# Efficient Graph-Based Active Learning with Probit Likelihood via Gaussian Approximations

**Kevin Miller** millerk22@math.ucla.edu
**Hao Li** lihao0809@math.ucla.edu
**Andrea L. Bertozzi** bertozzi@math.ucla.edu
*Department of Mathematics*
*University of California, Los Angeles*
*Los Angeles, CA 90025, USA*

## Abstract

We present a novel adaptation of active learning to graph-based semi-supervised learning (SSL) under non-Gaussian Bayesian models. We present an approximation of non-Gaussian distributions to adapt previously Gaussian-based acquisition functions to these more general cases. We develop an efficient rank-one update for applying "look-ahead" based methods as well as model retraining. We also introduce a novel "model change" acquisition function based on these approximations that further expands the available collection of active learning acquisition functions for such methods.

**Keywords:** Active Learning, Graph-Based Semi-Supervised Learning, Machine Learning

## 1. Introduction

Active learning in semi-supervised learning (SSL) seeks to alleviate the issue of trying to train machine learning classifiers with few labeled data but ubiquitous unlabeled data. While there are a few different formulations of active learning, we focus on the *pool-based* active learning paradigm as opposed to online or streaming-based active learning (Settles, 2012). That is, the active learner has access to a fixed "pool" of unlabeled data points from which it can decide the next training point. We consider querying only a single point at a time, as opposed to *batch-mode* active learning (Hoi et al., 2008). Let $Z = \{1, 2, \dots, N\}$ index a set of input feature vectors $X = \{\mathbf{x}_i\}_{i=1}^N$ among which $\mathcal{L} \subset Z$ have known labels $\{y_j\}_{j\in\mathcal{L}}$. We assume the *binary classification* case, in which the labels reside in $y_j \in \{\pm 1\}$ (or $\{0, 1\}$). In pool-based active learning, most methods alternate between: (1) training a model given the current labeled data $\mathcal{L}, \{y_j\}_{j\in\mathcal{L}}$ and (2) choosing an active learning query point $k^*$ in the unlabeled set $\mathcal{U} = Z - \mathcal{L}$ according to an *acquisition function*. We can classify most methods into a few categories: uncertainty (Settles, 2012; Houlsby et al., 2011; Gal et al., 2017), margin (Tong and Koller, 2001; Balcan et al., 2006; Jiang and Gupta, 2019), clustering (Dasgupta and Hsu, 2008; Maggioni and Murphy, 2019), and look-ahead (Zhu et al., 2003b; Cai et al., 2013) acquisition functions. Active learning methods have been proposed for graph-based SSL models which use a similarity graph to represent the geometric relationships between points in the dataset, such as Gaussian Random Field (GRF) models (Zhu et al., 2003a; Bertozzi and Flenner, 2016; Bertozzi et al., 2018). Active learners implementing look-ahead expected risk (Zhu et al., 2003b; Jun and Nowak, 2016),

model posterior covariance (Ji and Han, 2012; Ma et al., 2013), and other measures of uncertainty (Kushnir and Venturi, 2020) have been produced for the GRF model of (Zhu et al., 2003a). The conditional distribution of this foundational GRF model is a harmonic function on the graph and hence is referred to as the *Harmonic Functions* (HF) model.

Our contributions are (1) provide a unifying framework for active learning in many graph-based SSL models, (2) introduce an adaptation of non-Gaussian Bayesian models to allow for efficient calculations previously done only on Gaussian models, and (3) introduce a novel "model change" active learning acquisition function built around our adaptation.

