SEQUENTIAL MAXIMUM MARGIN CLASSIFIERS FOR PARTIALLY LABELED DATA

Elizabeth Hou, Alfred O. Hero

University of Michigan
Dept. of Electrical Engineering and Computer Science
1301 Beal Avenue, Ann Arbor, MI 48109-2122

ABSTRACT

In many real-world applications, data is not collected as one batch, but sequentially over time, and often it is not possible or desirable to wait until the data is completely gathered before analyzing it. Thus, we propose a framework to sequentially update a maximum margin classifier by taking advantage of the Maximum Entropy Discrimination principle. Our maximum margin classifier allows for a kernel representation to represent large numbers of features and can also be regularized with respect to a smooth sub-manifold, allowing it to incorporate unlabeled observations. We compare the performance of our classifier to its non-sequential equivalents in both simulated and real datasets.

***Index Terms*—** semi-supervised classification, support vector machines, maximum entropy, maximum margin classifiers

1. INTRODUCTION

As the popularity of big data increases and more data is being gathered, the importance of sequential models that are able to continuously update with new data has increased. These models are particularly crucial in high throughput real-time applications such as speech or streaming text classification. To this end, we propose a sequential framework to update the probabilistic maximum margin classifier built from the Maximum Entropy Discrimination (MED) principle of [1].

The proposed sequential MED framework can be cast as recursive Bayesian estimation where the likelihood function is a log-linear model formed from a series of constraints and weighted by Lagrange multipliers. In the Gaussian case it shares similarities with the problem of Gaussian process classification, which has been previously studied [2, 3, 4, 5, 6, 7], but to the best of our knowledge, a method to recursively update the Gaussian process classifier has not been developed. In the single time point case, sequential MED can be specialized to the support vector machine [4] and Laplacian support vector machine [8] as previously discussed in [1] and [9].

We are interested in situations where we receive a stream of data $\boldsymbol{X}_{(1)}, \boldsymbol{X}_{(2)}, \ldots$ over time t where each $X_{(t)}$ is a matrix of dimension $p \times n$, with p denoting the number of feature variables and n denoting the number of i.i.d. samples, where $n = n_{(t)}$ may vary with time. In the fully labeled scenario, the data has corresponding labels $y_i = [1, -1]\, \forall i$ and t; however in the partially labeled scenario, at each time point t, only $l_{(t)} < n_{(t)}$ of the samples have labels. We define the observed data at any time point t as $\mathcal{D}_{(t)} = \{\boldsymbol{X}_{(t)}, \boldsymbol{y}_{(t)}\}$ and all observed data up to time τ as $\{\mathcal{D}_{(t)}\}_{t=1}^{\tau}$. Such scenarios would arise in a variety of domains such as a satellite that only transmits its data daily or a government agency that only releases its data quarterly with their corresponding reports. The rest of the paper is organized as follows: Section 2 and Section 3 will discuss how to sequentially update the corresponding MED models for supervised and semi-supervised classification. Section 4 validates the method by simulation and we present an application to a dataset of spoken letters of the English alphabet.

This work was partially supported by the Consortium for Verification Technology under Department of Energy National Nuclear Security Administration award number DE-NA0002534 and partially by the University of Michigan ECE Departmental Fellowship.

2. SEQUENTIAL MED

Constrained relative entropy minimization is used to estimate the closest distribution to a given prior distribution subject to a set of moment constraints. The authors of [10] show that, if the prior distribution is from the exponential family, then the density that optimizes the constrained relative entropy problem is also a member of the exponential family. Similar to Bayesian conjugate priors, there exist relative entropy conjugate priors that facilitate evaluation of the closest distribution. These produce optimal constrained relative entropy densities, which can be thought of as posteriors, from the same parametric family as the prior. Maximum entropy discrimination (MED) [1] also admits conjugate priors as it a special case of constrained relative entropy minimization where one of the constraints is over a parametric family of discriminant functions $\mathcal{L}(\boldsymbol{X}|\boldsymbol{\Theta})$.

2.1. Review of MED for Maximum Margin Classification

In this paper, we are interested in maximum margin binary classifiers. In this case the discriminant function $\mathcal{L}(\boldsymbol{X}|\boldsymbol{\theta}, b) = f(\boldsymbol{X})\boldsymbol{\theta} + b$ is linear for some feature transformation $f(\cdot)$, feature weights vector $\boldsymbol{\theta}$, and bias term b. Slack variables γ_i are used to create a margin in the constraints $\mathrm{E}(y_i(f(\boldsymbol{X}_i)\boldsymbol{\theta} + b) - \gamma_i)$, the expected hinge loss with slack variables. The MED objective function is

