Active Learning Team
An Annotated Bibliography

Xoaquin Baca, Jack Mauro, Jason Setiadi, Zhan Shi
UCLA CAM REU

June 25, 2021

References

[1] A. Bertozzi and A. Flenner, "Diffuse interface models on graphs for classification of high dimensional data," *SIAM Review*, vol. 58, pp. 293–328, 01 2016.

This article first introduces the Ginzburg-Landau (GL) functional, its Γ-convergence to the total variation seminorm as well as its connection with the Allen-Cahn and Cahn-Hilliard equations and image processing. Convex splitting can be applied to minimizing GL functionals. (Section 1) Section 2 generalizes to graphs and Section 3 provides algorithms for convex splitting and Nyström extension. Experiments on House voting records, two-moon dataset and image segmentation show that this method can achieve high accuracy and good generalizability (Section 4). Connections with graph cuts, nonlocal means and geometric diffusion are made in Section 5, while related work including the MBO scheme is introduced in Section 6.

[2] A. L. Bertozzi, "Graphical models in machine learning, networks and uncertainty quantification," in *proceedings of 2018 International Congress of Mathematicians*, vol. 3. World Scientific, 2018, pp. 3853–3880.

This paper is a review article on semi-supervised and unsupervised graph models for classification using similarity

graphs and for community detection in networks. Graphical models generate useful information based on connections between nodes usually called the similarity graph. It helps to solve the problem of computational complexity in machine learning using efficient algorithms. The authors proposed the Ginzburg-Landau (GL) functional as a regularizer for semi-supervised learning on classification problems. The authors also proposed an even more efficient method called the MBO method for a graphical setting, showing significant speed-up in run time compared to the Ginzburg-Landau method. This paper reviews some of this literature and discusses future problem areas including crossover work between network modularity and machine learning and efforts in uncertainty quantification.

[3] J. Calder, B. Cook, M. Thorpe, and D. Slepčev, "Poisson learning: Graph based semi-supervised learning at very low label rates," in *37th International Conference on Machine Learning, ICML 2020*, ser. 37th International Conference on Machine Learning, ICML 2020, H. Daume and A. Singh, Eds. International Machine Learning Society (IMLS), 2020, pp. 1283–1293.

This paper introduces a graph based learning technique called Poisson Learning. The necessity of Poisson Learning is motivated by the poor performance of Laplace learning from [21] on low label rates. The authors identify a bias that is present in the solution to the Laplace equation at low label rates and introduce Possion Learning to rectify this bias. In this method, one computes the solution to the Poisson Equation $\mathcal{L}u(x_i) = \sum_{j=1}^{m}(y_j - \overline{y})\delta_{ij}$ satisfying $\sum_{i=1}^{n} d_i u(x_i) = 0$. The paper provides a random walk interpretation of graph based learning to demonstrate that the poor performance in the Laplace algorithm at low label rates is caused by the very large stopping time τ of the random walk. The paper also introduces a graph-cut Poisson MBO algorithm that slightly improves accuracy from the Poissson algorithm. Experimental results demonstrate that Poisson and Poisson MBO outperform other graph based learning methods on data with low label rates. The main contribution of this paper is the

introduction of the Poisson Algorithm that is stable at low label rates.

[4] Y. Chong, Y. Ding, Q. Yan, and S. Pan, "Graph-based semi-supervised learning: A review," *Neurocomputing*, vol. 408, pp. 216–230, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231220304938

This paper provides a comprehensive review of both semi-supervised learning as a whole as well as graph based semi-supervised learning. It offers an account of how to construct a similarity graph from the given data and different transductive and inductive techniques used to assign labels to the initially unlabeled vertices. The paper offers a brief overview of various semi-supervised and unsupervised methods for graph based learning. This paper is useful to provide an introduction to graph based learning as a whole before taking a deeper dive into the various methods used in graph based learning.

[5] A. Gadde, A. Anis, and A. Ortega, "Active semi-supervised learning using sampling theory for graph signals," *Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining*, 2014.

