Efficient Observation Selection in Probabilistic Graphical Models Using Bayesian Lower Bounds

Dilin Wang
Computer Science
Dartmouth College
dilin.wang.gr@dartmouth.edu

John Fisher III
CSAIL
MIT
fisher@csail.mit.edu

Qiang Liu
Computer Science
Dartmouth College
qiang.liu@dartmouth.edu

Abstract

Real-world data often includes rich relational information, which can be leveraged to help predict unknown variables using a small amount of observed variables via a propagation effect. We consider the problem of selecting the best subset of variables to observe to maximize the overall prediction accuracy. Under the Bayesian framework, the optimal subset should be chosen to minimize the Bayesian optimal error rate, which, unfortunately, is critically challenging to calculate when the variables follow complex and high dimensional probabilistic distributions such as graphical models. In this paper, we propose to use a class of Bayesian lower bounds, including Bayesian Cramér Rao bounds as well as a novel extension of it to discrete graphical models, as surrogate criteria for optimal subset selection, providing a set of computationally efficient algorithms. Extensive experiments are presented to demonstrate our algorithm on both simulated and real-world datasets.

1 INTRODUCTION

We consider the following optimal label selection problem: Given an unknown $\theta = [\theta_1, \dots, \theta_n]$ with posterior distribution $p(\theta \mid X)$ conditioning on observation X, select a best subset $C \subset [n]$ of size no larger than k on which the true values of θ_C are revealed, such that the prediction accuracy of $\theta_{\neg C}$ on the remaining set $\neg C = [n] \setminus C$ is maximized. We assume $p(\theta \mid X)$ to be a multivariate distribution with rich correlation structures, such as graphical models, so that the prediction of $\theta_{\neg C}$ can largely benefit from knowing θ_C via a "propagation effect".

Problems of this type appear widely in many important areas, including semi-supervised learning, experiment design, active learning, as well as application domains such as optimal sensor placement, and optimal budget allocation in crowdsourcing (e.g., Zhu et al., 2003; Krause & Guestrin, 2009; Settles, 2010; Bilgic et al., 2010; Liu et al., 2015).

The optimal subset C should be chosen to minimize certain uncertainty measures of the conditional model $p(\theta_{\neg C} \mid \theta_C \ ; \ X)$, and a natural choice is the conditional variance,

$$R_*(C) = \mathbb{E}_{\theta|X}\big(\sum_{i \in \neg C} \mathrm{var}(\theta_i \mid \theta_C \ ; \ X)\big),$$

which equals the mean squared error of the optimal Bayesian estimator of $\theta_{\neg C}$ given θ_C and X. Unfortunately, this objective is notoriously difficult to calculate in practice; Krause & Guestrin (2009) showed that it is #P-complete to calculate $R_*(C)$ for general discrete graphical models, even for simple tree structured models in many cases. Although practical approximations can be constructed using (approximate) posterior sampling via Markov chain Monte Carlo (MCMC), the estimation accuracy of the conditional variance can be poor when the size of C is large because each conditioning case only receives a small number of samples. Computationally, evaluating $R_*(C)$ requires both unconditioning sampling from $\theta \sim p(\theta \mid X)$, as well as conditional sampling from $\theta_{\neg C} \sim p(\theta_{\neg C} \mid \theta_C, X)$ for each value of θ_C that appears in the unconditioning sample; this makes it extremely difficult to optimize the conditional variance objective in practice, even when simple greedy search methods are used. Similar computational difficulty also appears in other uncertainty measures, such as conditional entropy and mutual information (Krause & Guestrin, 2009); what is worse, these information-theoretic objectives have the additional difficulty of depending on the normalization constant (known as the partition function) of $p(\theta \mid X)$, which is very challenging to calculate.

A special case when the computation can be largely simplified is when $p(\theta \mid X)$ is a multivariate Gaussian distribution; in this case, the conditional variance reduces a simple trace function of the inverse of the sub-matrix on $\neg C$, that is, $R_*(C) = \mathrm{tr}(Q[\neg C]^{-1})$, where Q is the inverse covariance (or the Fisher information) matrix of $p(\theta \mid X)$, and

$Q[\neg C]$ represents the sub-matrix of Q formed by the rows and columns in $\neg C$. This objective can be evaluated and optimized much more efficiently, because Q can be pre-calculated and the block-wise inversion can be calculated recursively using the block-wise matrix inversion formula (Horn & Johnson, 2012).

Contribution In this paper, we propose to solve the subset selection problem using information criteria of form $\mathrm{tr}(Q[\neg C]^{-1})$ for generic, non-Gaussian distribution $p(\theta \mid X)$, where Q is a generalized "information matrix" of $p(\theta \mid X)$ that we will define later; this is motivated by lower bounds of Bayesian risks of form

$$\mathbb{E}_{\theta|X}(\|\hat{\theta}_{\neg C} - \theta_{\neg C}\|_2^2) \geq \mathrm{tr}(Q[\neg C]^{-1}),$$

where $\hat{\theta}_{\neg C} = \hat{\theta}_{\neg C}(\theta_C, X)$ is any (deterministic or randomized) estimator of $\theta_{\neg C}$. Results of this type are the Bayesian version of the classical frequentist Cramér Rao bound, which, however, only works for unbiased estimators. For continuous θ with smooth densities, this bound is based on the van Trees inequality, or known as Bayesian Cramér Rao bound (Van Trees & Bell, 2007) and Q is a Bayesian version of the typical frequentist Fisher information matrix. For θ with discrete values, we derive a new form of Q based a new extension of van Trees inequality; our result appears to be the first bound of this type for discrete graphical models to the best of our knowledge.

Minimizing these Bayesian lower bounds provides new computationally efficient approaches for observation selection with complex $p(\theta \mid X)$. We provide extensive empirical results to demonstrate the advantage of our methods in two practical application settings, including selecting control questions in crowdsourcing and label propagation for graph-based semi-supervised learning.

Related Work Bayesian CR bounds have been widely used in signal processing and information fusion, but seem to be less well known in machine learning; we refer to Van Trees (2004); Van Trees & Bell (2007) for an overview of its theory and applications. Related to our work, Williams (2007) used the log-determinant (instead of the trace) of Bayesian Fisher information as the selection criterion, and studied its sub-modularity.

Outline This paper is organized as follows. Section 2 introduces backgrounds on the observation selection problem and Bayesian Cramér Rao (CR) bounds. We apply Bayesian CR bounds to solve the observation selection problem in Section 3 and propose the extension to discrete models in Section 4. We then discuss two examples of applications of our methods in Section 5, and present empirical results in Section 6. The paper is concluded in Section 7.

2 BACKGROUND

We introduce backgrounds on the observation selection problem in Section 2.1, and Bayesian Cramér-Rao bounds in Section 2.2. We restrict to the case when θ is a continuous variable in this section, and discuss the extension to discrete variables in Section 4.