## 2. Graph-Based SSL Models

Consider input data $X$ with index set $Z$; we create a similarity graph $G(Z, W)$ with edge weights $W_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}_j) \geq 0$ calculated by a similarity kernel $\kappa$. In this paper we consider both the *unnormalized graph Laplacian matrix* $L_u = D - W$ or *normalized graph Laplacian matrix* $L_n = D^{-1/2}(D - W)D^{-1/2}$, where $D = \text{diag}(d_1, d_2, \ldots, d_N), d_i = \sum_{j \neq i} W_{ij}$ is the diagonal *degree matrix*. As $L = L_u, L_n$ are both positive semi-definite, then with $\tau > 0$, $L_\tau = \tau^{-2}(L + \tau^2 I)$ is positive definite and therefore $\mathcal{N}(0, L_\tau^{-1})$ is a well-defined Bayesian prior distribution. Define a real-valued function on the nodes of the graph $u : Z \to \mathbb{R}$, $\mathbf{u} \in \mathbb{R}^N$ whose values reflect the classification of the data points. In graph-based SSL, given the current labeled set $\mathcal{L}$, one seeks the solution to the optimization problem

$$\mathbf{u}^* = \underset{\mathbf{u} \in \mathbb{R}^N}{\arg\min} \frac{1}{2}\langle \mathbf{u}, L_\tau \mathbf{u} \rangle + \sum_{j \in \mathcal{L}} \ell(u_j, y_j) =: \underset{\mathbf{u} \in \mathbb{R}^N}{\arg\min}\, J_\ell(\mathbf{u}; \mathbf{y}), \tag{1}$$

where $\ell : \mathbb{R} \times \mathbb{R} \to [0, \infty)$ is a chosen loss function and $\mathbf{y} \in \mathbb{R}^{|\mathcal{L}|}$ is a vector of labels $y_j$. Common loss functions include $\ell(x, y) = (x - y)^2/2\gamma^2$ and $\ell(x, y) = -\log \Psi_\gamma(xy)$, where $\Psi_\gamma(t) = \int_{-\infty}^{t} \psi_\gamma(s) ds$ is the cumulative distribution function (CDF) of a log-concave probability density function (PDF) $\psi_\gamma(s)$.

This variational perspective has a probabilistic counterpart, from which Bayesian statistical methods can provide useful ways for devising well-principled acquisition functions. We can view the objective function in Eqn 1 as the negative log of an associated Bayesian posterior distribution. In the case of $\ell(x, y) = (x-y)^2/2\gamma^2$, we model the likelihood of observations $\mathbf{y}|\mathbf{u}$ by $\mathcal{N}(P\mathbf{u}, \gamma^2 I_{|\mathcal{L}|})$ where $P : \mathbb{R}^N \to \mathbb{R}^{|\mathcal{L}|}$ is the projection of $\mathbf{u}$ onto the labeled indices $\mathcal{L}$. This likelihood is Gaussian and therefore the posterior $\mathbb{P}(\mathbf{u}|\mathbf{y})$ is Gaussian $N(\mathbf{m}, C)$ with covariance $C = \left(L_\tau^{-1} + P^T P/\gamma^2\right)^{-1}$ and mean $\mathbf{m} = CP^T\mathbf{y}/\gamma^2$. We refer to this as the *Gaussian Regression* (GR) model. The Gaussian structure of this posterior distribution allows us to efficiently calculate the posterior mean and covariance, including look-ahead calculations. Although the prior is Gaussian, the posterior distribution for general loss functions $\ell$ is not necessarily Gaussian. The key idea behind our method is to approximate a non-Gaussian distribution with a suitable Gaussian distribution to exploit the efficient calculations of the look-ahead posterior mean and covariance. This more general formulation allows us to use more realistic models for classification than just regression. An example of such a non-Gaussian posterior occurs when the loss function is $\ell(x, y) = -\log \Psi_\gamma(xy)$. In this case, the likelihood is derived from the model $y_j = \text{Sign}(u_j + \eta_j)$, where $\eta_j \sim \psi_\gamma$ (Hoffmann et al., 2020). We refer to this as the *Probit* model.

