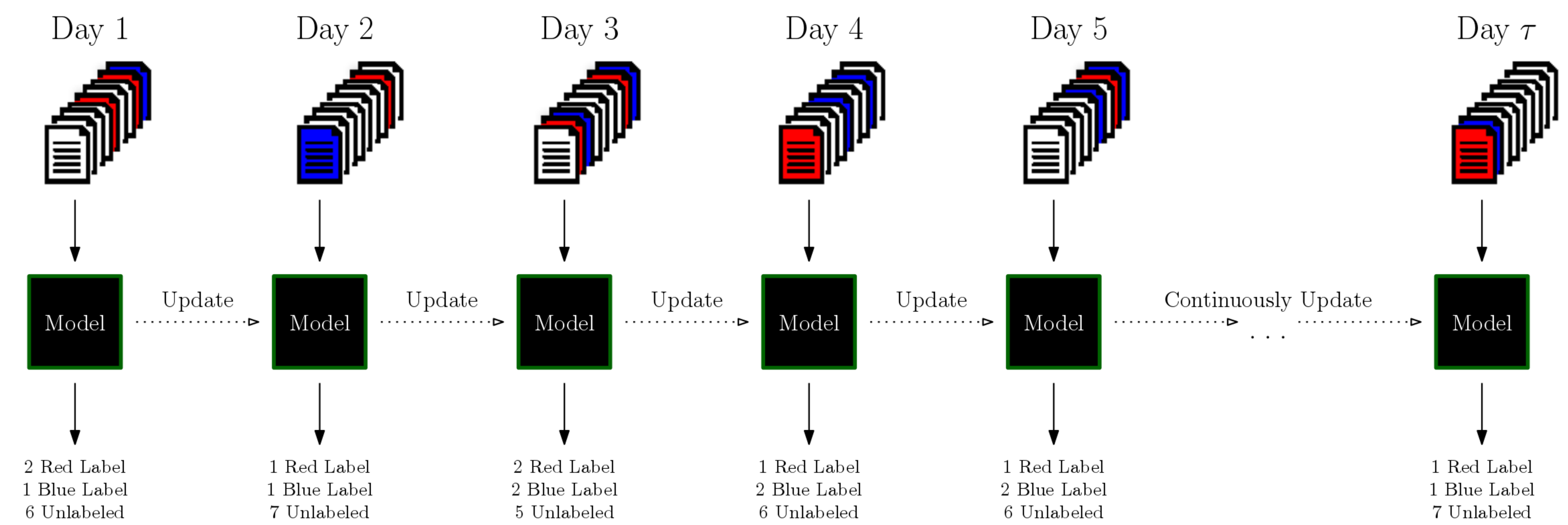
Sequential Maximum Margin Classifiers for Partially Labeled Data

Elizabeth Hou[†], Alfred O. Hero III[†]

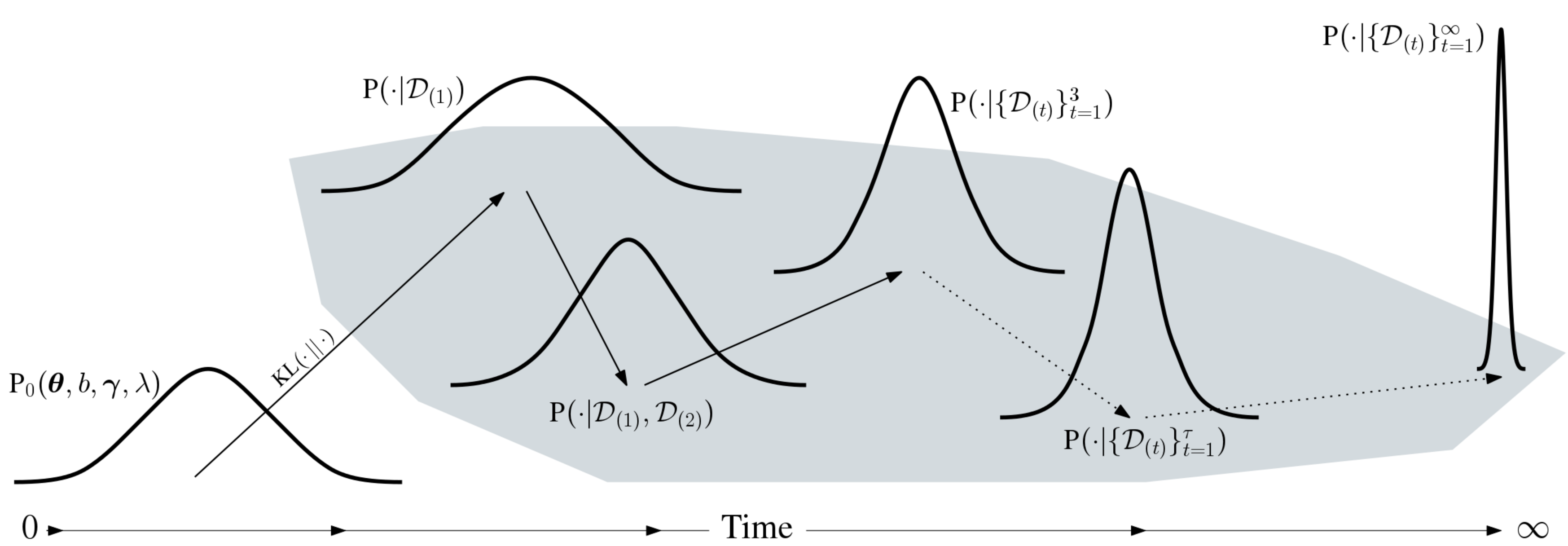
[†] University of Michigan, EECS Department