$$\min_{\mathrm{P}(\boldsymbol{\Theta},\boldsymbol{\gamma}|\mathcal{D})} \mathrm{KL}\left(\mathrm{P}(\boldsymbol{\Theta},\boldsymbol{\gamma}|\mathcal{D})||\mathrm{P}_0(\boldsymbol{\Theta},\boldsymbol{\gamma})\right) \qquad \text{subject to} \tag{1}$$

$$\iint \mathrm{P}(\boldsymbol{\Theta},\boldsymbol{\gamma}|\mathcal{D})\left(y_i(f(\boldsymbol{X}_i)\boldsymbol{\theta} + b) - \gamma_i\right) d\boldsymbol{\Theta} d\boldsymbol{\gamma} \geq 0 \;\forall i = 1, \ldots, n$$

whose solution $\mathrm{P}(\boldsymbol{\Theta}, \boldsymbol{\gamma}|\mathcal{D})$ is the constrained minimum relative entropy posterior. The associated MED decision rule $\hat{y}_{i'} = \mathrm{sgn}(\iint \mathrm{P}(\boldsymbol{\Theta}|\mathcal{D})(f(\boldsymbol{x}_{i'})\boldsymbol{\theta} + b)\, d\boldsymbol{\Theta})$ is a weighted combination of discriminant functions. The minimum relative entropy posterior has the form

$$\mathrm{P}(\boldsymbol{\Theta},\boldsymbol{\gamma}|\mathcal{D}) = \frac{\mathrm{P}_0(\boldsymbol{\Theta},\boldsymbol{\gamma})}{Z(\boldsymbol{\alpha})} \exp\left\{\sum_{i=1}^{n} \alpha_i \left(y_i(f(\boldsymbol{X})\boldsymbol{\theta} + b) - \gamma_i\right)\right\}$$

where $\boldsymbol{\alpha} = [\alpha_1, ..., \alpha_n]^T \geq 0$ are Lagrange multipliers that minimize the partition function $Z(\boldsymbol{\alpha})$. It is common to set the initial prior distribution to the separable form: $\mathrm{P}_0(\boldsymbol{\Theta}, \boldsymbol{\gamma}) = \mathrm{P}_0(\boldsymbol{\theta})\mathrm{P}_0(b)\prod_{i=1}^n \mathrm{P}_0(\gamma_i)$. If in addition, we specify that $\mathrm{P}_0(\gamma_i) = Ce^{-C(1-\gamma_i)}\mathcal{I}(\gamma_i \leq 1)$, $\mathrm{P}_0(\boldsymbol{\theta})$ is $N(\mathbf{0}, \mathbf{I})$, and $\mathrm{P}_0(b)$ is a zero mean Bayesian non-informative (diffuse) prior, denoted $N(0, \infty)$, then the Lagrange multipliers can be obtained as the solution $\hat{\boldsymbol{\alpha}}$ to the constrained optimization

$$\max_{\boldsymbol{\alpha}} -\frac{1}{2}\boldsymbol{\alpha}^T \boldsymbol{Y} f(\boldsymbol{X}) f(\boldsymbol{X})^T \boldsymbol{Y}\boldsymbol{\alpha} + \sum_{i=1}^n \alpha_i + \log(1 - \alpha_i/C)$$

$$\text{subject to } \sum_{i=1}^n y_i \alpha_i = 0 \text{ and } \alpha_1, \ldots, \alpha_n \geq 0$$

where $\boldsymbol{Y} = \mathrm{diag}(\boldsymbol{y})$. This objective function has a log barrier term $\log(1 - \alpha_i/C)$ instead of the inequality constraints $\alpha_i \leq C$ commonly found in the dual form of the SVM. Except in some ill-defined cases where the maximum lies near the boundary of the feasible set, the $\hat{\alpha}_i$ will be identical to the optimal support vectors that maximize the SVM objective. The authors in [1, 9] show that the *maximum a posteriori* (MAP) estimator for $\boldsymbol{\theta}$ of the MED posterior is related to the Lagrange multipliers by $\hat{\boldsymbol{\theta}} = f(\boldsymbol{X})^T\hat{\boldsymbol{\alpha}}$, so the MED posterior mode is equivalent to a maximum margin classifier.

2.2. Updating MED

Under the separable prior assumptions above, the MED posterior $\mathrm{P}(\boldsymbol{\Theta}, \boldsymbol{\gamma}|\mathcal{D})$ will take the factored form $\mathrm{P}(\boldsymbol{\theta}|\mathcal{D})\mathrm{P}(b|\mathcal{D})\mathrm{P}(\boldsymbol{\gamma})$. Due to the fact that the slack parameters γ_i do not depend on the data $\mathcal{D}$, the density $\mathrm{P}(\boldsymbol{\gamma})$ does not affect the MED decision rule given after (1). Hence only $\mathrm{P}(\boldsymbol{\theta}|\mathcal{D})$ and $\mathrm{P}(b|\mathcal{D}$ are important. This remaining part of the MED posterior has the form: $\mathrm{P}(\boldsymbol{\theta}|\mathcal{D})\mathrm{P}(b|\mathcal{D}) = N(f(\boldsymbol{X})^T\boldsymbol{Y}\boldsymbol{\alpha}, \mathbf{I})N(0, \infty)$, which is a conjugate distribution. Due to this conjugacy the posterior distribution optimizing the objective in (1) can be propagated forward in time in a recursive manner. The updating procedure is given in the following theorem and corollaries.

