with pairwise penalties, which is convex and can be optimized with any gradient method.

3.2 Expectation Penalties

As mentioned earlier the number of parameters for pairwise penalties can get intractable with the increasing size of data. In order to reduce the number of parameters, we consider a relaxed version of the pairwise penalties. Here, instead of minimizing the discrepancy of conditional distributions across all x, x' pairs, we minimize the discrepancy of distributions over local regions. In particular, we impose minimization of various norms of the following discrepancy

$$\left(\sum_{\bar{x} \in S_x} (s(\hat{x}, \bar{x}) p(y|\hat{x}) - s(\hat{x}, \bar{x}) p(y|\bar{x})) \right), \tag{11}$$

over $(\hat{x}, y)$ pairs. This corresponds to imposing the conditional distribution of an instance $\hat{x}$ to be similar to the weighted average of the conditional distribution of instances within the vicinity of $\hat{x}$.

As in the case of pairwise penalties, we can express (11) in terms of a linear operator $\Phi p = \sum_x \Phi_x p_x$ over similarity feature functions ϕ given by

$$\phi_{k,y}(x_i, y') = \begin{cases} s(x_k, x_i) & \text{if } y = y' \text{ and } i \neq k, \\ -\sum_j s(x_j, x_i) & \text{if } y = y' \text{ and } i = k, \\ 0 & \text{otherwise.} \end{cases} \tag{12}$$

for $i \in \{1, \dots, n\}$. Then $(\Phi p)_{i,y}$ yields (11) for $\hat{x} = x_i$. We augment the primal MaxEnt problem with some norm of Φp.

This formulation requires at most nC additional parameters and hence yields smaller models than the semi-supervised approach with pairwise penalties. Furthermore, it can be more robust to conflicting (true but hidden) labels of similar samples.

The semi-supervised logistic regression with ℓ_2^2 regularization and expectation semi-supervised penalty is given by (9) with F defined as

$$F(x, y; \lambda, \gamma) = \langle \lambda, \psi(x, y) \rangle + \sum_{\hat{x}} s(\hat{x}, x) \gamma_{xy}$$
$$- \sum_{\bar{x}} s(x, \bar{x}) \gamma_{\bar{x}y}.$$

The gradients of γ is given by

$$\frac{\partial \mathbf{Q}(\lambda, \gamma)}{\partial \gamma_{xy}} = \sum_{\breve{x}} p(y|\breve{x}) s(\breve{x}, x) - \sum_{\hat{x}} p(y|x) s(\hat{x}, x).$$

The kernel version follows as in Section 3.1.2.

4 Related Work

Recently constraint driven semi-supervised approaches have attracted attention, (Bellare et al., 2009; Chang et al., ; Liang et al., 2009; Mann & McCallum, 2007). Chang et al. were one of the first to guide semi-supervised algorithms with constraints (Chang et al.,). Their model is trained via an EM like procedure with alternating steps. The authors impose constraints on the outputs y rather than the model distribution $p(y|x)$, as proposed in this paper. They also have a constraint violation mechanism where the hyper-parameters are manually set.

Bellare et al. impose expectation constraints on unlabeled data (Bellare et al., 2009). They define an auxiliary distribution that respects general convex constraints and has low divergence with the model distribution. The fundamental difference with our approach is that the authors impose the penalty functions on the dual objective of the MaxEnt framework. This in turn yields a non-convex optimization problem which is solved by alternating projections. In contrast, we impose constraints on the target distribution directly to the primal problem which yields convex loss functions.

Liang et al. propose *measurements* (Liang et al., 2009), a mechanism for partial supervision that unifies labels and constraints. A measurement is the expectation of a function over the outputs of the unlabeled samples. This approach allows fully-labeled examples, partially-labeled examples and general constraints on the model predictions such as label proportions to be treated similarly as these can be cast as instances of measurements. However, the loss function computations become intractable and approximate inference methods are required. Our approach shares the basic principle to enforce constraints on the predicted model distribution using Fenchel's duality and the maximum entropy framework. Yet, we use such constraints to integrate prior information about the geometry of the data over local regions using a similarity metric which can also be interpreted as matching predicted moments of similarity features. Moreover, we analyze the primal-dual relations of model features in RHKS along with similarity features.