Some common acquisition functions originally derived for Gaussian models are

- **MBR** (Zhu et al., 2003b) $k_{MBR} = \arg\min_{k \in \mathcal{U}} \mathbb{E}_{y_k|\mathbf{m}} \left[ \sum_{i=1}^{N} \mathrm{Err}(i, \mathbf{m}^{k,y_k}) \right]$
- **VOpt** (Ji and Han, 2012) $k_V = \arg\max_{k \in \mathcal{U}} \frac{1}{\gamma^2 + C_{k,k}} \|C_{:,k}\|_2^2$
- **ΣOpt** (Ma et al., 2013) $k_\Sigma = \arg\max_{k \in \mathcal{U}} \frac{1}{\gamma^2 + C_{k,k}} \langle \mathbb{1}, C_{:,k} \rangle$

where $\mathrm{Err}(i, \mathbf{m}^{k,y_k})$ is the estimated risk on the $i^{th}$ data point of the look-ahead mean $\mathbf{m}^{k,y_k}$. These acquisition functions are originally defined on the HF model (Zhu et al., 2003b), but have been generalized here to fit the GR model. To recover the HF model's acquisition functions, let $\gamma = 0$, $y_j \in \{0, 1\}$, and the posterior covariance $C$ be defined only on the unlabeled nodes per the conditional nature of the HF model.

### 2.1 Laplace Approximation of the Probit Model

*Laplace approximation* is a popular technique for approximating non-Gaussian distributions with a Gaussian distribution (Rasmussen and Williams, 2006). We approximate the Probit posterior with the Gaussian distribution:

$$\hat{\mathbb{P}}(\mathbf{u}|\mathbf{y}) = \mathcal{N}(\hat{\mathbf{u}}, \hat{C}), \;\; \hat{\mathbf{u}} = \operatorname*{arg\,min}_{\mathbf{u} \in \mathbb{R}^N} J_\ell(\mathbf{u}; \mathbf{y}), \;\; \hat{C} = \left( \nabla\nabla J_\ell(\mathbf{u}; \mathbf{y})|_{\mathbf{u}=\hat{\mathbf{u}}} \right)^{-1}. \tag{2}$$

The mean of this Gaussian distribution $\hat{\mathbf{u}}$ is the *maximum a posteriori* (MAP) estimator of the true Probit posterior. This Gaussian distribution is in a form in which we can apply adaptations of acquisition functions of GR and HF models, such as VOpt (Ji and Han, 2012), Σ-Opt (Ma et al., 2013), and MBR (Zhu et al., 2003b). The Laplace approximations of the GR and HF models are indeed themselves, because the mean and MAP estimator (i.e. mode) are the same for Gaussian distributions. Furthermore, this Laplace approximation of non-Gaussian posterior distributions incorporates labeling information that is not contained in the GR and HF models' covariance matrices.

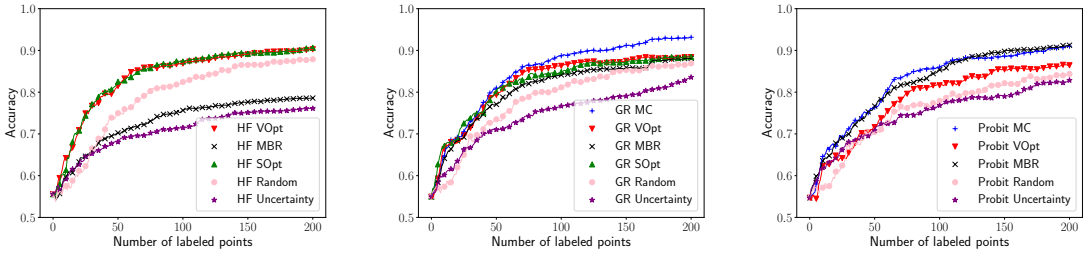
### 2.2 Look-Ahead Updates

Acquisition functions such as MBR need a *look-ahead model* with index $k$ and label $y_k$:

$$\operatorname*{arg\,min}_{\mathbf{u} \in \mathbb{R}^N} J^k(\mathbf{u}; \mathbf{y}, y_k) := \operatorname*{arg\,min}_{\mathbf{u} \in \mathbb{R}^N} \frac{1}{2} \langle \mathbf{u}, L_\tau \mathbf{u} \rangle + \sum_{j \in \mathcal{L}} \ell(u_j, y_j) + \ell(u_k, y_k).$$