Theorem 1 *Let the MED prior at $t = 0$ be $\boldsymbol{\theta} \sim N(\mathbf{0}, \mathbf{I}), b \sim N(0, \infty)$, and $\mathrm{P}_0(\gamma_i) = C_{(0)}e^{-C_{(0)}(1-\gamma_i)}\mathcal{I}(\gamma_i \leq 1)$. Then given data $\mathcal{D}_{(\tau)}$ at time point τ, the relative entropy conjugate priors are*

$$\mathrm{P}_0\left(\boldsymbol{\theta}|\{\mathcal{D}_{(t)}\}_{t=1}^{\tau-1}\right) = N\left(\sum_{t=1}^{\tau-1} f(\boldsymbol{X}_{(t)})^T\boldsymbol{Y}_{(t)}\hat{\boldsymbol{\alpha}}_{(t)}, \mathbf{I}\right)$$

$$\mathrm{P}_0\left(b|\{\mathcal{D}_{(t)}\}_{t=1}^{\tau-1}\right) = N(0, \infty)$$

$$\mathrm{P}_0(\boldsymbol{\gamma}) = \prod_{i=1}^n C_{(\tau-1)} \exp\left\{-C_{(\tau-1)}(1 - \gamma_i)\right\}\mathcal{I}(\gamma_i \leq 1)$$

and the MED posterior $\mathrm{P}(\boldsymbol{\Theta}|\{\mathcal{D}\}_{t=1}^{\tau})$ can represented as

$$\mathrm{P}\left(\boldsymbol{\theta}|\{\mathcal{D}\}_{t=1}^{\tau}\right) = N\left(\boldsymbol{\mu}_0 + f(\boldsymbol{X}_{(\tau)})^T\boldsymbol{Y}_{(\tau)}\boldsymbol{\alpha}_{(\tau)}, \mathbf{I}\right)$$

where $\boldsymbol{\mu}_0 = \sum_{t=1}^{\tau-1} f(\boldsymbol{X}_{(t)})^T\boldsymbol{Y}_{(t)}\hat{\boldsymbol{\alpha}}_{(t)}$ is the prior mean and $\mathrm{P}(b|\{\mathcal{D}\}_{t=1}^{\tau})$ is the same as the Bayes non-informative prior.

Introducing the kernel function $k(\boldsymbol{x}, \boldsymbol{x}') = \langle f(\boldsymbol{x}), f(\boldsymbol{x}')\rangle$ and the parameter transformation $\boldsymbol{\omega} = f(\boldsymbol{X})\boldsymbol{\theta}$, the posterior at time $\tau > 0$ can be represented in terms of this kernel.

Corollary 1.1 *The equivalent prior at $t = 0$ for the transformed parameter is $\boldsymbol{\omega} \sim N(\mathbf{0}, \boldsymbol{K}_{(1)})$ where $\boldsymbol{K}_{(1)} = f(\boldsymbol{X}_{(1)})f(\boldsymbol{X}_{(1)})^T$. Furthermore, the posterior at time τ is of Gaussian form $\mathrm{P}(\boldsymbol{\omega}|\{\mathcal{D}_{(t)}\}_{t=1}^{\tau}) = N(\boldsymbol{\mu}_{(\tau)}, \boldsymbol{K}_{(\tau)})$ where the mean parameter satisfies the recursions $\boldsymbol{\mu}_{(\tau)} = \boldsymbol{\mu}_{(\tau-1)} + \boldsymbol{K}_{(\tau)}\boldsymbol{Y}_{(\tau)}\boldsymbol{\alpha}_{(\tau)}$.*

Since $\mathrm{P}(\boldsymbol{\theta}|\{\mathcal{D}_{(t)}\}_{t=1}^{\tau})$ is Gaussian, the MAP estimator is simply the mean parameter $\boldsymbol{\mu}_{(\tau)}$ given in the Corollary 1.1. Thus the decision rule reduces to $\hat{y}_{i'} = \mathrm{sgn}(f(\boldsymbol{x}_{i'})\hat{\boldsymbol{\theta}} + \hat{b})$ where the MAP estimator $\hat{\boldsymbol{\theta}}$ is a function of the previously estimated Lagrange multipliers $\hat{\boldsymbol{\alpha}}_{(1)}, \ldots, \hat{\boldsymbol{\alpha}}_{(\tau-1)}$ and the maximizing values $\hat{\boldsymbol{\alpha}}_{(\tau)}$ and $\hat{b}$ for the current time point τ.