2.1 OBSERVATION SELECTION

Assume $\theta = [\theta_1, \ldots, \theta_n] \in \mathbb{R}^n$ is a continuous random parameter of interest with posterior distribution $p(\theta \mid X) \propto p(X \mid \theta)p(\theta)$ conditioning on observed data X. We are interested in the setting when we have the option of revealing the true value θ_C of a subset $C \subset [n]$ of size no larger than k, such that we can get the best estimation on the unknown parameter $\theta_{\neg C}$ in the remaining set $\neg C = [n] \setminus C$. To be concrete, let $\hat{\theta}_{\neg C} = \hat{\theta}_{\neg C}(\theta_C, X)$ be an estimator of $\theta_{\neg C}$ based on θ_C and X, the optimal C should ideally minimize the mean squared Bayesian risk:

$$\min_{C\colon |C|\leq k} \big\{ R_{\hat{\theta}}(C) \equiv \mathbb{E}_{\theta|X}(\|\hat{\theta}_{\neg C} - \theta_{\neg C}\|_2^2) \big\}.$$

However, this objective depends on the choice of the estimator $\hat{\theta}_{\neg C}$ and is not easy to estimate in practice. Consider the Bayesian estimator $\hat{\theta}_{\neg C} = \mathbb{E}(\theta_{\neg C} \mid \theta_C,\ X)$, then $R_{\hat{\theta}}(C)$ reduces to the trace of the conditional variance:

$$R_*(C) = \mathrm{tr}(\mathbb{E}_{\theta|X}(\mathrm{cov}(\theta_{\neg C} \mid \theta_C, X))), \qquad (1)$$

which is also the minimum Bayesian risk one can possibly achieve. This objective function is called the A-optimality ("average" or trace) in the experiment design literature (e.g., Chaloner & Verdinelli, 1995).

There also exist other similar objective functions, but with less direct connection to the mean squared Bayesian risk; this includes the information-theoretic quantities such as the conditional entropy $H(\theta_{\neg C} \mid \theta_C,\ X)$ and the mutual information $I(\theta_{\neg C},\ \theta_C \mid X)$. The negative of these objective functions are often shown to be submodular and monotonic under certain conditions, for which an $(1-1/e)$ optimality approximation can be obtained using a simple greedy algorithm, that is, starting with an empty set $C = \emptyset$, and sequentially add the best item i so that $R(C \cup \{i\})$ is minimized (Nemhauser et al., 1978).

The major challenge in implementing the greedy algorithm for objectives like (1) is the computational cost of the objective. Although it is possible to draw approximate sample from $p(\theta \mid X)$ using MCMC, the estimation quality of the conditional variance $\mathrm{var}(\theta_{\neg C} \mid \theta_C)$ can be poor, especially when the size of C is large, because it requires samples from $\theta_{\neg C} \sim p(\theta_{\neg C} \mid \theta_C,\ X)$ for each value of θ_C that appears in the unconditioning sample of $p(\theta \mid X)$. Further, the Monte Carlo estimates are required for every candidate C considered, making the optimization algorithm very time consuming.

The information-theoretic objective functions, such as conditional entropy and mutual information, also suffer from the similar difficulty due to the need for estimating the conditional distribution $\log p(\theta_{\neg C} \mid \theta_C,\ X)$; in addition, they also involve calculating the normalization constant $Z = \int p(X \mid \theta)p(\theta)d\theta$, which is known to be critically difficult (e.g., Chen et al., 2012).

The computation can be largely simplified when $p(\theta \mid X)$ is a multivariate normal distribution, e.g., $\mathcal{N}(\mu, \Sigma)$, in which case the objective (1) reduces to a matrix function $\mathrm{tr}(Q[\neg C]^{-1})$, where $Q = \Sigma^{-1}$ is the inverse covariance matrix, and the greedy selection can be implemented efficiently based on the recursive relation,

$$R_*(C \cup \{i\}) = R_*(C) + \sum_{j \in \neg C} \frac{\sigma_{ij}^2}{\sigma_{ii}},$$

where $Q[\neg C]^{-1} = \{\sigma_{ij}\}$ and can also be calculated recursively using the block-wise matrix inversion formula (Horn & Johnson, 2012).

2.2 BAYESIAN CRAMÉR RAO LOWER BOUND

Bayesian Cramér Rao bounds (Van Trees, 2004), also known as van Trees inequalities, are lower bounds of Bayesian risks for any estimator $\hat{\theta}$ in terms of Fisher information matrix; it is the Bayesian version of the classical Cramér Rao bound, but does not restrict to unbiased estimators.

Let $\hat{\theta} = \hat{\theta}(X)$ be any (randomized or deterministic) estimator, then under mild regularity conditions (Van Trees & Bell, 2007, page 35), the Bayesian Cramér Rao bound guarantees

$$\mathbb{E}_{\theta \mid X}[\|\hat{\theta} - \theta\|^2] \geq \mathrm{tr}(H^{-1}), \tag{2}$$

where $H = -\mathbb{E}_{\theta \mid X}\left[\nabla_\theta^2 \log p(\theta \mid X)\right]$ and is called the *Bayesian Fisher information matrix*; compared to the classical Fisher information, Bayesian Fisher information takes expectation on the parameter θ and does not require a true value θ^*. We note that H can be rewritten into

$$H = -\mathbb{E}_\theta[\nabla_\theta^2 \log p(X \mid \theta)] - \mathbb{E}_\theta[\nabla_\theta^2 \log p(\theta)],$$

where the first term represents the information brought by the observed data, and the second term is the information from the prior knowledge.

Nuisance Parameter In many practical cases, there exist additional nuisance parameters $\eta \triangleq \{\eta_1, \cdots, \eta_{n'}\}$ of no direct interest. Ideally, this can be handled by applying Bayesian CR bound on the marginalized probability $p(\theta \mid X) = \int p(\theta, \eta \mid X)d\eta$. This, however, can be difficult to calculate because $\nabla_\theta^2 \log p(\theta \mid X)$ may have no closed form and require another Monte Carlo approximation. A weaker, but more computationally efficient, lower bound (Van Trees & Bell, 2007, Section 1.2.6) can be

$$\mathbb{E}_{\theta \mid X}[\|\hat{\theta} - \theta\|^2] \geq \mathrm{tr}([H^{-1}]_{\theta\theta}), \tag{3}$$

where $[H^{-1}]_{\theta\theta} = (H_{\theta\theta} - H_{\theta\eta}H_{\eta\eta}^{-1}H_{\eta\theta})^{-1}$ is the $\theta\theta$-submatrix of H^{-1}, with H being the joint Bayesian Fisher information matrix of $[\theta, \eta]$:

$$H = \begin{bmatrix} H_{\theta\theta} & H_{\theta\eta} \\ H_{\eta\theta} & H_{\eta\eta} \end{bmatrix} = \mathbb{E}_{\theta,\eta \mid X} \begin{bmatrix} \nabla_{\theta\theta}\ell & \nabla_{\theta\eta}\ell \\ \nabla_{\eta\theta}\ell & \nabla_{\eta\eta}\ell \end{bmatrix}, \tag{4}$$

where $\ell = -\log p(\theta, \eta \mid X)$.