This paper establishes an active learning scheme based on sampling theory involving the cutoff frequency, the optimal sampling set and the reconstruction algorithm. For the experimental results, at 10% of data being labeled, the proposed method reaches $\sim 85\%$ accuracy in handwritten digits classification, $\sim 75\%$ accuracy in spoken letters classification and $\sim 55\%$ accuracy in text classification. In all three tasks, the proposed method leads other methods by at least 5% accuracy. In short, this method can produce higher accuracy in many cases, but text classification remains a challenge for active learning schemes.

[6] Q. Gu and J. Han, "Towards active learning on graphs: An error bound minimization approach," *2012 IEEE 12th International Conference on Data Mining*, pp. 882–887, 2012.

This paper derives a generalization error bound for a graph-based learning method called *Learning with Local*

and Global Consistency (LLGC), inducing an active learning scheme via minimizing the error bound by solving $\text{argmin}_{\mathcal{L}\subset\mathcal{V}}\text{tr}\big((\mu\mathbf{L}_{\mathcal{L}\mathcal{L}}+\mathbf{I})^{-2}\big)$ where $\mathbf{L}_{\mathcal{L}\mathcal{L}}$ denotes the principle submatrix of $\mathbf{L}$ corresponding to the labeled set $\mathcal{L}$ and $\mathcal{V}$ being the set of all nodes. By introducing a *selection matrix* $\mathbf{S}$ with $S_{ij}=1$ if the data point $\mathbf{x}_i$ is selected as the j-th point in $\mathcal{L}$ and 0 otherwise and $\mathcal{S}$ being the set of all selection matrices, the minimization problem becomes $\text{argmin}_{\mathbf{S}\subset\mathcal{S}}\text{tr}\big((\mu\mathbf{S}^T\mathbf{L}\mathbf{S}+\mathbf{I})^{-2}\big)$. Eigen-decomposing $\mathbf{L}=\mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$ with $\mathbf{\Lambda}=\text{diag}(\lambda_1,\cdots,\lambda_n)$ and letting $\mathbf{\Gamma}=\text{diag}\big((\mu\lambda_1+1)^2-1,\cdots,(\mu\lambda_n+1)^2-1\big)$, the minimization problem becomes $\text{argmin}_{\mathcal{L}\subset\mathcal{V}}\text{tr}\big((\mathbf{\Gamma}^{-1}+\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U})^{-1}\mathbf{\Gamma}^{-1}\big)$. Here $\mathbf{\Gamma}^{-1}$ is approximated by replacing the first entry (of value zero as $\lambda_1\equiv 0$) of $\mathbf{\Gamma}$ with 10^{-6}. This problem is then easy to solve with iterative methods using the Sherman-Morrison formula, where the updating step is efficient (without matrix inversion). The performance of this method is generally better than that of methods like random sampling and variance minimization, reaching an accuracy of $\sim 70\%-80\%$ with 160 labeled data for all the datasets tested.

[7] G. Iyer, J. Chanussot, and A. L. Bertozzi, "A graph-based approach for data fusion and segmentation of multimodal images," *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

This article proposes a graph-based algorithm for feature extraction and segmentation of multimodal images. The authors applied this method on Data Fusion Challenge 2015 Images (DFC2015), DFC2013, and DFC2018, which are optical and LiDAR images. The algorithms being performed are spectral clustering, semisupervised graph MBO, and k-means as a more naive algorithm. The result shows that graph-based methods can give a more data-driven model by reducing the information from each modality. Then, we can calculate and compare different weight matrices to create the graph Laplacian of the data and extract features in the form of eigenvectors. These features can then be used as a part of many different data- segmentation algorithms.