Motivation



- Data is not collected as a batch, but sequentially over time
  - Speech or streaming text classification
  - Satellite only transmits data daily
  - Agency's quarterly reports
- Often not possible/desirable to wait until complete before analyzing
- Additionally, some of the labels may be missing
- At any time point t, observed data as $\mathcal{D}_{(t)} = \{\boldsymbol{X}_{(t)}, \boldsymbol{y}_{(t)}\}$
  - $\boldsymbol{X}_{(t)}$ is a matrix of $n_{(t)}$ samples and p feature variables
  - $l_{(t)} < n_{(t)}$ of the samples have label $y_i = [1, -1]$

Sequential Laplacian MED



Parameters:

- $\boldsymbol{\theta}$ are weights for decision boundary
- b is a bias term for decision boundary
- γ_i are margin parameters
- λ is a regularizer

Objective:

$$\min_{\mathsf{P}(\cdot|\{\mathcal{D}_{(t)}\}_{t=1}^{\tau})} \mathsf{KL}\left(\mathsf{P}(\cdot|\{\mathcal{D}_{(t)}\}_{t=1}^{\tau}) || \mathsf{P}_0(\cdot|\{\mathcal{D}_{(t)}\}_{t=1}^{\tau-1})\right)$$

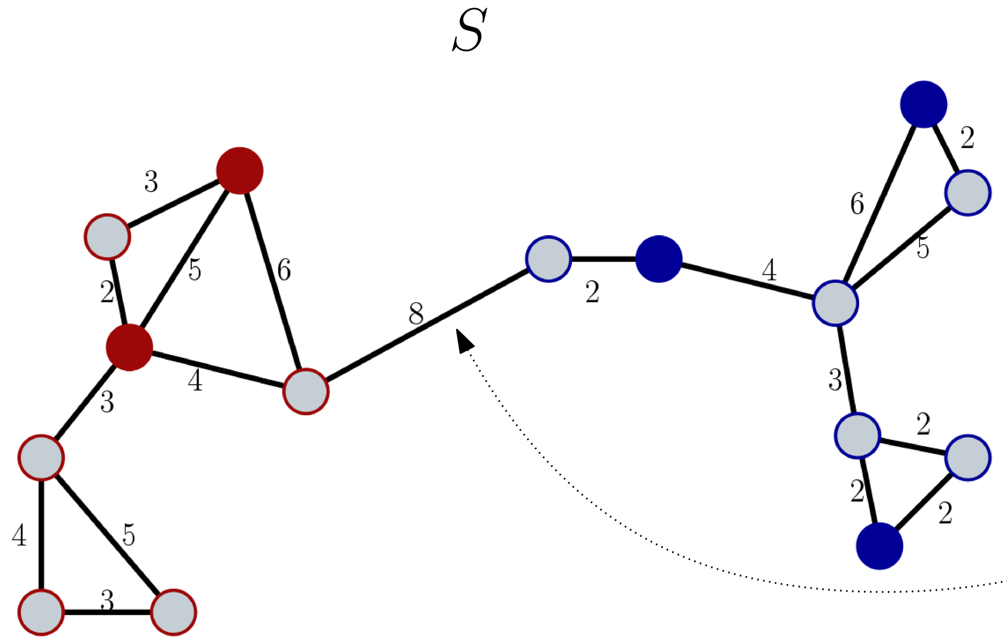
subject to

$$\mathsf{E}_{\theta,b,\gamma}\big(\underbrace{y_{(\tau)i}(\boldsymbol{X}_{(\tau)i}\boldsymbol{\theta} + b) - \gamma_i}_{H_i} |\{\mathcal{D}_{(t)}\}_{t=1}^{\tau}\big) \geq 0 \ \ \forall i$$