3 BAYESIAN CR BOUND FOR LABEL SELECTION

We apply Bayesian CR bounds to define an objective function of form $\mathrm{tr}(Q[\neg C]^{-1})$ for the observation selection problems, allowing more efficient computation.

Proposition 1. *For any subset $C \subseteq [n]$ and estimator $\hat{\theta}_{\neg C} = \hat{\theta}_{\neg C}(\theta_C, X)$, assume the conditions for Bayesian Cramér Rao bound holds, we have*

$$\mathbb{E}_{\theta \mid X}[\|\hat{\theta}_{\neg C} - \theta_{\neg C}\|^2] \geq \mathrm{tr}(Q[\neg C]^{-1}),$$

where $Q[\neg C]$ is the submatrix of a matrix Q with rows and columns in $\neg C$ and Q can be one of the following two cases:

1. With no nuisance parameter, Q is the Bayesian Fisher information of θ, that is, $Q = -\mathbb{E}_{\theta \mid X}[\nabla_\theta^2 \log p(\theta \mid X)]$.

2. With a nuisance parameter η, we have $Q = H_{\theta\theta} - H_{\theta\eta}H_{\eta\eta}^{-1}H_{\eta\theta}$ and H is the joint Bayesian Fisher information of $[\theta, \eta]$ as defined in (4).

Proof. Apply (2) and (3) by treating $[\theta_C, X]$ as the fixed observation and $\theta_{\neg C}$ as the random parameter to be estimated. □

Remark Because the conditional variance in (1) is the Bayesian risk obtained by the Bayesian estimator $\hat{\theta}_{\neg C} = \mathbb{E}(\theta_{\neg C} \mid \theta_C;\ X)$, it should also be lower bounded by the Bayesian CR bound, that is,

$$\mathrm{tr}(\mathbb{E}_{\theta \mid X}(\mathrm{cov}(\theta_{\neg C} \mid \theta_C;\ X))) \geq \mathrm{tr}(Q[\neg C]^{-1}).$$

The above result suggests a method for finding the optimal subset C by minimizing the lower bound in Proposition 1, reducing to the observation selection problem to a sub-matrix selection problem:

$$\max_{C:\ |C| \leq k} \big\{ f_Q(C) \equiv -\mathrm{tr}(Q[\neg C]^{-1}) \big\}, \tag{5}$$

where k is the maximum size of C that defines our budget.

We now introduce conditions under which the objective function $f_Q(C)$ is a monotonically non-increasing and submodular function, so that the simple greedy selection algorithm yields an $(1 - 1/e)$-approximation. See Algorithm 1.

Proposition 2. *(i). Assume Q is positive definite. For any $i \notin C$, we have*

$$f_Q(C \cup \{i\}) = f_Q(C) + \frac{\sum_{j\in\neg C}\sigma_{ij}^2}{\sigma_{ii}}.$$

where σ_{ij} is the ij-element of $Q[\neg C]^{-1}$, and hence we have $f_Q(C) \geq f_Q(C')$ for any $C' \subseteq C$.

(ii). If Q is positive definite and also satisfies $Q_{ij} \leq 0$ for $i \neq j$ (i.e.,it is a Stieltjes matrix, equivalently a symmetric M-matrix), then $\Sigma = Q^{-1}$ is element-wise nonnegative, and $f_Q(C)$ is a sub-modular function.

Proof. (1) is an elementary fact, and (2) is a special case of Friedland & Gaubert (2013, Theorem 3). □

Since $\Sigma = Q^{-1}$ corresponds to the covariance matrix in the Gaussian case, Proposition 2(ii) suggests that we need Σ to be element-wise nonnegative, that is, θ_i are positive related to each other (in a rough sense), to make $f_Q(C)$ a submodular function. We remark that this element-wise positive condition is necessary; see Friedland & Gaubert (2013, Example 18) for a counter example. Similar "suppressor-free" conditions also appear when considering the submodularity of conditional variance functions in other settings (e.g., Das & Kempe, 2008; Ma et al., 2013).

The greedy algorithm for optimizing (5) is shown in Algorithm 1, in which we use Proposition 2(i) to reduce the greedy update $i^* = \arg\max_i f_Q(C \cup \{i\})$ to a simpler form:

$$i^* = \arg\max_i \left\{\sigma_{ii} + \frac{\sum_{j\in\neg C, j\neq i}\sigma_{ij}^2}{\sigma_{ii}}\right\}. \tag{6}$$

Intuitively, the first term of the above selection criterion corresponds to a *local effect*, representing the uncertainty σ_{ii} of θ_i itself, while the second term corresponds to a *global effect*, representing how much knowing the true value of θ_i can help in estimating the remaining parameters. Note that Algorithm 1 also updates $Q[\neg C]^{-1} = \{\sigma_{ij}\}$ recursively using the sub-matrix inverse formula (Line 9).

We should point out that evaluating the expectation in $Q = -\mathbb{E}[\nabla_\theta^2 \log p(\theta \mid X)]$ still requires drawing samples from $p(\theta \mid X)$ (or $p(\theta, \eta \mid X)$), but this can be pre-calculated before the greedy search starts, and is much more efficient than optimizing the exact conditional variance objective function, which requires expensive Monte Carlo or MCMC sampling for each candidate C evaluated during the optimization process.

4 EXTENSION TO DISCRETE VARIABLES

The Bayesian CR bound above works only for continuous random parameters, since it requires to calculate the derivatives and Hessian matrices. In this section, we introduce a new class of lower bounds that apply to general discrete probabilistic graphical models.

Algorithm 1 Greedy Subset Selection based on Bayesian CR bound

```
1: Input: Posterior distribution $p(\theta, \eta \mid X)$; budget size $k$.
2: Denote $H(\theta, \eta) = -\nabla^2_{[\theta,\eta]} \log p(\theta, \eta \mid X)$.
3: Draw sample $[\theta^\ell, \eta^\ell]_{\ell=1}^m \sim p(\theta, \eta \mid X)$.
4: $H = \frac{1}{m}\sum_\ell H(\theta^\ell, \eta^\ell)$ and $Q = H_{\theta\theta} - H_{\theta\eta}H_{\eta\eta}^{-1}H_{\eta\theta}$.
5: Initialize $C = \emptyset$. $\Sigma = Q^{-1}$.
6: while $|C| < k$ do
7:     $i^* \leftarrow \arg\max_{i\in\neg C} \sum_{j\in\neg C} \sigma_{ij}^2/\sigma_{ii}$ .
8:     $C \leftarrow C \cup \{i^*\}$.
9:     $\sigma_{ij} \leftarrow \sigma_{ij} - \sigma_{ii^*}\sigma_{i^*j}/\sigma_{i^*i^*}, \quad \forall i, j \in \neg C$.
10: end while
```