[8] M. Ji and J. Han, "A variance minimization criterion to active learning on graphs," in *Proceedings of the Fifteenth International Conference*

on Artificial Intelligence and Statistics, ser. Proceedings of Machine Learning Research, N. D. Lawrence and M. Girolami, Eds., vol. 22. La Palma, Canary Islands: PMLR, 21–23 Apr 2012, pp. 556–564. [Online]. Available: http://proceedings.mlr.press/v22/ji12.html

This paper introduces a method for graph based learning known as variance minimization. The variance minimization approach is based solely on the graph structure. The authors analyze the multivariate Gaussian distribution over the unlabeled nodes of the graph representation of the dataset. Using this multivariate Gaussian distribution for the unlabeled nodes, the authors introduce a formula to calculate the expected prediction error of the unlabeled nodes, and they select the nodes to label by solving the optimization problem arg min $\mathrm{Tr}(L_{uu}^{-1})$ for $\mathcal{L} \subset \mathcal{V}$. The paper also provides a methodology for selecting the first datapoint to label. When compared with empirical risk minimization, random selection, label selection based on clustering, and uncertainty sampling, variance minimization performed significantly better at lower label rates and outperformed other methods when there were many (20+) classification categories. The authors used MNIST, DBLP, and Isolet for their experiment.

[9] H. Li, "Graph-based learning and data analysis," 2020. [Online]. Available: https://escholarship.org/uc/item/6nk296hc

Section 4 of this dissertation provides a method for uncertainty quantification for a multi-class setting. This method utilizes Bayesian statistics where a the unlabeled vertices adopt a Gaussian prior distribution. The author introduces a quantity $S(i)$ that represents the total measure of uncertainty averaged across all of the classes. Data points with low confidence scores are then randomly sampled to be labeled, but because data with the lowest confidence scores are often outliers, no data is sampled from the lowest 10% of confidence scores. This uncertainty quantification method is run on MNIST and EgoSeg, and it is shown that using the uncertainty quantification method to pick which points are sent to the oracle in an active learning setting outperforms randomly selecting points.

[10] H. Li, H. Chen, M. Haberland, A. L. Bertozzi, and P. J. Brantingham, "PDEs on graphs for semi-supervised learning applied to first-person activity recognition in body-worn video," *Discrete and Continuous Dynamical Systems*, vol. 41, no. 9, pp. 4351–4373, 2021.

This paper develops a PDE-based method that can classify body-worn videos that only requires a modest amount of training to achieve effectiveness and can protect privacy. The motion descriptor first uses the Horn-Schunk method to solve for the optical flow field and then places the optical flow vectors in each of the $s_x \times s_y \times s_t$ number of $dx \times dy \times dt$ volumes into eight histogram bins based on the orientation ($\frac{\pi}{4}$ per bin); to reduce the dimension of the system X of size $m \times n = 8s_x s_y \times s_t$, non-negative matrix factorization (NMF) is used to factorize it into a product of matrices V of size $m \times \hat{k}$ and H of size $\hat{k} \times n$ (as opposed to PCA in prior works) with $\hat{k} = 50$ producing reasonably good results and fixing V for each X further reduces computational costs. The classification step is based on minimizing the graph total variation balanced by a least-squares data fidelity term, which uses Ginzburg-Landau relaxation to construct a sequence of functions that approaches the minimizer. The optimization step involves solving a force-driven heat equation for $u^{k+\frac{1}{2}}$ and thresholding $u^{k+\frac{1}{2}}$ to obtain u^{k+1}, which is similar to the MBO scheme. The numerical method is coupled with the Nyström extension, which can approximate the $n \times n$ weight matrix W with number of data points $\ll n$. Experiments demonstrate performance better than most methods and significantly less training data required (6% vs. $\sim 60\%$).

[11] Y. Ma, R. Garnett, and J. Schneider, "Σ-optimality for active learning on Gaussian random fields," in *Advances in Neural Information Processing Systems*, vol. 26, 2013.