This is simply the updated graph-based SSL problem, having added the index $k$ and associated label $y_k$ to the labeled data. As mentioned previously, one convenience of Gaussian models is that we can solve for the look-ahead posterior distribution's parameters from the current posterior distribution *without expensive model retraining*. This is a crucial property for computing acquisition functions like MBR (Zhu et al., 2003b), that consider the effects of adding an index $k$ with label $y_k$ to the labeled data. There is no simple, closed-form solution for computing the look-ahead MAP estimator $\hat{\mathbf{u}}^{k,y_k}$ from the current $\hat{\mathbf{u}}$ in the Probit model (Eqn 2) because of the loss function $-\ln \Psi_\gamma(xy)$. We approximate the look-ahead update $\tilde{\mathbf{u}}^{k,y_k}$ by computing a single step of Newton's Method on the look-ahead objective $J^k(\mathbf{u}; \mathbf{y}, y_k)$, starting with the current MAP estimator $\hat{\mathbf{u}}$:

$$\tilde{\mathbf{u}}^{k,y_k} = \hat{\mathbf{u}} - \left( \nabla\nabla J^k(\hat{\mathbf{u}}; \mathbf{y}, y_k) \right)^{-1} \left( \nabla J^k(\hat{\mathbf{u}}; \mathbf{y}, y_k) \right) = \hat{\mathbf{u}} - \frac{F(\hat{u}_k, y_k)}{1 + \hat{C}_{k,k} F'(\hat{u}_k, y_k)} \hat{C}_{:,k}, \tag{3}$$




(a) HF (b) GR (c) Probit

Figure 1: Checkerboard Dataset Results

where $F, F'$ are the first and second derivatives of the loss function with respect to the first argument. We call this single step of Newton's method as a *Newton Approximation (NA) update*. This is a simple rank-one update of the MAP estimator. The update requires storing the posterior covariance matrix $\hat{C}$; this is needed for all the aforementioned Gaussian-based acquisition functions, in this context. Due to the second-order nature of Newton's method, this NA update $\tilde{\mathbf{u}}^{k,y_k}$ empirically is a good approximation of the true look-ahead MAP estimator $\hat{\mathbf{u}}^{k,y_k}$. We also derive a *NA posterior covariance update* similar to the GR model:

$$\hat{C}^{k,y_k} = \left(\nabla\nabla J^k(\hat{\mathbf{u}}^{k,y_k}; \mathbf{y}, y_k)\right)^{-1} \approx \hat{C} - \frac{F'(\tilde{u}_k^{k,y_k}, y_k)}{1 + \hat{C}_{k,k} F'(\tilde{u}_k^{k,y_k}, y_k)} \hat{C}_{:,k} \hat{C}_{:,k}^T =: \tilde{C}^{k,y_k}. \tag{4}$$

With these simple NA updates, we can straightforwardly apply the Gaussian-based acquisition functions to our approximation (Eqn 2) of the Probit model. Furthermore, model retraining is approximated by using these NA updates of the MAP estimator and posterior covariance, as we demonstrate in Sec. 3.

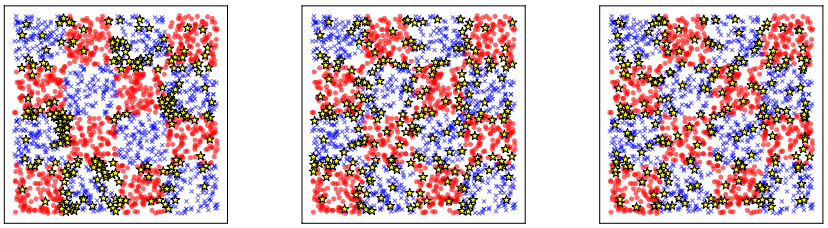
### 2.3 Model Change (MC) Acquisition Function