Proposition 3. *(i). Assume $\theta = [\theta_1, \ldots, \theta_n]$ takes values in a discrete set $\theta \in \{a_1, \ldots, a_d\}^n$, and $p(\theta|X) > 0$ for any θ. Let a^* be the solution of $\sum_{k=1}^d \frac{1}{a_k - a^*} = 0$. Define*

$$s_i(\theta,\ X) = \frac{1}{d(\theta_i - a^*)p(\theta_i \mid \theta_{\neg i}\ ;\ X)},$$

and $Q = \mathbb{E}_{\theta|X}[ss^\top]$, then for any estimator $\hat{\theta}(X)$, we have

$$\mathbb{E}_{\theta|X}[\|\hat{\theta}(X) - \theta\|^2] \geq \mathrm{tr}(Q^{-1}).$$

(ii). For any subset $C \subsetneq [n]$ and conditional estimator $\hat{\theta}_{\neg C} = \hat{\theta}_{\neg C}(\theta_C, X)$, we have

$$\mathbb{E}_{\theta|X}[\|\hat{\theta}_{\neg C} - \theta_{\neg C}\|^2] \geq \mathrm{tr}(Q[\neg C]^{-1}),$$

where $Q[\neg C]$ is the submatrix of Q with rows and columns in $\neg C = [n] \setminus C$.

Proof. (i). Denote by $\delta = \theta - \hat{\theta}$, we have by Cauchy's inequality,

$$\mathbb{E}_{\theta|X}[\delta\delta^\top] \succeq \mathbb{E}_{\theta|X}[\delta s^\top] \cdot [\mathbb{E}_{\theta|X}(ss^\top)]^{-1} \cdot \mathbb{E}_{\theta|X}[s\delta^\top].$$

Since $\|\hat{\theta}(X) - \theta\|^2 = \mathrm{tr}(\delta\delta^\top)$, we just need to show that $\mathbb{E}_{\theta|X}[\delta s^\top] = \mathbb{E}_{\theta|X}[(\theta - \hat{\theta})s^\top] = I$ where I is the identity matrix. To see this, note that

$$\begin{aligned}\mathbb{E}_{\theta|X}[s_i] &= \sum_\theta \frac{p(\theta_{\neg i} \mid X)}{d(\theta_i - a^*)} \\ &= \sum_{\theta_i} \frac{1}{d(\theta_i - a^*)} \sum_{\theta_{\neg i}} p(\theta_{\neg i} \mid X) = 0,\end{aligned}$$

where the last step is because $\sum_{\theta_i} \frac{1}{\theta_i - a^*} = 0$ by the definition of a^*. Therefore, we have $\mathbb{E}_{\theta|X}[s] = 0$, and hence $\mathbb{E}_{\theta|X}[\delta s^\top] = \mathbb{E}_{\theta|X}[\theta s^\top]$. Further, note that

$$\mathbb{E}_{\theta|X}[\theta_i s_i] = \mathbb{E}_{\theta|X}[(\theta_i - a^*)s_i] = \frac{1}{d}\sum_\theta p(\theta_{\neg i} \mid X) = 1,$$

$$\begin{aligned}\mathbb{E}_{\theta|X}[\theta_j s_i] &= \sum_{\theta} \frac{\theta_j - a^*}{d(\theta_i - a^*)} p(\theta_{\neg i} \mid X) \\ &= \sum_{\theta_i} \frac{1}{d(\theta_i - a^*)} \sum_{\theta_{\neg i}} (\theta_j - a^*) p(\theta_{\neg i} \mid X) \\ &= 0 \quad \forall i \neq j.\end{aligned}$$

This gives $\mathbb{E}_{\theta|X}[\theta s^\top] = I$ and the result follows.

(ii). Apply the result in (ii) by treating $\theta_{\neg C}$ as the random parameter and (θ_C, X) as the observed data. □

Note that s_i depends on $p(\theta \mid X)$ only through the conditional distribution $p(\theta_i \mid \theta_{\neg i},\ X)$, which is often computationally tractable since it does not depend on the troublesome normalization constant $Z = \sum_\theta p(X|\theta)p(\theta)$.

Example Consider the case of binary parameter $\theta \in \{0,1\}^n$, then solving $\frac{1}{0-a^*} + \frac{1}{1-a^*} = 0$ gives $a^* = 1/2$. Therefore, we have $s_i(\theta,\ X) = \frac{1}{(2\theta_i - 1)p(\theta_i|\theta_{\neg i},\ X)}$ in this case.

We remark that there exist variants of Bayesian CR bounds that use finite differences to replace the derivatives, including Borrovsky-Zakai bound (Bobrovsky & Zakai, 1975) and Weiss-Weinstein bound (Weiss & Weinstein, 1985); these bounds can be naturally applied when θ takes values in the integer lattice $\mathbb{Z}^n$, but does not work well when θ takes values a finite set due to the boundary problem.

5 APPLICATIONS

The subset selection problem has wide applications in many important areas. In this section, we describe two examples of applications that involve continuous and discrete random variables, respectively; empirical results on real datasets are presented in Section 6.

5.1 CONTINUOUS LABEL SELECTION FOR CROWDSOURCING

Crowdsourcing has been widely used in data-driven applications for collecting large amounts of labeled data (Howe, 2006). A major challenge, however, is that the (often anonymous) crowd labelers tend to give unreliable, even strongly biased, answers. Probabilistic modeling has been widely used to estimate the workers' reliabilities and downweight or eliminate the unreliable workers (e.g., Raykar et al., 2010; Karger et al., 2011; Zhou et al., 2012; Liu et al., 2012). However, to correct the biases, it is often necessary to reveal a certain amount of true labels, raising the problem of deciding which questions should be chosen to reveal the true labels (e.g., Liu et al., 2013, 2015).

To set up the problem, we follow the setting in Liu et al. (2013, 2015). Assume we have a set of questions $\{i\}$, each relates to an unknown continuous quantity θ_i that we want to estimate (e.g., price, point spreads, GDP). Let $\{j\}$ be a set of crowd workers that we hire to estimate $\{\theta_i\}$, and each worker j is characterized by a parameter $\eta_j = [b_j, v_j]$, where b_j and v_j represent the bias and variance of worker j, respectively; we assume the crowd label $\{x_{ij}\}$ of question i given by worker j is generated by

$$x_{ij} = \theta_i + b_j + \sqrt{v_j}\xi_{ij}, \quad \xi_{ij} \sim \mathcal{N}(0,1). \tag{7}$$

Using a Bayesian approach, we assume Gaussian priors $p(\theta_i) = \mathcal{N}(0, \sigma_\theta^2)$, $p(b_j) = \mathcal{N}(0, \sigma_b^2)$ on θ_i and b_j, and an inverse Gamma prior $p(v_j) = \text{Inv-Gamma}(\alpha, \beta)$ on v_j. The posterior distribution of θ and η can be written as

$$\begin{aligned}p(\theta, \eta \mid X) \propto \prod_j \exp\left[-\frac{b_j^2}{2\sigma_b^2}\right] \prod_j v_j^{-\alpha - \frac{d_j}{2} + 1} \exp\left[-\frac{\beta}{v_j}\right] \\ \prod_{i,j} \exp\left[-\frac{(X_{ij} - \theta_i - b_j)^2}{2v_j}\right] \prod_i \exp\left[-\frac{\theta_i^2}{2\sigma_\theta^2}\right].\end{aligned}$$

However, the crowd labels X may not carry enough information for predicting θ, and we hence consider the option of acquiring the ground truth labels of a subset C of questions (called the *control questions*), which can be incorporated into Bayesian inference to help evaluate the bias and variance of the workers, and hence improve the prediction of the remaining questions.