Corollary 1.2 *Given all previous $\hat{\boldsymbol{\alpha}}_{(1)}, \ldots, \hat{\boldsymbol{\alpha}}_{(\tau-1)}$, the current optimal Lagrange multipliers $\hat{\boldsymbol{\alpha}}_{(\tau)}$ are the solution to*

$$\max_{\boldsymbol{\alpha}_{(\tau)}} -\frac{1}{2}\boldsymbol{\alpha}_{(\tau)}^T\boldsymbol{Y}_{(\tau)}\boldsymbol{K}_{(\tau)}\boldsymbol{Y}_{(\tau)}\boldsymbol{\alpha}_{(\tau)} + \sum_{i=1}^{n_{(\tau)}} \log(1 - \alpha_{(\tau)i}/C_{(\tau)}) + \boldsymbol{\alpha}_{(\tau)}^T\left(\mathbf{1} - \boldsymbol{Y}_{(\tau)}\sum_{t=1}^{\tau-1} k(\boldsymbol{X}_{(\tau)}, \boldsymbol{X}_{(t)})\boldsymbol{Y}_{(t)}\hat{\boldsymbol{\alpha}}_{(t)}\right)$$

subject to $\boldsymbol{y}_{(\tau)}^T\boldsymbol{\alpha}_{(\tau)} = 0$ and $\alpha_{(\tau)i} \geq 0$ for all $i = 1, \ldots, n_{(\tau)}$

and, holding the Lagrange multipliers fixed, the optimal bias $\hat{b} =$

$$\arg\min_b \sum_{s\in\{i|\hat{\alpha}_{(\tau)i}\neq 0\}} \left|\left(y_{(\tau)s} - \sum_{t=1}^{\tau} k(\boldsymbol{X}_{(\tau)s}, \boldsymbol{X}_{(t)})\boldsymbol{Y}_{(t)}\hat{\boldsymbol{\alpha}}_{(t)}\right) - b\right|$$

ensures that the expectation constraints in the objective hold.

The above dual formulation for the Lagrange multipliers $\boldsymbol{\alpha}_{(\tau)}$ has some interesting implications. Since the Lagrange multipliers from the previous time points are fixed at time step τ, the factor $\mathbf{1} - \boldsymbol{Y}_{(\tau)}\sum_{t=1}^{\tau-1} k(\boldsymbol{X}_{(\tau)}, \boldsymbol{X}_{(t)})\boldsymbol{Y}_{(t)}\hat{\boldsymbol{\alpha}}_{(t)}$ are constants and can be thought of as (unnormalized) weights for $\boldsymbol{\alpha}_{(\tau)}$, the Lagrange multipliers from the current time point. Thus the corresponding Lagrange multipliers for samples that are easily predicted using only the prior information will have lower weight than the Lagrange multipliers for samples that are difficult or incorrect.

3. MANIFOLD REGULARIZATION

Next we consider the case wheres some of the labels are missing. Without loss of generality we will assume the first l points are labeled and the latter $n - l$ points are unlabeled.

We will adopt the semi-supervised MED classification framework of [9], called Laplacian MED (LapMED). LapMED introduces an additional "geometric" constraint

$$\iint \mathrm{P}(\boldsymbol{\theta}, \lambda)\left(\int_{x\in\mathcal{M}} \boldsymbol{\theta}^T f(\boldsymbol{x})\Delta_{\mathcal{M}} f(\boldsymbol{x})\boldsymbol{\theta}\, d\mathcal{P}_x - \lambda\right) d\boldsymbol{\theta} d\lambda \leq 0 \quad (2)$$

to (1) where $\mathcal{M} = \mathrm{supp}(\mathcal{P}_X) \subset \mathbb{R}^n$ is a compact submanifold, $\Delta_{\mathcal{M}}$ is the Laplace-Beltrami operator on $\mathcal{M}$, and λ controls the complexity of the decision boundary in the intrinsic geometry of $\mathcal{P}_X$. This constraint was motivated by the semi-supervised framework of [8] to encourage the function $f(x)$ to be smooth over the support set of the feature distribution $\mathcal{P}_X$, inducing a geometric interpolation of unlabeled points. Since the marginal distribution is unknown, from [11]

$$f(\boldsymbol{X})^T\boldsymbol{L}f(\boldsymbol{X}) \to \int_{x\in\mathcal{M}} f(\boldsymbol{x})\Delta_{\mathcal{M}} f(\boldsymbol{x})\, d\mathcal{P}_x, \text{ as } n \to \infty$$

where $\boldsymbol{L}$ is the normalized graph Laplacian formed with a heat kernel. The LapMED posterior can be approximated as