Soft Margin Hinge Loss

$$\mathsf{E}_{\theta,\lambda}\big(\underbrace{\boldsymbol{\theta}^T \boldsymbol{X}_{(\tau)}^T \boldsymbol{L}_{(\tau)} \boldsymbol{X}_{(\tau)} \boldsymbol{\theta} - \lambda}_{S} |\{\mathcal{D}_{(t)}\}_{t=1}^{\tau}\big) \leq 0$$

Regularize Smoothness w.r.t. distribution of x



$\boldsymbol{L}_{(\tau)}$ is the normalized graph Laplacian

- Form $\boldsymbol{L}_{(\tau)}$ using knn from labeled and unlabeled $\boldsymbol{X}_{(\tau)i}$
- Points with same label are geometrically close
- Want decision boundary to prefer "minCuts" $(e^{-dist.})$

$$\mathsf{P}(\boldsymbol{\theta}, b, \boldsymbol{\gamma}, \lambda|\{\mathcal{D}_{(t)}\}_{t=1}^{\tau}) = \frac{\mathsf{P}_0(\cdot|\{\mathcal{D}_{(t)}\}_{t=1}^{\tau-1})}{Z(\boldsymbol{\alpha}_{(\tau)}, \beta_{(\tau)})} \exp\left\{\sum_{i=1}^{l_{(\tau)}} \alpha_{(\tau)i} H_i + \beta_{(\tau)} S\right\}$$

If prior is exp. family, the minimizing density is of same family [2, 3]

Update Theorem [4]

At time τ, let the MED priors be $\boldsymbol{\theta} \sim N(\boldsymbol{\mu}_{(\tau-1)}, (\boldsymbol{G}_{(\tau-1)})^{-1})$, $b \sim N(0, \sigma^2)$ and $\lambda \sim Exp.(\nu)$ where $\sigma, \nu \to \infty$, and $\gamma_i \sim \mathsf{P}_0(C_{(\tau)})$.

Then the posterior also factorizes with $\mathsf{P}(\boldsymbol{\theta}|\{\mathcal{D}_{(t)}\}_{t=1}^{\tau})$ as Gaussian:

mean $\boldsymbol{\mu}_{(\tau)} = \boldsymbol{G}_{(\tau)}^{-1}\left(\boldsymbol{G}_{(\tau-1)}\boldsymbol{\mu}_{(\tau-1)} + \boldsymbol{X}_{(\tau)}^T \boldsymbol{J}^T \boldsymbol{Y}_{(\tau)} \hat{\boldsymbol{\alpha}}_{(\tau)}\right)$ and

precision matrix $\boldsymbol{G}_{(\tau)} = \boldsymbol{G}_{(\tau-1)} + 2\beta_{(\tau)} \boldsymbol{X}_{(\tau)}^T \boldsymbol{L}_{(\tau)} \boldsymbol{X}_{(\tau)}$.

Decision Rule Corollary [4]

Assume $\beta_{(t)}$ are fixed parameters. Then the decision rule reduces to $\hat{y}_{i'} = \mathsf{sgn}(\boldsymbol{X}_{i'}\boldsymbol{\mu}_{(\tau)} + \hat{b})$, which is a function of the previous mean and optimal parameters $\hat{\boldsymbol{\alpha}}_{(\tau)} = \arg\max_{\boldsymbol{\alpha}_{(\tau)}} Z(\boldsymbol{\alpha}_{(\tau)}, \beta_{(\tau)}) =$