We can use Algorithm 1 to select the optimal subset C, where the greedy update (6) strikes a balance between selecting the most uncertain questions to myopically improve the overall MSE, and the most "influential" questions (e.g., these labeled by a lot of workers) whose ground truth labels can significantly improve the estimation of the workers' bias and variance, and hence improve the prediction of the unlabeled questions via a *propagation effect*.

5.2 DISCRETE LABEL SELECTION ON GRAPHS

Numerous real-world applications produce networked data with rich relational structures, such as web data and communication networks, and these relational information can be used to improve the prediction accuracy of unlabeled data using a small amount of labeled data. Various methods have been developed to exploit this effect, including graph-based semi-supervised learning (e.g., Zhu et al., 2003; Zhou et al., 2004) and collective, or graph-based, classification (e.g., Lu & Getoor, 2003). A related important question is how to select the best labeling subset to enable the best prediction on the remaining data.

We set up the problem using undirected graphical models. Assume $G = (V, E)$ is an undirected graph, and θ_i is a discrete label associated with node $i \in V$. It is common to model the posterior distribution using a pairwise graphical model,

$$p(\theta) \propto \exp\big[\sum_{(i,j)\in E} J_{ij}\theta_i\theta_j + \sum_{i\in V} h_i\theta_i\big], \tag{8}$$

where J_{ij} represents the correlation between θ_i and θ_j and h_i the local information of θ_i. We are interested in the problem of selecting the best subset $C \subseteq V$ so that the prediction accuracy based $p(\theta_{\neg C} \mid \theta_C)$ is maximized. In the semi-supervised learning settings, θ_i is often assumed to be a continuous variable, and $p(\theta)$ reduces to a simple Gaussian Markov random field. We instead assume θ to be discrete labels (e.g., $\theta \in \{-1, +1\}$) which is much more challenging to deal with. Our bound in Section 4 and Algorithm 1 (but with Q defined in Proposition 3) provide a novel tool for solving this problem efficiently.

6 EXPERIMENTS

We present experiments to better understand the performance of our proposed observation selection methods based on Bayesian lower bounds. To achieve this, we first illustrate our method using a toy example based on Gaussian mixture, and then apply our method to the two application areas described in Section 5, including selecting optimal control questions in crowdsourcing, as well as discrete label selection in graph-based classification.

6.1 CONTINUOUS VARIABLES

We test our method in the case when θ is a continuous variable, first on a toy Gaussian mixture model, and then on the model for selecting control questions in crowdsourcing. We implement our method `BayesianCRB(Gibbs)` as shown in Algorithm 1 with the sample $[\theta^\ell, \eta^\ell]_{\ell=1}^m$ generated using Gibbs sampler. In addition, the following baseline selection methods are compared:

`Random`, in which a random set C of size k is selected uniformly.

`BayesianOpt(Gibbs)`, which greedily minimizes the trace of the conditional variance in (1); to estimate the conditional variance we draw $[\theta^\ell, \eta^\ell]_{\ell=1}^m \sim p(\theta, \eta, |X)$ using Gibbs sampler, and then for each candidate set C evaluated during the greedy search, we further draw $[\theta_{\neg C}^{\ell,r}, \eta^{\ell,r}]_{r=1}^{m'} \sim p(\theta_{\neg C}, \eta, |\theta_C^\ell, X)$ using another Gibbs sampler, and estimate the objective in (1) by

$$\frac{1}{m(m'-1)} \sum_{\ell=1}^{m} \sum_{r=1}^{m'} \sum_{i \in \neg C} (\theta_i^{\ell,r} - \bar{\theta}_i^\ell)^2,$$

where $\bar{\theta}_i^\ell = \frac{1}{m'} \sum_{r=1}^{m'} \theta_i^{\ell,r}$. This method aims to minimize the Bayesian optimal risk, but is obviously much more expensive than our `BayesianCRB(Gibbs)` because it needs a large size m' of MCMC sample to get a good approximation for evaluating each candidate C, while it tends to degenerate significantly when m' is small.

`MaxVar(Gibbs)`, which greedily finds a subset C with the largest uncertainty in the sense of maximizing the variance $\sum_{i \in C} \mathrm{var}(\theta_i | X)$, instead of minimizing the conditional variance. The variance is estimated by the empirical variance using the MCMC samples $[\theta^\ell, \eta^\ell]_{\ell=1}^m$. This algorithm is computationally as fast as our `BayesianCRB(Gibbs)`, but does not consider the "propagation effect" that the information in θ_C can help improve the inference on $\theta_{\neg C}$.

`Laplacian`, which uses a Laplacian approximation to approximate the posterior $p(\theta, \eta \mid X)$ with a multivariate normal distribution (Liu et al., 2015), under which the objective (1) reduces to the matrix form in (5). This algorithm is the same as our Algorithm 1, except that the H in Line 4 is instead estimated by $H = H(\theta^*, \eta^*)$, where $[\theta^*, \eta^*]$ is the mode of the posterior distribution $p(\theta, \eta \mid X)$. Obviously, `Laplacian` would perform similarly to `BayesianCRB(Gibbs)` when the posterior $p(\theta, \eta | X)$ is close to normal, but would otherwise perform poorly, especially when $p(\theta, \eta \mid X)$ is multimodal.

6.1.1 Gaussian Mixture Model

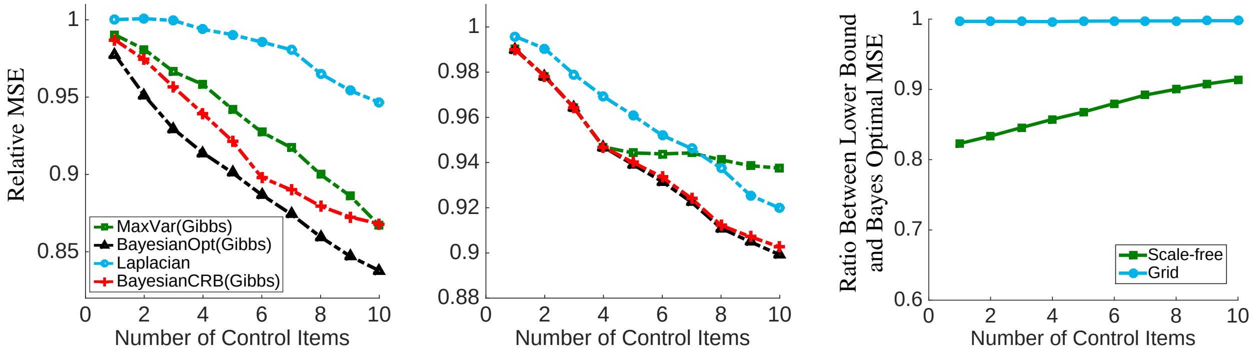
We start with the following toy example of Gaussian mixture model,