Calculating the approximate change in a model (i.e. classifier) from the addition of an index $k$ and associated label $y_k$ has been investigated previously (Cai et al., 2013; Karzand and Nowak, 2020). Employing our NA update (Eqn 3), we propose a MC acquisition function for our approximated Probit model in a *max-min* framework:

$$k_{MC-P} = \arg\max_{k \in \mathcal{U}} \min_{y_k \in \{\pm 1\}} \|\hat{\mathbf{u}} - \hat{\mathbf{u}}^{k,y_k}\|_2 \approx \arg\max_{k \in \mathcal{U}} \min_{y_k \in \{\pm 1\}} \left\| \frac{F(\hat{u}_k, y_k)}{1 + \hat{C}_{k,k} F'(\hat{u}_k, y_k)} \hat{C}_{:,k} \right\|_2 .$$

## 3. Results

We present numerical results demonstrating our Gaussian approximations and subsequent NA updates in the Probit model on a synthetic dataset (Checkerboard) and a real-world dataset (MNIST). In each of the **HF**, **GR**, and **Probit** models, we show the performance of the **MC** method of Sec. 2.3, **VOpt** (Ji and Han, 2012), **MBR** (Zhu et al., 2003b), **Uncertainty** (Settles, 2012), and **Random**. We calculate the average accuracies over five trials according to the underlying SSL classifier of the acquisition function. After comparing accuracies across all methods with a common classifier (of the Probit model), we find that each method's query choices better improve the accuracy of its underlying classifier. In Fig. 3, we demonstrate how closely the NA updates $\tilde{\mathbf{u}}^{k,y_k}, \tilde{C}^{k,y_k}$ (Eqns 3, 4) approximate the active learning choices from retraining the model (i.e. $\hat{\mathbf{u}}^{k,y_k}, \hat{C}^{k,y_k}$).



(a) HF-MBR (b) GR-MC (c) Probit-MC

(d) HF-Vopt (e) GR-VOpt (f) Probit-Uncertainty

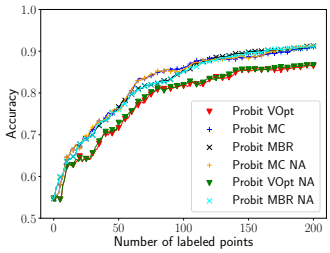
Figure 2: Acquisition function choices on the Checkerboard dataset. Yellow stars show the 200 points chosen by each of the given acquisition functions.

### 3.1 Checkerboard Dataset

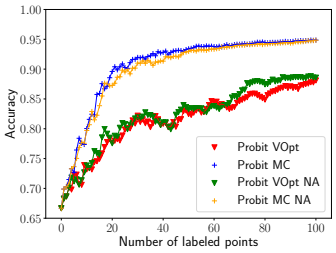
The Checkerboard dataset consists of $2,000$ points uniformly sampled on the unit square $[0,1]^2 \subset \mathbb{R}^2$, and we divide into two classes based on a $4 \times 4$ checkerboard pattern. For each of the five trials, we choose ten points uniformly at random to label initially (five from each class), and then sequentially choose 200 query points via our list of acquisition functions. Similar to (Kushnir and Venturi, 2020), we showcase this dataset because successful active learning in this dataset requires properly "exploring" the many different clusters as well as "exploiting" the learned decision boundaries efficiently. The best performing methods are the **MC** methods in the **GR** and **Probit** models, as well as **Probit-MBR**. These methods not only identify each of the clusters in the grid (Fig. 2b, 2c) but also explore the decision boundaries between clusters. In the **Probit-Uncertainty**(Fig. 2f) and **HF-MBR**(Fig. 2a), the methods have not explored the extent of the clustering structure and do not reach as high of accuracy (Fig. 1). Conversely, the **VOpt** acquisition function in each model only identifies points that are representative of each of the clusters. As seen in Figs. 2d and 2e, these acquisition functions have not explored the boundaries between the clusters and so do not achieve as high of accuracy.