$$\mathrm{P}(\boldsymbol{\theta}, b, \boldsymbol{\gamma}, \lambda|\mathcal{D}) = \frac{\mathrm{P}_0(\boldsymbol{\theta}, b, \boldsymbol{\gamma}, \lambda)}{Z(\boldsymbol{\alpha}, \beta)} \exp\left\{ \sum_{i=1}^{l} \alpha_i \left(y_i(f(\boldsymbol{X})\boldsymbol{\theta} + b) - \gamma_i\right) + \beta\left(\lambda - \boldsymbol{\theta}^T f(\boldsymbol{X})^T \boldsymbol{L} f(\boldsymbol{X})\boldsymbol{\theta}\right)\right\}$$

where $\beta \geq 0$ is a Lagrange multiplier for the smoothness constraint.

3.1. Sequential Laplacian MED

The distribution $\mathrm{P}(\boldsymbol{\Theta}, \boldsymbol{\gamma}, \lambda|\mathcal{D})$ that minimizes the objective with the additional constraint (2) can similarly be factorized and, like the distribution of slack parameters considered in Section 2, the distribution of the smoothness parameter λ is also independent of the data $\mathcal{D}$. Likewise, the distribution of the decision rule coefficients $\mathrm{P}(\boldsymbol{\Theta}|\mathcal{D})$ are conjugate distributions with their priors. Thus the updating procedure for the LapMED problem is similar to the updating procedure in Section 2.

Theorem 2 *At $t = 0$, the MED priors for $\boldsymbol{\theta}$ (or $\boldsymbol{\omega}$), b, and γ_i are the same as in Theorem 1, and the prior for λ is a Bayesian zero mean point prior, denoted $Exp.(\infty)$. Then given data $\mathcal{D}_{(\tau)}$ at time point τ, the MED conjugate prior and posterior are still $Exp.(\infty)$ for λ, the same as in Theorem 1 for b and γ_i, and Gaussian of form $N\left(\boldsymbol{\mu}_{(\tau)}, \boldsymbol{\Sigma}_{(\tau)}\right)$ for $\boldsymbol{\theta}$ (or $\boldsymbol{\omega}$). Define a $l \times n$ expansion matrix as $\boldsymbol{J} = [\mathbf{I}\ \mathbf{0}]$. Then the mean and covariance parameters for the distribution of $\boldsymbol{\theta}$ are*

$$\boldsymbol{\mu}_{(\tau)} = \boldsymbol{G}_{(\tau)}^{-1} \sum_{t=1}^{\tau} f(\boldsymbol{X}_{(t)})^T \boldsymbol{J}^T \boldsymbol{Y}_{(t)} \boldsymbol{\alpha}_{(t)}, \quad \boldsymbol{\Sigma}_{(\tau)} = \boldsymbol{G}_{(\tau)}^{-1},$$

where $\boldsymbol{G}_{(\tau)} = \boldsymbol{G}_{(\tau-1)} + 2\beta_{(\tau)} f(\boldsymbol{X}_{(\tau)})^T \boldsymbol{L}_{(\tau)} f(\boldsymbol{X}_{(\tau)})$ is a recursive graph of vertex disjoint subgraphs, and for the distribution of $\boldsymbol{\omega}$ are

$$\boldsymbol{\mu}_{(\tau)} = \sum_{t=1}^{\tau} k_{(\tau)}\left(\boldsymbol{X}_{(\tau)}, \boldsymbol{X}_{(t)}\right) \boldsymbol{J}^T \boldsymbol{Y}_{(t)} \boldsymbol{\alpha}_{(t)}, \ \boldsymbol{\Sigma}_{(\tau)} = k_{(\tau)}\left(\boldsymbol{X}_{(\tau)}, \boldsymbol{X}_{(\tau)}\right)$$

where $k_{(\tau)}(\boldsymbol{x}, \boldsymbol{x}') = \langle f(\boldsymbol{x}), \boldsymbol{G}_{(\tau)}^{-1} f(\boldsymbol{x}')\rangle$ is a kernel function that can be recursively defined as

$$k_{(\tau)}(\boldsymbol{x}, \boldsymbol{x}') = k_{(\tau-1)}(\boldsymbol{x}, \boldsymbol{x}') - k_{(\tau-1)}(\boldsymbol{x}, \boldsymbol{X}_{(\tau)}) \Big(\left(2\beta_{(\tau)} \boldsymbol{L}_{(\tau)}\right)^{-1} + k_{(\tau-1)}\left(\boldsymbol{X}_{(\tau)}, \boldsymbol{X}_{(\tau)}\right) \Big)^{-1} k_{(\tau-1)}(\boldsymbol{X}_{(\tau)}, \boldsymbol{x}'). \quad (3)$$