$$p(\theta) = \sum_{\kappa=1}^{2} \omega_\kappa \mathcal{N}(\theta \mid \mu_\kappa, \Sigma_\kappa),$$

where we ignore the dependence on observed data X. We draw μ_κ randomly from a zero-mean normal distribution with variance 0.1, and set $\Sigma_\kappa = 0.1(\alpha D_\kappa - W_\kappa)^{-1}$, where W_κ corresponds to an adjacency matrix of an undirected graph and D_κ is a diagonal matrix where $D_{\kappa,ii} = \sum_j W_{ij}$ and α is a constant larger than one to enforce L to be positive definite (we set $\alpha = 1.1$). We consider two different graph structures: (1) both W_1 and W_2 are the 30 by 30 2D grid graph, in which case we set $\omega = [1, 1]/2$; (2) W_1 and W_2 are scale-free networks of size 30 generated using the Barabási-Albert (BA) model (Barabási & Albert, 1999), with average degrees of 1 and 4, respectively, in which case we set $\omega = [0.9, 0.1]$. We simulate the ground truth of θ by drawing samples from $p(\theta \mid X)$, and plot the relative MSE of different algorithms compared to the random selection baseline in Figure 1(a)-(b); the results are averaged over 50 random trials.

As shown in Figure 1 (a)-(b), `BayesianOpt(Gibbs)` achieves the best performance, since it minimizes the conditional variance objective, which is the Bayesian optimal MSE. `Laplacian` performs the worst because $p(\theta)$ has multiple modes and the Laplacian approximation can only capture one of the mode. On the other hand, our `BayesianCRB(Gibbs)`, which takes the advantage of minimizing Bayesian CR bound, and is closer to `BayesianOpt(Gibbs)` than all the other methods.

It's also worth studying the tightness of the Bayesian CR bound. This is shown in Figure 1(c) where we plot the ratio



(a) *Relative MSE on scale-free graph* (b) *Relative MSE on grid graph* (c) *Tightness of the lower bounds*

Figure 1: (a-b) Results on the toy Gaussian mixture model; the y-axis are the MSE of different algorithms divided by the MSE of the random selection algorithm. (c) The ratio between the Bayesian CR lower bound and Bayesian optimal MSE (the trace of the conditional variance) on both the scale-free and grid graphs.

between the Bayesian optimal MSE (the trace of the conditional variance) and the Bayesian CR lower bound for both the scale-free and grid graphs; here the bounds are evaluated using the subsets C with different size k, selected by our `BayesianCRB(Gibbs)` method. We can see that the ratios are very close to one (≥ 0.8), suggesting that the bayesian CR bounds are very tight in these cases. In particular, we note that the bound is very tight for the grid graph example (ratio ≈ 1), explaining the good performances of `BayesianCRB(Gibbs)` in figure 1(b).

6.1.2 Application to Crowdsourcing

We further apply our method to the problem of selecting the optimal control questions in crowdsourcing as described in Section 5. We evaluate our selection algorithms on both simulated datasets and real-world datasets.

Toy dataset: We first generate a simulated dataset according to the Gaussian model described in (7), where θ_i and b_j are $i.i.d$ drawn from normal distribution with standard deviation of 4, and the labelers' variances v_j are generated from an inverse Gamma distribution Inv-Gamma$(1, 1)$. The dataset contains 30 questions and 30 labelers, and we assume the i-th question is answered only by the first i labelers; in this way, the first question is answered only by the first labeler and hence has the most *uncertain* result, and the last question is answered by all the 30 workers, and hence is the most *influential*, in that knowing its true value can help evaluate the bias and variance of all the 30 workers and hence improve the prediction on all the other items.

Figure 2(a) shows the average MSE given by the different methods. In this case, we can see that `BayesianOpt(Gibbs)`, `BayesianCRB(Gibbs)` and `Laplacian` tend to perform similarly, all of which significantly outperform `Random` and `MaxVar(Gibbs)`. Note that `MaxVar(Gibbs)` is even worse than `Random` at the beginning, since it myopically selects the most uncertain questions (the first few questions labeled by a small number of workers in this case), while much more significant improvements could be obtained by selecting the more influential items (these labeled by more workers). We find that `Laplacian` performs as well as `BayesianCRB(Gibbs)`, probably because the posterior distribution tends to be unimodal in this case. Figure 2 (b) shows the tightness of our lower bound as the size k of subset C increases, evaluated on the C given by `BayesianCRB(Gibbs)`, and we can see that the lower bound is again very tight in this case (ratio ≥ 0.93).

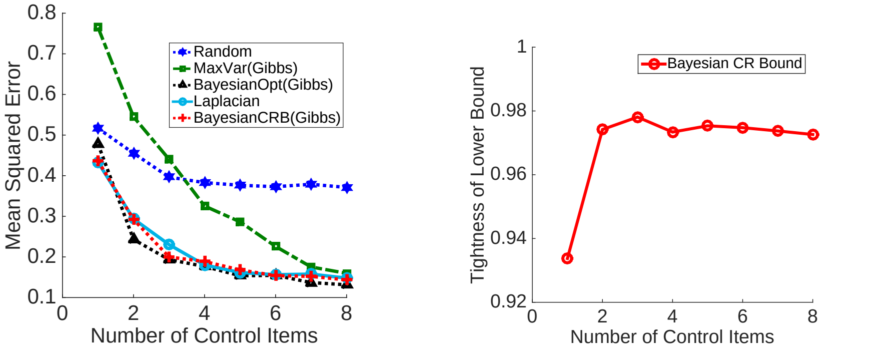
Real-world datasets: We also evaluate our approach on three real-world datasets:

The *PriceUCI* dataset (Liu et al., 2013). It consists of 80 household items collected from Internet, and whose prices are estimated by 155 UCI undergraduate students. As suggested in Liu et al. (2013), a log transform is performed on the prices before using the Gaussian models.

The national football league (*NFL*) forecasting dataset used in Massey et al. (2011). It consists predictions of point differences of 245 NFL games given by 386 participants; the point spreads determined by additional professional bookmakers are used as the ground truth.