Various other techniques with information theoretic justification have been previously proposed in the SSL literature. Information regularization (IR) (Grandvalet & Bengio, 2005) minimizes the conditional entropy of the label distribution predicted on unlabeled data, favoring minimal class overlap, along with the negative log-likelihood of the labeled data. Despite its high empirical performance, IR is criticized for its sensitivity to hyper-parameter tuning to balance the loss and regularization terms. Furthermore, if the labeled data is very scarce, IR tends to assign all unlabeled data with the same label. Expectation Regularization (ER) (Mann & McCallum, 2007) augments the negative conditional log-likelihood with a regularization term, enforcing the model expectation on features from unlabeled data to match either user-provided or empirically computed expectations. The authors provide experimental results for label features minimizing the KL divergence between the expected class distribution and the desired class proportions. In (Mann & McCallum, 2008) the same authors extend ER to CRFs for semi-supervised structured prediction. Note that both ER and IR algorithms use unlabeled data to regularize the log-likelihood, i.e., manipulate the dual objective of MaxEnt. We believe enforcing the data to match certain conditions directly in the primal is a more natural approach as it enables an easier interpretation and yields a convex optimization problem.

Using similarities to encode the data geometry is reminiscent of the similarity graphs used in label propagation methods and manifold methods (e.g.,(Sindhwani et al., 2005)). As in these methods, we use similarity graphs to impose smoothness, in the sense that similar inputs should have similar outputs. However, our approach of imposing this assumption in the primal MaxEnt problem leads to a feature representation of the similarities and allows the model to choose which similarity evaluations are useful via optimizing γ parameters, as opposed to the uniform treatment of similarities in previous SSL method. This renders our approach less sensitive to the choice of similarity function and yields good performance across many data sets as opposed to other SSL methods (see Section 5).

5 Experiments

5.1 Similarity Metric

For the empirical evaluation we use the following similarity definition

$$s(x_i, x_j) = \begin{cases} K(x_i, x_j) & \text{if } x_j \in N_{\kappa_{x_i}}, \\ 0 & \text{otherwise.} \end{cases} \tag{13}$$

where K is a Mercer kernel and $N_{\kappa_{x_i}}$ is the κ-nearest neighborhood of x_i with respect to K. Note that this similarity metric is sparse and non-symmetric.

Table 1: Properties of data sets. See (Chapelle et al., 2006; Chapelle & Zien, 2005) for more details. C: Number of classes, Dim: Data dimension.

	C	$\lvert U\rvert$	$\lvert L\rvert$	Dim.	Splits
Digit1	2	1400	100	241	12
COIL	6	1400	100	241	12
USPS$_2$	2	1400	100	241	12
USPS$_{10}$	10	1957	50	256	10
text	2	1856	50	7511	10
MNIST	10	5000	100/250	784	10

5.2 Data Sets

We present experiments on data sets that have been extensively analyzed in previous SSL work for fair and extensive comparison. We chose Digit1, USPS$_2$ and COIL data sets among the benchmarks data sets from (Chapelle et al., 2006), USPS$_{10}$ and text data sets from (Chapelle & Zien, 2005) and MNIST (LeCun et al., 1998). Table 1 summarizes essential properties of the data sets. For further details see (Chapelle et al., 2006; Chapelle & Zien, 2005).

5.3 Model Selection

The hyper-parameters of our algorithm are the neighborhood size κ in (13), the regularization constant ϵ_1 for the model feature parameters and ϵ_2 for the similarity feature parameters and finally the kernel bandwidth α in the case of a RBF kernel. We performed cross validation on a subset of labeled samples for model selection. From each data split we transferred 25% of the labeled samples to the corresponding unlabeled data split and found the model parameters that give the best average transduction performance on these samples only. In other words, *model selection is completely blind to the true labels of the unlabeled samples* in order to reflect the real-life scenario as closely as possible. We considered a range of hyper-parameters for model selection, $\kappa \in \{5, 15, 20, 30\}$ and $\epsilon_1, \epsilon_2 \in \{e^{-1}, e^{-2}, e^{-3}, e^{-4}\}$. We set $\alpha = \eta^{-2}$ where η is the median of pairwise distances. Subsequently, we retrained the algorithm with these parameters on the original set of labeled and unlabeled samples. In the following section, we report transduction error on the unlabeled samples averaged over all splits. Following previous work, we used cosine kernel, $K(\mathbf{x_i}, \mathbf{x_j}) = \langle \mathbf{x_i}, \mathbf{x_j} \rangle / \|\mathbf{x_i}\| \|\mathbf{x_j}\|$ for *text* and RBF kernel, $K(\mathbf{x_i}, \mathbf{x_j}) = \exp(-\alpha \|\mathbf{x_i} - \mathbf{x_j}\|^2)$ for all other data sets. In all experiments the same kernel is used for the kernel logistic regression (KLR) and the similarity metric.