Theorem 2 gives the posterior distribution for semi-supervised classification whose form is comparable to the form given in Corollary 1.1 for the supervised case. Indeed the forms are identical except for the presence of the precision matrix term $G_{(\tau)}$ in the semi-supervised case. As the sparsity of $G_{(\tau)}$ is associated with the graph Laplacian, the kernel function of the semi-supervised case is a regularized version of the kernel function that appears in Corallary 1.1. If we let $\beta_{(t)}$ be a fixed parameter, then $\hat{\boldsymbol{\alpha}}_{(t)}$ and $\hat{b}$ optimize an objective of the same form as in Corollary 1.2, but with kernel function $k_{(\tau)}(\boldsymbol{x}, \boldsymbol{x}')$. If $\beta_{(t)}$ is chosen to be 0, the sequential LapMED simply ignores the unlabeled data of time point t, and if all $\beta_{(i)}$'s are 0, then the unlabeled data is always ignored and the updating procedure is exactly the same as in the supervised scenario. These parameters are functions of the γ_A and γ_I, which are identical to the penalty parameters in the Laplacian SVM [8], associated with the reproducing kernel Hilbert space and data distribution respectively: $C_{(t)} = \frac{1}{2l_{(t)}\gamma_A}$ and $\beta_{(t)} = \frac{\gamma_I}{2\gamma_A n_{(t)}^2}$.

3.2. Approximating the Kernel Function

Because the kernel function in (3) is a function of the previous kernel functions, calculating a map to its associated Hilbert space $\mathcal{H}_{(\tau)}$ can be computationally expensive. Thus in this subsection, we derive an approximation to the map to $\langle \boldsymbol{x}, \boldsymbol{x}'\rangle_{\mathcal{H}_{(\tau)}}$, which is computationally easier than direct recursive calculation.

Recall that we approximate the constraint in (2), at any time point t, empirically with the graph Laplacian $\boldsymbol{L}_{(t)}$ formed using the data from that time point $\boldsymbol{X}_{(t)}$. However, the non-empirical constraint using the Laplace-Beltrami operator over the unknown marginal distribution $\mathcal{P}_x$, is actually the same at every time point. Thus as $n_{(\tau-1)} \to \infty$, the prior graph $\boldsymbol{G}_{(\tau-1)}$ converges to

$$B \int_{x \in \mathcal{M}} f(\boldsymbol{x}) \Delta_{\mathcal{M}} f(\boldsymbol{x})\, d\mathcal{P}_x \approx B \sum_{i=1}^{\infty} \delta_i \xi_i^2 \upsilon_i(z) \upsilon_i(z) \quad (4)$$

where $B = 2\sum_{t=1}^{\tau} \beta_{(t)}$, δ_i are the eigenvalues of the Laplace-Beltrami operator, and $\upsilon_i(z)$ and ξ_i are the infinite sequence of right singular functions and singular values of $f(x) = \int k(x, z) f(z)\, dz$. The approximate decomposition arises since the left singular functions of f are the eigenfunctions of the Laplace-Beltrami operator [12] and [8]. Thus instead of empirically approximating the Laplacian as a sum of subgraphs $\boldsymbol{G}_{(\tau-1)} = \mathbf{I} + \sum_{t=1}^{\tau-1} 2\beta_{(t)} f(\boldsymbol{X}_{(t)})^T \boldsymbol{L}_{(t)} f(\boldsymbol{X}_{(t)})$, we can instead implement approximations to the eigen/singular values and singular functions in (4).

Assuming that the sample size n is large enough, the average eigenvalues of the $\tau - 1$ graph Laplacians would be a good estimator for the eigenvalues of the Laplace-Beltrami operator. Additionally the rows of the matix $\boldsymbol{V}^T$ from the singular value decomposition of $\boldsymbol{X}$ will contain the basis for its row space. Thus because the right singular functions form an orthonormal basis for the coimage of f, if the mapping approximately preserves the basis, the mapped average singular vectors $f(\bar{\boldsymbol{V}}_i)$ would be good estimators for the right singular functions $\upsilon_i(z)$ and correspondingly so for the singular values.