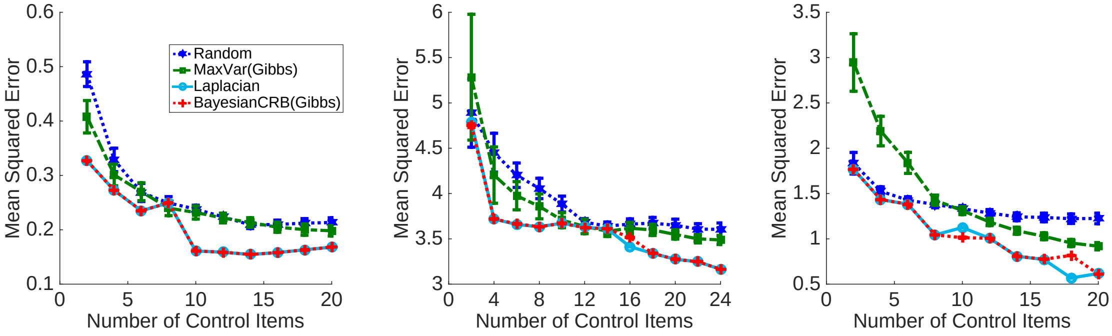
The *GDP Growth* dataset used in Budescu & Chen (2014). It contains the forecasts of GDP growth nine months ahead by professional forecasters surveyed by European Central Bank (ECB). A total of 98 forecasters made forecasts for 50 quarters of GDP growth.

The results on these three real-world datasets are shown in figure 3(a)-(c). `BayesianOpt(Gibbs)` is not evaluated because it is too slow on these real world datasets. We can see that both `BayesianCRB(Gibbs)` and `Laplacian` tend to outperform the other methods significantly. Again, `Laplacian` tends to perform as well as `BayesianCRB(Gibbs)` because the posteriors are very close to Gaussian distribution in these cases.



(a) *Average MSE on simulated dataset* (b) *Ratio between bounds and the Bayes optimal MSE*

Figure 2: Results on crowdsourcing with simulated data. (a) The MSE given by different selection algorithms as the budget k increases. (b) The ratio between the Bayesian lower bound and Bayesian optimal MSE. We can see that the lower bound is very close to the Bayesian optimal MSE (ratio ≥ 0.93), and the tightness tends to increase as the size k of C increases.



(a) *PriceUCI* (b) *NFL* (c) *GDP Growth*

Figure 3: (a)-(c) Results on three real datasets, *PriceUCI, NFL* and *GDP Growth*, respectively. The y-axes are the average MSE on the remaining items as the size of the subset C increases. The error bars show the standard deviation over the random trials.

6.2 DISCRETE VARIABLES

In this section, we use our algorithm to select optimal subsets on binary Ising graphical models defined in (8) with $\theta \in \{-1, +1\}$. We use `DiscreteLB(Gibbs)` to denote the greedy optimization algorithm on our discrete Bayesian lower bound (it is the same as Algorithm 1, except with Q defined in Proposition 3). We again compare our algorithm with several baselines, including:

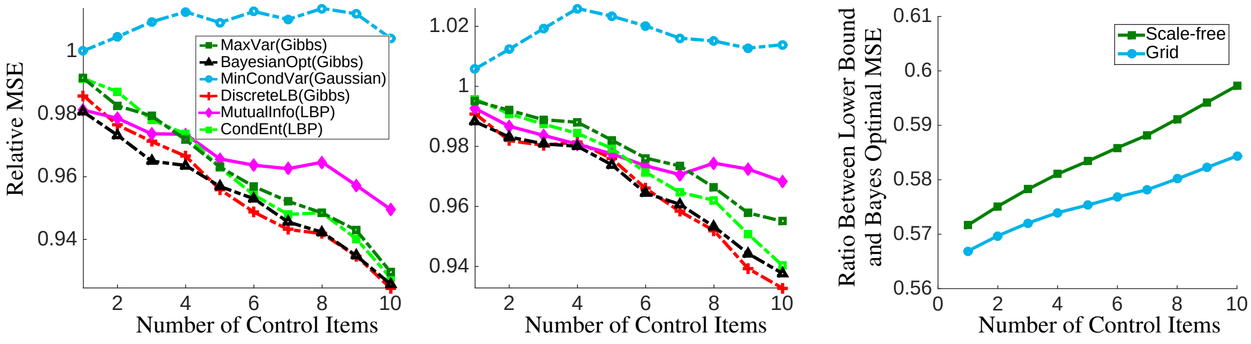
`CondEnt(LBP)`, which greedily selects a subset C to minimum the conditional entropy $H(\theta_{\neg C} \mid \theta_C, X)$; it is equivalent to maximizing the marginal entropy $H(\theta_C \mid X)$. The entropy is approximated using loopy belief propagation (LBP). This algorithm is similar to `MaxVar (Gibbs)` for continuous variables, in that both myopically find the subset with the largest uncertainty, ignoring the propagation effect that the added true labels can help predict the remaining unlabeled items (Krause et al., 2008).

`MutualInfo(LBP)`, which maximizes the mutual information $I(\theta_C; \theta_{\neg C} \mid X)$; this was proposed by Krause et al. (2008) to avoid the myopic property of the entropy objective. The mutual information is again approximated using loopy belief propagation (LBP).

`MinCondVar(Gaussian)`, which minimizes $\mathrm{tr}(L_{\neg C}^{-1})$ where $L = \Lambda - J$, where Λ is a diagonal matrix chosen to make L positive definite. This method is equivalent to treating θ as a continuous variable, and hence (8) a multivariate Gaussian distribution.

Comparisons are made on both simulated and real-world datasets:

Simulated data: We set $p(\theta \mid X)$ to be the binary graphical model in (8) (there is no actual observed data X in this case), with both J and h in (8) drawn from Gaussian distributions: we draw each element of h from $\mathcal{N}(0, 0.2)$, and set $J = 0.1W$, where W is an adjacency matrix of an undirected graph with values drawn from standard normal distribution. The graph structure is defined to be either a 30×30 2D grid, or a scale free network of size 30 generated using the Barabási-Albert (BA) model (Barabási & Albert, 1999) with the preferential attachment mechanism.



(a) *Relative error rate on scale-free graph* (b) *Relative error rate on grid graph* (c) *Tightness of the lower bounds*

Figure 4: Results on binary Ising models with simulated data. (a)-(b) The Relative MSE of different algorithms on the synthetic datasets simulated from the scale-free and the grid graph, respectively. (c) The ratio between our discrete Bayesian lower bound and the Bayesian optimal MSE on the simulated dataset; the bounds are evaluated on the subsets C selected by our `DiscreteLB(Gibbs)` method.

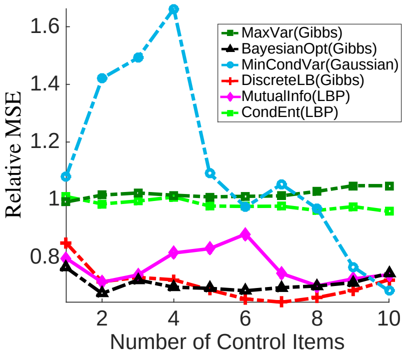


Figure 5: Comparison of the relative MSE of different methods on the PubMed Diabetes dataset. Our algorithm `DiscreteLB(Gibbs)` achieves similar results as `BayesianOpt(Gibbs)`, but with much lower computational cost.