5.4 Results

Tables 2, 3, 4 demonstrate our empirical evaluation. In Table 2, we report transduction error on Digit 1, USPS_2 and COIL data sets from (Chapelle et al., 2006) for Logistic Regression (LR) and Kernel Logistic Regression (KLR) both augmented with pairwise (PW) and expectation (EXP) penalties. All results are averages over all splits for the model parameters selected with cross validation as discussed previously. The first four lines correspond to the supervised methods, namely 1-nearest neighborhood (1-NN), Support Vector Machine (SVM), LR and KLR, where the algorithms are trained only on the labeled samples. At the bottom of the table, the performances of the most competitive semi-supervised algorithms reported in (Chapelle et al., 2006), namely Transductive SVM (TSVM) (Vapnik, 1998), Cluster Kernel (Chapelle et al.,), Discrete Regularization (Chapelle et al., 2006), Data Dependent Regularization (Chapelle et al., 2006), Low Density Separation (LDS) (Chapelle & Zien, 2005). The reader may refer to (Chapelle et al., 2006) for a comparison with a wider selection of algorithms.

A comparison of the results of our framework to supervised learning methods shows a consistent improvement for all data sets. This is not the case for many semi-supervised learning methods. We conjecture that this is due to the *feature* representation of the similarities, where the model can choose which similarity evaluations are useful. Regarding the relative performance with respect to other SSL methods, we observe that our approach is very competitive. In particular, it yields the best performance in Digit1 data set with 20% error reduction. For the other data sets, the method achieves the second and third best results. Interestingly the linear logistic regression algorithm is as good as the kernel logistic regression algorithm in most cases, indicating that using similarity features captures the non-linearities sufficiently. Investigating the differences between pairwise and expectation penalties, we observe that pairwise constraints are almost always more informative.

Table 3 reports the 10 class USPS data set and the text data. Performances of ∇TSVM, a variant of TSVM (Chapelle & Zien, 2005), Laplacian SVM (Sindhwani et al., 2005), LDS (Chapelle & Zien, 2005), Label Propagation (Zhu & Ghahramani, 2002), Transductive Neural Network (TNN) (Karlen et al., 2008) and Manifold Transductive Neural Network (Karlen et al., 2008) (ManTNN) algorithms are provided for comparison. The comparative analysis yields a similar pattern to Table 2. On text data, the performance of our approach is not as good as the most competitive methods reported for this data set.

Table 2: Transduction error on benchmark data sets averaged over all splits. Here we report only the most competitive results from previous work, for the full comparison table see the analysis of benchmarks chapter in (Chapelle et al., 2006). 1-NN: 1-nearest neighborhood.

	Digit1	USPS_2	COIL
1-NN	3.89	5.81	17.35
SVM	5.53	9.75	22.93
LR	7.31	12.83	35.17
KLR	6.02	9.20	24.63
LR+EXP	2.35	5.69	15.33
LR+PW	2.27	**5.18**	12.37
KLR+EXP	**1.94**	6.44	15.22
KLR+PW	2.26	5.54	**11.34**
TSVM	6.15	9.77	25.80
Discrete Reg.	2.77	**4.68**	**9.61**
Cluster-Kernel	3.79	9.68	21.99
Data-Dep. Reg.	**2.44**	5.10	11.46
LDS	3.46	4.96	13.72

Finally Table 4 reports the performance on a randomly chosen subset of the MNIST data set for LR with pairwise and expectation penalties. Here, we see that expectation penalties are preferable over pairwise penalties, which can be explained by the larger size of the unlabeled data set as opposed to other benchmark data sets. Hence, the empirical marginal distribution and its aggregate is more informative. The performance on USPS_{10}, COIL, MNIST data sets indicates that our algorithm can successfully handle multi-class problems.

6 Conclusions and Future Work

We presented a novel approach to integrate unlabeled data within the generalized maximum entropy framework through modifications to the potential functions. We demonstrated two such modifications, namely pairwise and expectation penalties on the MaxEnt objective. These penalties restrict the entropy maximization problem using the similarity relationships between data samples reflecting our prior knowledge.

Generalized MaxEnt framework encompasses a family of inference algorithms. We provided details for two special cases, logistic regression and kernel logistic regression for semi-supervised learning. Our approach can yield semi-supervised formulations of other instances of the MaxEnt framework such as Conditional Random Fields (CRFs) and kernel CRFs. Future work includes the development of these methods and further theoretical analysis.

Table 3: Transduction error averaged over all splits of USPS_{10} and text data sets. Supervised training error for single layer neural network and SVM and other semi-supervised methods have been provided for comparison. NN stands for neural network. Results of previous work obtained from (Karlen et al., 2008).

	USPS_{10}	text
SVM	23.18	18.86
NN	24.57	15.87
LR	26.07	15.64
KLR	28.81	15.70
LR+EXP	20.02	13.03
LR+PW	**14.96**	12.87
KLR+EXP	19.76	13.20
KLR+PW	16.15	**12.06**
∇TSVM	17.61	5.71
LapSVM	12.70	10.40
LDS	15.80	**5.10**
Label Propagation	21.30	11.71
TNN	16.06	6.11
ManTNN	**11.90**	5.34

Table 4: Transduction error on MNIST data set.