$$-\frac{1}{2}\boldsymbol{\alpha}_{(\tau)}^T \boldsymbol{Y}_{(\tau)} \boldsymbol{J} \boldsymbol{X}_{(\tau)} \boldsymbol{G}_{(\tau)}^{-1} \boldsymbol{X}_{(\tau)}^T \boldsymbol{J}^T \boldsymbol{Y}_{(\tau)} \boldsymbol{\alpha}_{(\tau)} + \boldsymbol{\alpha}_{(\tau)}^T \left(\boldsymbol{1} - \boldsymbol{Y}_{(\tau)} \boldsymbol{J} \boldsymbol{X}_{(\tau)} \boldsymbol{\mu}_{\tau-1}\right)$$

$$+ \sum_{i=1}^{l_{(\tau)}} \log\left(1 - \frac{\alpha_{(\tau)i}}{C_{(\tau)}}\right) \quad \text{s.t.} \quad \boldsymbol{y}_{(\tau)}^T \boldsymbol{\alpha}_{(\tau)} = 0, \ \alpha_{(\tau)i} \geq 0 \ \forall i = 1, \ldots, l_{(\tau)}$$

and the $\hat{b}$ that statisfy the KKT conditions.

The above can be extended to non-linear decision boundaries with the kernel trick [4].

ACKNOWLEDGMENTS This work was funded partially 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.

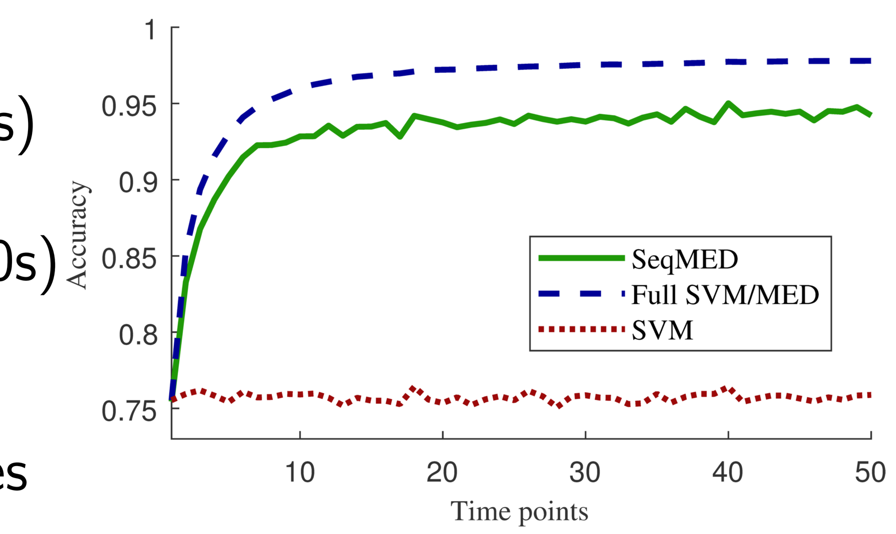
Experiments

Simulation

- Training: ~ 100 per time point $(n_{(t)} = [97, 103])$
- Test: 1000 test points, Accuracy $= (TP + TN)/1000$

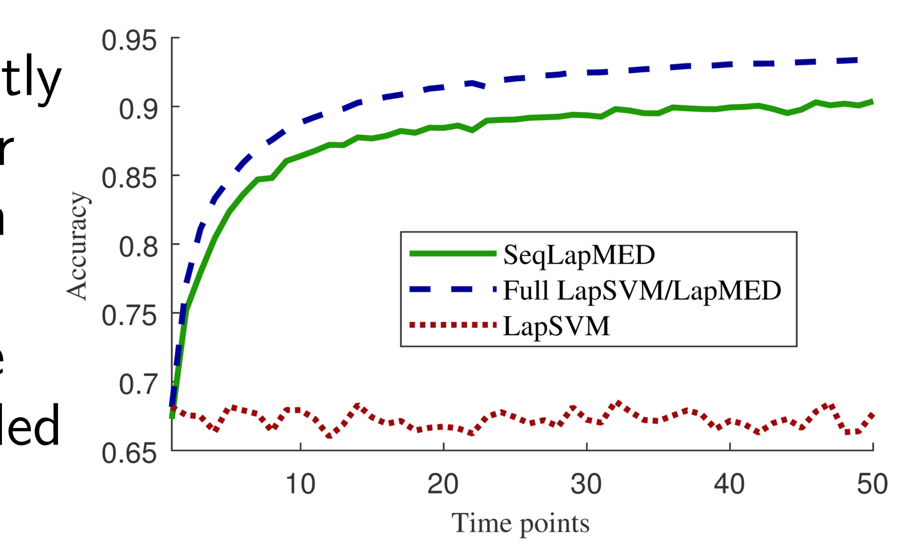
Supervised: Data from 200 categorical distributions