The posterior kernel function $k_{(\tau)}(\boldsymbol{x}, \boldsymbol{x}')$ using an approximation to the decomposition in (4) will no longer be a recursive function of prior kernel functions $k_{(\tau-1)}(\boldsymbol{x}, \boldsymbol{x}')$ that have the same form, like in (3). Instead for $\tau > 2$, it uses a prior kernel function

$$\tilde{k}_{(\tau-1)}(\boldsymbol{x}, \boldsymbol{x}') = k(\boldsymbol{x}, \boldsymbol{x}') - k(\boldsymbol{x}, \bar{\boldsymbol{V}}_{(\tau-1)}) \Big(\frac{\mathrm{diag}(\bar{\boldsymbol{s}}_{(\tau-1)}^2 \bar{\boldsymbol{d}}_{(\tau-1)})^{-1}}{B} + k(\bar{\boldsymbol{V}}_{(\tau-1)}, \bar{\boldsymbol{V}}_{(\tau-1)}) \Big)^{-1} k(\bar{\boldsymbol{V}}_{(\tau-1)}, \boldsymbol{x}').$$

where $k(\boldsymbol{x}, \boldsymbol{x}') = \langle f(\boldsymbol{x}), f(\boldsymbol{x}')\rangle$ is the non-regularized kernel function. So at time τ, the singular vectors of $\boldsymbol{X}_{(\tau-1)}$ are used to update the average singular vectors, in the above function, through

$$\bar{\boldsymbol{V}}_{(\tau-1)} = \bar{\boldsymbol{V}}_{(\tau-2)} + \frac{\boldsymbol{V}_{(\tau-1)} - \bar{\boldsymbol{V}}_{(\tau-2)}}{\tau - 1}$$

and similarly so for the average corresponding singular values $\bar{\boldsymbol{s}}_{(\tau-1)}$ and the average eigenvalues of the graph Laplacians $\bar{\boldsymbol{d}}_{(\tau-1)}$.

4. EXPERIMENTS

In this section, we compare the proposed sequential maximum margin classifiers to popular supervised and semi-supervised maximum margin classifiers (SVM [4] and LapSVM [8]) where the model is trained using just the current time points data and where the model has been re-trained on all previous data. The former type of model is a lower bound on performance since it ignores all previous data and the latter type of model is an upper bound since it is re-trained on all previous data at every time point. Note the MED and SVM models only differ by a weak log-barrier term in the objective function making their performance identical, and similarly so for LapMED and LapSVM. Thus their performance curves will referred to as Full SVM/MED and Full LapSVM/LapMED.

4.1. Simulations

In both of the following simulations, the models receive roughly 100 samples ($n_{(t)} = [97, 103]$) at every time point, the parameters are empirically chosen with a validation set, and then the models are tested on an independent data set of 1000 test points. The test accuracy $\frac{TP+TN}{1000}$ is the average accuracy over 100 trials of simulation.

In the first simulation, we generate data from 200 categorical distributions where 100 of the variables are sparse so they have high probability of being 0, another 50 of the variables have lower probability of being 0, and the final 50 variables are used to distinguish between the two classes. We use the term frequency - inverse document frequency (TF-IDF) kernel of [13], which is used in document processing and topic models. Figure 1 shows that the accuracy of the sequential model (SeqMED) improves as the model is updated with more training data and has much better results even after one model update versus the independent model (SVM) that ignores previous training data. Of course the sequential model does not improve as rapidly as the model that is re-trained on all the data (Full SVM/MED), but this is the price paid for lower computational complexity. For example, at $t = 30$, SeqMED updates and fits 100 coefficients for the new data whereas Full SVM/MED fits 3,000 coefficients for all the data.

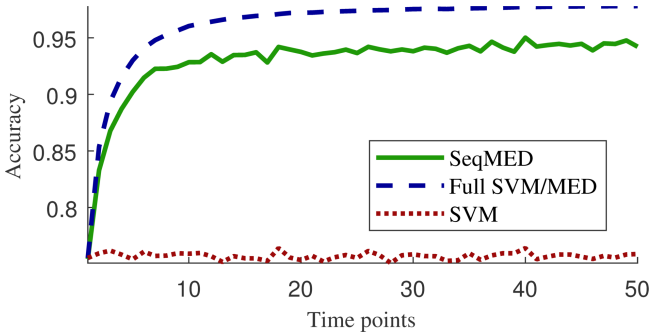

Fig. 1. Accuracy of prediction for categorical fully labeled simulated data. The proposed sequential MED (SeqMED) classifier performs almost as well as the full batch implementation of the SVM/MED (Full SVM/MED).

In the second simulation, we generate data from the interior of a 3-dimensional sphere where one class is roughly at the center of the sphere and the other class is on the shell, but only 10% of the samples are labeled. We use a rbf kernel with width 1 for the kernel function and a heat kernel with width 0.01 and a 20 nearest neighbors graph for the graph Laplacian. Figure 2 shows improvement in performance of the sequential model similar to in Figure 1. We use the approximate kernel function of Subsection 3.2 to perform each update, establishing that the approximation is adequate.