	$\lvert L\rvert = 100$	$\lvert L\rvert = 250$
LR	27.23	19.69
LR+EXP	21.21	12.78
LR+PW	24.01	13.53

References

Altun, Y., & Smola, A. J. (2006). Unifying divergence minimization and statistical inference via convex duality. *COLT* (pp. 139–153).

Bellare, K., Druck, G., & McCallum, A. (2009). Alternating projections for learning with expectation constraints. *UAI*.

Borwein, J. M., & Zhu, Q. (2005). *Techniques of variational analysis*. Springer.

Chang, M.-W., Ratinov, L., & Roth, D. *ACL* (pp. 280–287). Prague, Czech Republic.

Chapelle, O., Schölkopf, B., & Zien, A. (Eds.). (2006). *Semi-supervised learning*. Cambridge, MA: MIT Press.

Chapelle, O., Weston, J., & Schölkopf, B. Cluster kernels for semi-supervised learning. *NIPS* (pp. 585–592). MIT Press.

Chapelle, O., & Zien, A. (2005). Semi–supervised classification by low density separation. *AISTATS* (pp. 57–64).

Dudík, M., & Schapire, R. E. (2006). Maximum entropy distribution estimation with generalized regularization. *COLT* (pp. 123–138).

Friedlander, M. P., & Gupta, M. R. (2006). On minimizing distortion and relative entropy. *IEEE Transactions on Information Theory*, *52*, 238–245.

Grandvalet, Y., & Bengio, Y. (2005). Semi-supervised learning by entropy minimization. *NIPS* (pp. 529–536). Cambridge, MA: MIT Press.

Karlen, M., Weston, J., Erkan, A., & Collobert, R. (2008). Large scale manifold transduction. *ICML* (pp. 448–455).

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE* (pp. 2278–2324).

Liang, P., Jordan, M. I., & Klein, D. (2009). Learning from measurements in exponential families. *ICML* (pp. 641–648). Montreal, Quebec, Canada.

Mann, G. S., & McCallum, A. (2007). Simple, robust, scalable semi-supervised learning via expectation regularization. *ICML* (pp. 593–600). Corvalis, Oregon.

Mann, G. S., & McCallum, A. (2008). Generalized expectation criteria for semi-supervised learning of conditional random fields. *ACL*. Columbus, Ohio.

Schölkopf, B., & Smola, A. J. (2001). *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. The MIT Press.

Sindhwani, V., Niyogi, P., & Belkin, M. (2005). Beyond the point cloud: from transductive to semi-supervised learning. *ICML*.

Vapnik, V. (1998). *Statistical learning theory*. John Wiley & Sons.

Zhu, X., & Ghahramani, Z. (2002). *Learning from labeled and unlabeled data with label propagation* (Technical Report). Carnegie Mellon University.

A. Proof of Lemma 1

Proof Let $f_q(p) = \mathbb{D}(p|q)$, $A_x p_x = \mathbb{E}_{p_x}[\psi]$ and $Ap = \mathbb{E}_p[\psi]$. Fenchel's Duality (Borwein & Zhu, 2005, Theorem (4.4.3)) states that $\inf_{p\in\mathcal{P}}\{f_q(p) + g(Ap)\} = \sup_{\lambda\in\mathcal{B}^*}\{-f^*(A^*\lambda) - g^*(-\lambda)\}$ via strong duality. The adjoint transformation A^* is given by $\langle Ap, \lambda\rangle = \langle A^*\lambda, p\rangle$. For the expectation operator, $\left\langle \sum_x \tilde{p}(x) \sum_y p(y|x)\psi(x,y), \lambda \right\rangle = \sum_x \tilde{p}(x) \sum_y p(y|x) \langle \psi(x,y), \lambda\rangle = \sum_x \tilde{p}(x) \langle A_x^*\lambda, p_x\rangle$ for $A_x^*\lambda = \langle \lambda, \psi(x,.)\rangle$. Then, $f^*(A^*\lambda) = \sup_p\{\langle p, A^*\lambda\rangle - f(p)\} = \sup_{\{p_x\}}\{\sum_x \tilde{p}(x)\langle A_x p_x, \lambda\rangle - \sum_x \tilde{p}(x) f(p_x)\} = \sum_x \tilde{p}(x) \sup_{p_x}\{\langle A_x^* p_x, \lambda\rangle - f(p_x)\}$ for independent x. This is in turn equal to $\sum_x \tilde{p}(x) f^*(A_x^*\lambda)$. Plugging values to Fenchel's duality completes the proof. ■