We show the results of different algorithms in Figure 4, where we find that our `DiscreteLB(Gibbs)` is comparable with `BayesianOpt(Gibbs)` which minimizes the Bayesian optimal error rate, and outperforms all the other baselines. Both `MaxVar(Gibbs)` and `CondEnt(LBP)` tend to myopically select the most uncertain items first and hence have similar performance. Figure 4(c) shows the tightness of our Bayesian lower bound compared to the Bayesian optimal error; we can see that it is less tight (ratio $\geq$ 0.56) compared with the Bayesian CR bound for continuous variables, but the good performance of `DiscreteLB(Gibbs)` seems to suggest that the lower bound still represents a good surrogate for the Bayesian optimal error.

PubMed Diabetes:[1] This is a citation graph of scientific papers from the PubMed database (Sen et al., 2008), in which each paper is classified into one of three classes: "Diabetes Mellitus, Experimental", "Diabetes Mellitus Type 1", "Diabetes Mellitus Type 2". For our experiment, we select the top 100 nodes with the highest degrees from class Diabetes Mellitus Type 1 and Type 2, and then took the largest connected component, with 93 nodes in total and 376 edges; this gives 43 nodes from class Type 1 and 50 nodes from class Type 2, and a graph with an average degree of 4. In this case, since we don't have any prior knowledge, we set h to be a vector of small random numbers, and let $J = 0.05W$, where W denotes the adjacency matrix. The results is shown in Figure 5, in which we observe a similar trend as that in the simulated data.

7 CONCLUSION

The Bayesian optimal risk is the ideal objective function for optimal subset selection, which, however, is extremely difficult to calculate and optimize in practice. In this paper, we proposed to use Bayesian lower bounds as surrogate criteria, and derived a class of computationally more efficient algorithms for observation selection. We discussed both continuous and discrete scenarios: for continuous models, we based our bound on the classical Bayesian Cramér Rao bound; for discrete models, we derived a new form based on an novel extension of van Trees inequality. We presented a number of experiments for both continuous and discrete models in various practical settings, and showed that the selection algorithms based on the Bayesian lower bounds tend to outperform most baseline algorithms, and are comparable with the selection based on the Bayesian optimal risk.

Acknowledgement This work is supported in part by NSF CRII 1565796, and VITALITE, which receives support from Army Research Office (ARO) Multidisciplinary Research Initiative (MURI) program (Award number W911NF-11-1-0391).

[1] http://linqs.umiacs.umd.edu/projects//projects/lbc/

References

Barabási, A.-L. and Albert, R. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

Bilgic, M., Mihalkova, L., and Getoor, L. Active learning for networked data. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pp. 79–86, 2010.

Bobrovsky, B. Z. and Zakai, M. A lower bound on the estimation error for markov processes. *Automatic Control, IEEE Transactions on*, 20(6):785–788, 1975.

Budescu, D. V. and Chen, E. Identifying expertise to extract the wisdom of crowds. *Management Science*, 61(2):267–280, 2014.

Chaloner, K. and Verdinelli, I. Bayesian experimental design: A review. *Statistical Science*, pp. 273–304, 1995.

Chen, M.-H., Shao, Q.-M., and Ibrahim, J. G. *Monte Carlo methods in Bayesian computation*. Springer Science & Business Media, 2012.

Das, A. and Kempe, D. Algorithms for subset selection in linear regression. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pp. 45–54. ACM, 2008.

Friedland, S. and Gaubert, S. Submodular spectral functions of principal submatrices of a hermitian matrix, extensions and applications. *Linear Algebra and its Applications*, 438(10):3872–3884, 2013.

Horn, R. A. and Johnson, C. R. *Matrix analysis*. Cambridge university press, 2012.

Howe, J. The rise of crowdsourcing. *Wired magazine*, 14 (6):1–4, 2006.

Karger, D. R., Oh, S., and Shah, D. Iterative learning for reliable crowdsourcing systems. In *Advances in neural information processing systems*, pp. 1953–1961, 2011.

Krause, A. and Guestrin, C. Optimal value of information in graphical models. *Journal of Artificial Intelligence Research*, pp. 557–591, 2009.

Krause, A., Singh, A., and Guestrin, C. Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies. *The Journal of Machine Learning Research*, 9:235–284, 2008.

Liu, Q., Peng, J., and Ihler, A. Variational inference for crowdsourcing. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 701–709, 2012.

Liu, Q., Ihler, A. T., and Steyvers, M. Scoring workers in crowdsourcing: How many control questions are enough? In *Advances in Neural Information Processing Systems*, pp. 1914–1922, 2013.

Liu, Q., Ihler, A., and Fisher, J. Boosting crowdsourcing with expert labels: Local vs. global effects. In *Information Fusion (Fusion), 2015 18th International Conference on*, pp. 9–14. IEEE, 2015.

Lu, Q. and Getoor, L. Link-based classification. In *ICML*, volume 3, pp. 496–503, 2003.

Ma, Y., Garnett, R., and Schneider, J. σ-optimality for active learning on gaussian random fields. In *Advances in Neural Information Processing Systems*, pp. 2751–2759, 2013.

Massey, C., Simmons, J. P., and Armor, D. A. Hope over experience desirability and the persistence of optimism. *Psychological Science*, 22(2):274–281, 2011.

Nemhauser, G. L., Wolsey, L. A., and Fisher, M. L. An analysis of approximations for maximizing submodular set functions. *Mathematical Programming*, 14(1):265–294, 1978.

Raykar, V., Yu, S., Zhao, L., Valadez, G., Florin, C., Bogoni, L., and Moy, L. Learning from crowds. *The Journal of Machine Learning Research*, 11:1297–1322, 2010.

Sen, P., Namata, G. M., Bilgic, M., Getoor, L., Gallagher, B., and Eliassi-Rad, T. Collective classification in network data. *AI Magazine*, 29(3):93–106, 2008.

Settles, B. Active learning literature survey. *University of Wisconsin, Madison*, 52(55-66):11, 2010.

Van Trees, H. L. *Detection, estimation, and modulation theory*. John Wiley & Sons, 2004.

Van Trees, H. L. and Bell, K. L. Bayesian bounds for parameter estimation and nonlinear filtering/tracking. *AMC*, 10:12, 2007.

Weiss, A. J. and Weinstein, E. A lower bound on the mean-square error in random parameter estimation (corresp.). *Information Theory, IEEE Transactions on*, 31(5):680–682, 1985.

Williams, J. L. *Information theoretic sensor management*. PhD thesis, Massachusetts Institute of Technology, 2007.

Zhou, D., Bousquet, O., Lal, T. N., Weston, J., and Schölkopf, B. Learning with local and global consistency. *Advances in neural information processing systems*, 16(16):321–328, 2004.

Zhou, D., Platt, J., Basu, S., and Mao, Y. Learning from the wisdom of crowds by minimax entropy. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 2204–2212, 2012.

Zhu, X., Ghahramani, Z., Lafferty, J., et al. Semi-supervised learning using gaussian fields and harmonic functions. In *ICML*, volume 3, pp. 912–919, 2003.