### 3.2 MNIST

MNIST (Lecun et al., 1998) is a data set of 70,000 grayscale $28 \times 28$ pixel images of handwritten digits (0-9). Each image is represented by a 784-dimensional vector $\mathbf{x}_i$ and we normalize the pixel values to range from 0 to 1. We form a set of 4,000 data points by choosing uniformly at random 400 images from each digit. We construct a 15-nearest neighbor graph among the data points with weights $w_{ij} = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|_2^2/380^2)$. We consider the binary classification problem of classifying even digits versus odd digits. For each of the five trials, we start with ten initial training points evenly distributed between the two classes (not necessarily among the digits) and use the active learners to query 100 points. The average

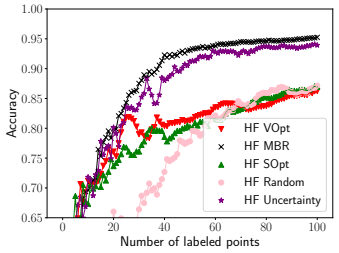



(a) Checkerboard

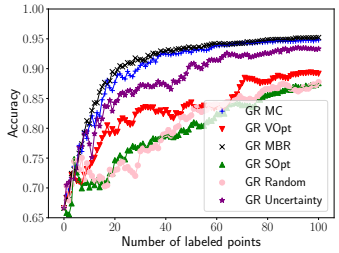



(b) MNIST

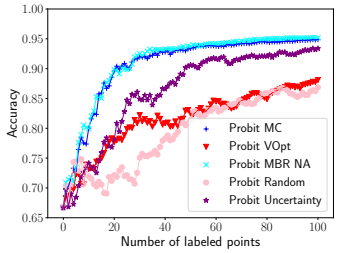
Figure 3: Accuracy comparison for query choices using the *true* posterior updates $\hat{\mathbf{u}}^{k,y_k}, \hat{C}^{k,y_k}$ compared to the NA updates $\tilde{\mathbf{u}}^{k,y_k}, \tilde{C}^{k,y_k}$. NA update denoted with "NA" in legend.




(a) HF




(b) GR




(c) Probit

Figure 4: MNIST Dataset Results

classification accuracies are presented in Fig. 4. Though the **MBR** methods perform the best, they are more costly to compute than our competitive **MC** acquisition functions.

## 4. Conclusion and Future Directions

Under this unifying Bayesian perspective of active learning in graph-based SSL, we use Laplace and Newton approximations to allow non-Gaussian models to employ acquisition functions previously only used in Gaussian models. We introduce a novel MC acquisition function that is both efficient to compute and provides competitive results. Future work could extend these results to batch-mode active learning, multi-class classification, and kernel methods other than graph-based SSL.

## Acknowledgments

KM is supported by the DOD's National Defense Science and Engineering Graduate (NDSEG) Fellowship, while HL and AB are supported by DARPA (grant FA8750-18-2-0066).

## References

Maria-Florina Balcan, Alina Beygelzimer, and John Langford. Agnostic active learning. In *Proceedings of the 23rd international conference on Machine learning*, ICML '06, pages

65–72, Pittsburgh, Pennsylvania, USA, June 2006. Association for Computing Machinery. ISBN 978-1-59593-383-6. doi: 10.1145/1143844.1143853. URL https://doi.org/10.1145/1143844.1143853.

Andrea L. Bertozzi and Arjuna Flenner. Diffuse Interface Models on Graphs for Classification of High Dimensional Data. *SIAM Review*, 2016. doi: 10.1137/16M1070426.

Andrea L. Bertozzi, Xiyang Luo, Andrew M. Stuart, and Konstantinos C. Zygalakis. Uncertainty quantification in graph-based classification of high dimensional data. *arXiv:1703.08816 [cs, stat]*, February 2018. URL http://arxiv.org/abs/1703.08816.

Wenbin Cai, Ya Zhang, and Jun Zhou. Maximizing Expected Model Change for Active Learning in Regression. In *2013 IEEE 13th International Conference on Data Mining*, pages 51–60, December 2013. doi: 10.1109/ICDM.2013.104. ISSN: 2374-8486.

Sanjoy Dasgupta and Daniel Hsu. Hierarchical sampling for active learning. In *Proceedings of the 25th international conference on Machine learning*, ICML '08, pages 208–215, Helsinki, Finland, July 2008. Association for Computing Machinery. ISBN 978-1-60558-205-4. doi: 10.1145/1390156.1390183. URL https://doi.org/10.1145/1390156.1390183.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep Bayesian active learning with image data. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pages 1183–1192, Sydney, NSW, Australia, August 2017. JMLR.org.