On Gaussian random fields (GRFs), Σ-optimality minimizes the survey risk $R_\Sigma(\boldsymbol{l}) = \mathbf{1}^T L_{\boldsymbol{u}}^{-1} \mathbf{1}$ as opposed to V-optimality which minimizes the L^2 risk $R_V(\boldsymbol{l}) = \mathrm{tr}(L_{\boldsymbol{u}}^{-1})$, with $\boldsymbol{l}$ being the set of all labeled points and $\boldsymbol{u}$ that of all unlabeled one. The former is more directly minimizing the 0/1 classification risk. This paper explores the greedy sequential application

of the two kinds of optimality. It is observed that both select nodes that have high variance and are highly correlated to high-variance nodes, whereas Σ-optimality additionally favors nodes that have consistent golbal influence, i.e. that are more likely to be in cluster centers. A near-optimal guarantee for greedy applications of the two kinds of optimality can be shown either using the normalization, monotonicity and submodularity of $R_\delta(\boldsymbol{l}) := R(\emptyset) - R(\boldsymbol{l})$ or the *suppressor-free* condition of GRFs. Experiments show that Σ-optimality does outperform V-optimality and other methods such as uncertainty sampling, expected error reduction and random selection on the datasets tested.

[12] Z. Meng, E. Merkurjev, A. Koniges, and A. L. Bertozzi, "Hyperspectral image classification using graph clustering methods," *Image Processing On Line*, vol. 7, pp. 218–245, 2017.

From this paper we learn of new techniques to segment hyperspectral images effectively and efficiently, and without losing much information that comes along with typical dimension reduction techniques. The researchers here present two custom training models: one that is semi-supervised and one that is unsupervised. The SSL model is used when some amount of truthfully labeled data is given, while the USL model can train off of no given labels. Embedded within both of these algorithms is the greatly-emphasized addition known as the Nyström extension technique. This technique greatly reduces computation by approximating the eigendecomposition of the graph Laplacian, through the actual eigendecomposition of a much smaller system whose solution is then scaled up to the N dimensions of the original system. Furthermore, in order to make these algorithms more efficient so that they can scale well to the large amounts of data that hyperspectral videos contain, the authors parallelize their code so that separate computations can be carried out at the same time across components of advanced processing cores. This is done using OpenMP directive-based parallelism. Ultimately, the results they obtain show that the models are effective and accurate at segmenting hyperspectral data, and that with the

aid of OpenMP these models scale almost ideally to larger batches of hyperspectral data.

[13] E. Merkurjev, A. L. Bertozzi, and F. Chung, "A semi-supervised heat kernel pagerank MBO algorithm for data classification," *Communications in Mathematical Sciences*, vol. 16, no. 5, pp. 1241–1265, 2018.

This paper presents a graph-based algorithm for semi-supervised classification called the kernel pagerank MBO. Two different techniques are used to compute the pagerank which are using random walks and using series. The authors also experimented with a simple method that uses heat kernel pagerank directly as a classifier. These algorithms are compared with other state-of-the-art methods on benchmark data sets such as Two Moons, MNIST, LFR, and COIL. The result shows that the heat kernel pagerank MBO algorithm performed competitively if not better than those of the state-of-the-art procedures. The algorithm also performed very well even when the number of labeled nodes is very small. The computational complexity is linear in the number of nodes for sparse graphs; the algorithm requires O(rnK) operations in the worst case for a dense graph, for a fixed number of classes m, where n is the size of the data set, and r and K are variables that describe the number of and the maximum length of random walks, respectively.

[14] E. Merkurjev, J. Sunu, and A. L. Bertozzi, "Graph MBO method for multiclass segmentation of hyperspectral stand-off detection video," in *2014 IEEE International Conference on Image Processing (ICIP)*, 2014, pp. 689–693.

This paper addresses the challenge of detecting chemical plumes in hyperspectral image data using graphical learning models. This method represents the pixels in a hyperspectral image as vertices of a weighted similarity graph. The Nyström extension method is used to calculate eigenvectors and eigenvalues of the graph Laplacian in very computationally efficient times ($<$ 1 minute) despite the very large size of the graph Laplacian. A major advantage of the method introduced in this paper is that it does not require any preprocessing of the hyperspectral data. The paper uses an MBO

scheme on the hyperspectral images and shows that accurate results can be reached with using only 10 eigenvectors.

[15] K. Miller, H. Li, and A. L. Bertozzi, "Efficient graph-based active learning with probit likelihood via Gaussian approximations," *arXiv preprint arXiv:2007.11126*, 2020.