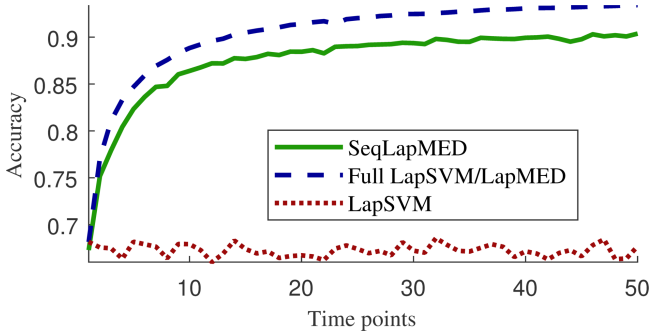

Fig. 2. Accuracy of prediction for continuous simulated data with 10% labeled.

4.2. Data

We compare the proposed algorithms on the Isolet speech database from the UCI machine learning repository [14] following the experimental framework used in [8]. To train the models, we take the entire training set of 120 speakers (isolet1 - isolet4) and break them into 24 groups (time points) of 5 speakers where only the first speaker is labeled. At each time point, the models train on 260 samples ($t = 21$ and 23 only have 259) where 52 of the samples are labeled. The parameters are set in the same way as in [8] and the test set is similarly composed of the 1,559 samples from isolet5. Figure 3 shows that, after two time points, the sequential model always performs better than the model that ignores previous data, and comes close to performing as well as the fully re-trained model as time progresses.

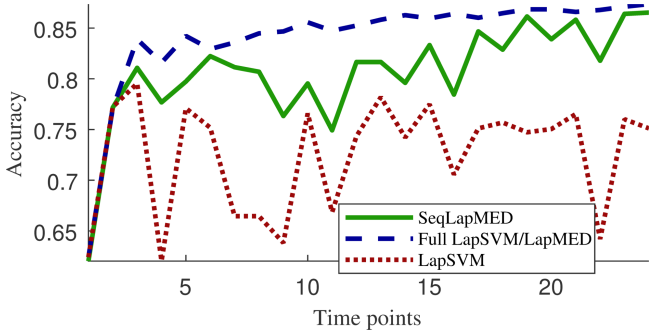

Fig. 3. Accuracy of prediction on isolet5 for models trained on partially labeled speech isolets 1-4. The proposed semi-supervised sequential Laplacian MED classifier (SeqLapMED) comes close to the full Laplacian SVM [8] as time progresses.

5. CONCLUSIONS

We have proposed recursive versions of supervised and semi-supervised maximum margin classifiers in the minimum entropy discrimination (MED) classification framework. The proposed sequential maximum margin classifiers perform nearly as well as a much more computationally expensive fully re-trained maximum margin classifiers and significantly better than a classifier that ignores previous data.

6. REFERENCES

[1] Tommi Jaakkola, Marina Meila, and Tony Jebara, "Maximum entropy discrimination," in *Advances in Neural Information Processing Systems 12*, S.A. Solla, T.K. Leen, and K. Müller, Eds. 2000, pp. 470–476, MIT Press.

[2] Grace Wahba, "Support vector machines, reproducing kernel hilbert spaces and the randomized gacv," *Advances in Kernel Methods-Support Vector Learning*, vol. 6, pp. 69–87, 1999.

[3] Tommi S Jaakkola and David Haussler, "Probabilistic kernel regression models.," in *AISTATS*, 1999.

[4] Alex J Smola, Bernhard Schölkopf, and Klaus-Robert Müller, "The connection between regularization operators and support vector kernels," *Neural networks*, vol. 11, no. 4, pp. 637–649, 1998.

[5] Manfred Opper and Ole Winther, "Gaussian process classification and svm: Mean field results and leave-one-out estimator," *Advances in Large Margin Classifiers*, 1999.

[6] Peter Sollich, "Bayesian methods for support vector machines: Evidence and predictive class probabilities," *Machine Learning*, vol. 46, no. 1, pp. 21–52, 2002.

[7] Carl Edward Rasmussen and Christopher K. I. Williams, *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*, The MIT Press, 2005.

[8] Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *J. Mach. Learn. Res.*, vol. 7, pp. 2399–2434, Dec. 2006.

[9] Elizabeth Hou, Kumar Sricharan, and Alfred O Hero, "Latent laplacian maximum entropy discrimination for detection of high-utility anomalies," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 6, pp. 1446–1459, June 2018.

[10] Oluwasanmi Koyejo and Joydeep Ghosh, "A representation approach for relative entropy minimization with expectation constraints," in *ICML WDDL workshop*, 2013.

[11] Alexander Grigoryan, "Heat kernels on weighted manifolds and applications," *Cont. Math*, vol. 398, pp. 93–191, 2006.

[12] R. R. Lederman and V. Rokhlin, "On the analytical and numerical properties of the truncated laplace transform i.," *SIAM Journal on Numerical Analysis*, vol. 53, no. 3, pp. 1214–1235, 2015.

[13] Charles Elkan, "Deriving tf-idf as a fisher kernel," in *SPIRE*. Springer, 2005, vol. 3772, pp. 295–300.

[14] M. Lichman, "UCI machine learning repository," 2013.