Franca Hoffmann, Bamdad Hosseini, Zhi Ren, and Andrew M. Stuart. Consistency of semi-supervised learning algorithms on graphs: Probit and one-hot methods. *arXiv:1906.07658 [cs, math, stat]*, March 2020. URL http://arxiv.org/abs/1906.07658.

Steven C.H. Hoi, Rong Jin, Jianke Zhu, and Michael R. Lyu. Semi-supervised SVM batch mode active learning for image retrieval. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–7, June 2008. doi: 10.1109/CVPR.2008.4587350. ISSN: 1063-6919.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian Active Learning for Classification and Preference Learning. *arXiv:1112.5745 [cs, stat]*, December 2011. URL http://arxiv.org/abs/1112.5745.

Ming Ji and Jiawei Han. A Variance Minimization Criterion to Active Learning on Graphs. In *Artificial Intelligence and Statistics*, pages 556–564, March 2012. URL http://proceedings.mlr.press/v22/ji12.html. ISSN: 1938-7228 Section: Machine Learning.

Heinrich Jiang and Maya Gupta. Minimum-Margin Active Learning. *arXiv:1906.00025 [cs, stat]*, May 2019. URL http://arxiv.org/abs/1906.00025.

Kwang-Sung Jun and Robert Nowak. Graph-Based Active Learning: A New Look at Expected Error Minimization. *arXiv:1609.00845 [cs, stat]*, September 2016. URL http://arxiv.org/abs/1609.00845.

Mina Karzand and Robert D. Nowak. MaxiMin Active Learning in Overparameterized Model Classes}. *arXiv:1905.12782 [cs, stat]*, April 2020. URL http://arxiv.org/abs/1905.12782.

Dan Kushnir and Luca Venturi. Diffusion-based Deep Active Learning. *arXiv:2003.10339 [cs, stat]*, March 2020. URL http://arxiv.org/abs/2003.10339.

Yann Lecun, Corinna Cortes, and Christopher C.J. Burges. The MNIST Database of Handwritten Digits, 1998. URL http://yann.lecun.com/exdb/mnist/.

Yifei Ma, Roman Garnett, and Jeff Schneider. $\sigma$-Optimality for Active Learning on Gaussian Random Fields. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2751–2759. Curran Associates, Inc., 2013. URL http://papers.nips.cc/paper/4951-optimality-for-active-learning-on-gaussian-random-fields.pdf.

Mauro Maggioni and James M. Murphy. Learning by Active Nonlinear Diffusion. *ArXiv*, 2019. doi: 10.3934/fods.2019012.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian processes for machine learning*. Adaptive computation and machine learning. MIT Press, Cambridge, Mass, 2006. ISBN 978-0-262-18253-9.

Burr Settles. Active Learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 6(1):1–114, June 2012. ISSN 1939-4608, 1939-4616. doi: 10.2200/S00429ED1V01Y201207AIM018. URL http://www.morganclaypool.com/doi/abs/10.2200/S00429ED1V01Y201207AIM018.

Simon Tong and Daphne Koller. Support Vector Machine Active Learning with Applications to Text Classification. *Journal of Machine Learning Research*, 2(Nov):45–66, 2001. ISSN ISSN 1533-7928. URL http://www.jmlr.org/papers/v2/tong01a.html.

Xiaojin Zhu, Zoubin Ghahramani, and John Lafferty. Semi-supervised learning using Gaussian fields and harmonic functions. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, ICML'03, pages 912–919, Washington, DC, USA, August 2003a. AAAI Press. ISBN 978-1-57735-189-4.

Xiaojin Zhu, John Lafferty, and Zoubin Ghahramani. Combining Active Learning and Semi-Supervised Learning Using Gaussian Fields and Harmonic Functions. In *ICML 2003 workshop on The Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining*, pages 58–65, 2003b.