The work of Miller, Li, and Bertozzi introduce what they refer to as the Probit model that uses non-Gaussian likelihood functions and loss functions. They then compare this to the Harmonic Functions model, and Gaussian Regression model, when using several different acquisition functions. Some of these acquisition functions in the past could only be applied to Gaussian models, however our researchers employ Laplacian and Newtonian approximations so that non-gaussian models (Probit) can utilize them. One of the acquisition functions the researchers have the models use is their own new Model Change (MC) function which approximates the change from the addition of a new index and corresponding label for it. Experimentation done on MNIST and a synthetic checkerboard dataset shows that even though some methods (like MBR) slighly outperform MC, MC has very close accuracy results and is much more computationally efficient.

[16] Y. Qiao, C. Shi, C. Wang, H. Li, M. Haberland, X. Luo, A. M. Stuart, and A. L. Bertozzi, "Uncertainty quantification for semi-supervised multi-class classification in image processing and ego-motion analysis of body-worn videos," *Electronic Imaging*, vol. 2019, no. 11, pp. 264–1, 2019.

This paper proposes a uncertainty quantification (UQ) method for graph-based semi-supervised multi-class classification problems. The authors adopted a Bayesian approach and propose a graphical multi-class probit model along with an effective Gibbs sampling procedure. Moreover, they also propose a confidence measure for each data point that correlates with classification performance, which means data points with high confidence measure are more likely to be classified correctly. This uncertainty quantification algorithm is tested on a handwritten digit data set (MNIST data set) as well as a body-worn video data set (HUJI EgoSeg data set).

The results on the MNIST data set contrasts the differences between high and low confidence images. The proposed confidence score also shows a correlation with the classification performance in experiments on the HUJI EgoSeg data set.

[17] B. Settles, "Active learning literature survey," 2010.

This guide provides a comprehensive overview of all the knowledge and literature concerning Active Learning, and its related fields, in 2010. The central idea behind Active Learning (which it makes sure to emphasize right from the beginning) is that if the machine model is allowed to choose the data points that it will be trained on, then it will be able to reach a desired performance level using less data points, and consequently more quickly. This is important in situations where labeling data points is expensive, whether it be time-wise, effort-wise, or simply fiscally. And these instances of expensive labeling account for a decent amount of activities where machine learning would be very beneficial—instances such as speech pattern recognition and categorical classification.

[18] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

Here our authors wish to improve on grouping algorithms that work on images, through focusing on the global framework of the image first before partitioning the image into discrete groups. Rather than employing a simple minimized-cut cost strategy to partition the graph, our authors develop a formula where the cut cost is a fraction of the total edge connections between the rest of the graph. Thus we obtain the Normalized Cut formula $Ncut(A,B) = \frac{cut(A,B)}{assoc(A,V)} + \frac{cut(A,B)}{assoc(B,V)}$, where A and B are subsets of the graph's nodes V and $assoc(A,V)$ is the sum of the connections between A and all the nodes in the graph. This prevents a bias towards partitions of very small subsets of nodes that min-cut has. The authors then describe their grouping algorithm, which consists of creating a graph with edge weights based on brightness and distance, solving $(D-W)x = \lambda Dx$ (this is the optimization of

Ncut), then from that finding the second smallest eigenvalue to split the graph into two groups, and then recurse further if needed. Results show that actual computation of the algorithm optimizing the Normalized Cut on images seems to successfully partition the images from the "top-down" in a hierarchical fashion.

[19] J. Sunu, J.-M. Chang, and A. L. Bertozzi, "Simultaneous spectral analysis of multiple video sequence data for LWIR gas plumes," in *Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XX*, M. Velez-Reyes and F. A. Kruse, Eds., vol. 9088, International Society for Optics and Photonics. SPIE, 2014, pp. 239 – 246. [Online]. Available: https://doi.org/10.1117/12.2050149

This paper is trying to tackle the challenge of detection of chemical plumes in hyperspectral image data. One challenge is to improve spectral clustering results while the other is to incorporate temporal information. There are two techniques performed to tackle these challenges which are manifold denoising and multiframe Nystrom method. The dataset being used is collected at the Dugway Proving Ground by long wave infrared spectrometers about two kilometers away from gas plume releases, at an elevation of about 1300 feet. The result shows that manifold denoising successfully improve the quality of spectral algorithms applied to hyperspectral data. It allows for better delineation in clustering results, improving the ordering of graph Laplacian eigenvectors, and making connections between broken segments. While multiframe Nystrom method provides a nice continuous segmentation across different frames, allowing for the incorporation of temporal information.