- 100 are sparse (high prob. of 0s)
- 50 are relevant (lower prob. of 0s)
- 50 are used to distinguish between 2 classes



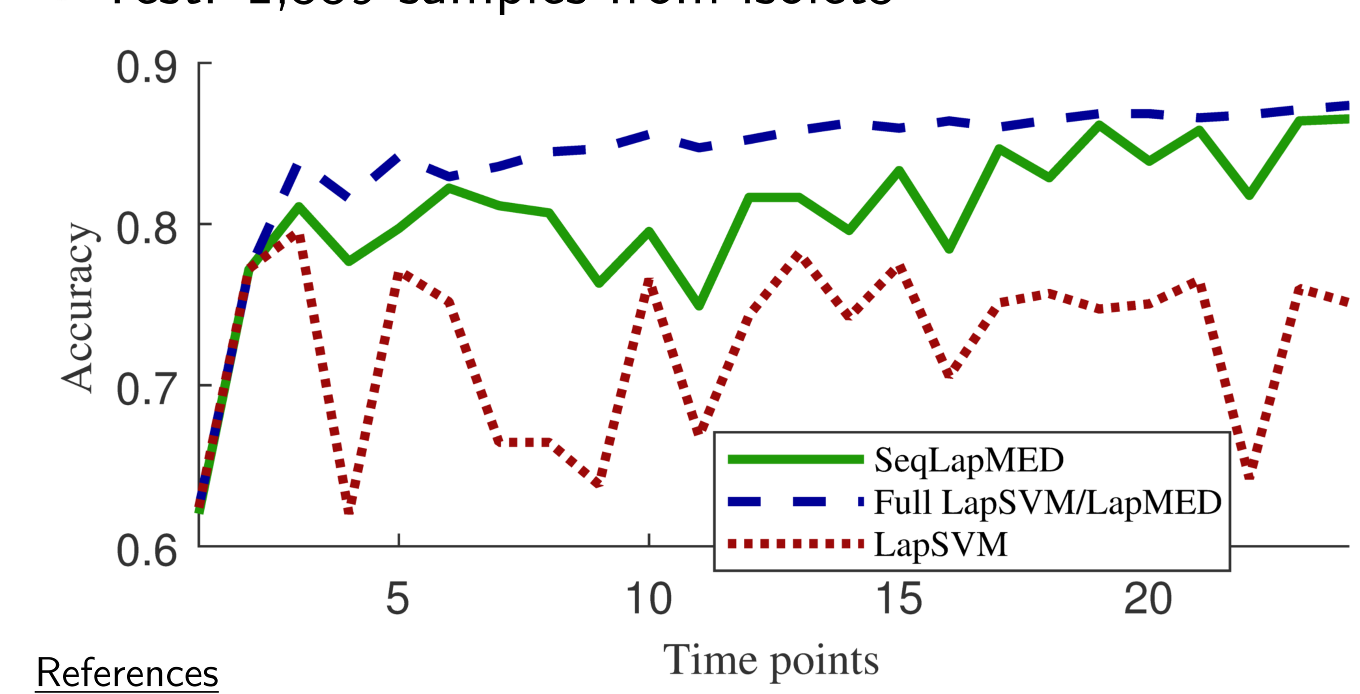
Semi-supervised: Data from interior of a 3-D sphere

- One class is mostly in center/interior
- Other class is on the shell
- Only 10% of the samples are labeled



Isolet Speech Database

- Follows the experimental framework used in [1]
- Training: (isolet1 - isolet4) broken into 24 time points of 5 speakers, only first speaker is labeled
- Test: 1,559 samples from isolet5



References

[1] M. Belkin, P. Niyogi, and V. Sindhwani. "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. JMLR (2006).

[2] O. Koyejo, J. Ghosh, "A representation approach for relative entropy minimization with expectation constraints. In 2013 ICML WDDL workshop.

[3] E. Hou, K. Sricharan, A. O. Hero. "Latent Laplacian Maximum Entropy Discrimination for Detection of High-Utility Anomalies". IEEE TIFS (2018).

[4] E. Hou, A. O. Hero. "Sequential Maximum Margin Classifiers for Partially Labeled Data". In 2018 IEEE ICASSP.