[20] U. von Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, pp. 395–416, 2007.

This article first introduces the unnormalized graph Laplacian L and the normalized ones $L_{\text{sym}} = D^{-\frac{1}{2}}LD^{-\frac{1}{2}} = I - D^{-\frac{1}{2}}WD^{-\frac{1}{2}}$ and $L_{\text{rw}} = D^{-1}W$ and their properties. Then three spectral clustering algorithms are introduced and several points of view are presented, including graph cut, random walks and perturbation theory. The article also talks

about some practical details, including the behaviors of different similarity graphs such as the ϵ-neighborhood graph, k-nearest neighbors graph and k-nearest neighbors graph as well as how to choose from them, and the considerations when choosing the appropriate graph Laplacians.

[21] X. Zhu, Z. Ghahramani, and J. Lafferty, "Semi-supervised learning using Gaussian fields and harmonic functions," in *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, ser. ICML'03. AAAI Press, 2003, p. 912–919.

This is one of the first algorithms developed in graph based learning. Data is modeled graphically where data points that are closer together in Euclidean space are given larger weights between the two vertices. The paper introduces the harmonic energy function $\frac{1}{2}\sum_{i,j} w_{ij}(f(i)-f(j))^2$ and forms a Gaussian field in order to assign a probability distribution on f. The minimization of the harmonic energy function seeks to extend the known labeled vertices in order to assign labels to the unlabeled vertices. The method used in this paper, called Harmonic Energy Minimization, is shown to follow the data structure more than an algorithm like KNN would. The paper also recommends utilizing the class prior distribution in order to produce more accurate labels. At the time of publication (2003), this method yielded more accurate results on text and digit classification than other methods being used at the time.

[22] X. Zhu, J. Lafferty, and Z. Ghahramani, "Combining active learning and semi-supervised learning using Gaussian fields and harmonic functions," *Proceedings of the ICML-2003 Workshop on The Continuum from Labeled to Unlabeled Data*, 08 2003.

This paper show that for energy $E(y) = \frac{1}{2}\sum_{i,j} w_{ij}(y(i) - y(j))^2)$, the minimum energy function $f = \arg\min_{y|L=y_L}$ of the Gaussian field is harmonic, i.e. $\Delta f = 0$ on unlabeled data points U and $\Delta f = y_L$ on the labeled data points L. Harmonicity then gives the mean-value property, i.e. the value at each unlabeled node is the average of neighboring nodes, and the maximum principle, which further implies uniqueness of f. With Δ being the graph Laplacian,

we get $f_u = -\Delta_{uu}^{-1}\Delta_{ul}f_l$ with $f_l = y_L$. It can be shown that f is both the mode and the mean of the field, so the Bayes classification rule specifies that node i is labeled as class 1 if $f(i) > 0.5$ and 0 otherwise. Selecting the query k that minimized the expected estimated risk and using the fact that $y_u \sim \mathcal{N}(f_u, \Delta_{uu}^{-1})$, we arrive at $f_u^{+(x_k,y_k)} = f_u + (y_k - f_k)\frac{(\Delta_{uu}^{-1})_{\cdot k}}{(\Delta_{uu}^{-1})_{kk}}$, where $(\Delta_{uu}^{-1})_{\cdot k}$ is the k-th column of Δ_{uu}^{-1}. Experiments on classification of 20 newsgroups and that of handwritten digits show that the proposed method has better performance than random query and SVM, and the highlight is that very few labels (~ 5) can result in an accuracy very close to the asymptotic value (i.e. no significant improvements by adding more labeled data).