Proceedings of the 2019 Winter Simulation Conference
N. Mustafee, K.-H.G. Bae, S. Lazarova-Molnar, M. Rabe, C. Szabo, P. Haas, and Y.-J. Son, eds.

OPTIMAL COMPUTING BUDGET ALLOCATION FOR BINARY CLASSIFICATION WITH NOISY LABELS AND ITS APPLICATIONS ON SIMULATION ANALYTICS

Weizhi Liu
Haobin Li
Loo Hay Lee
Ek Peng Chew

Department of Industrial Systems
Engineering and Management
National University of Singapore
1 Engineering Drive 2
Singapore, 117576, SINGAPORE

Hui Xiao

Department of Management Science
School of Statistics
Southwestern University of Finance
and Economics
Chengdu, 611130, CHINA

ABSTRACT

In this study, we consider the budget allocation problem for binary classification with noisy labels. The classification accuracy can be improved by reducing the label noises which can be achieved by observing multiple independent observations of the labels. Hence, an efficient budget allocation strategy is needed to reduce the label noise and meanwhile guarantees a promising classification accuracy. Two problem settings are investigated in this work. One assumes that we do not know the underlying classification structures and labels can only be determined by comparing the sample average of its Bernoulli success probability with a given threshold. The other case assumes that data points with different labels can be separated by a hyperplane. For both cases, the closed-form optimal budget allocation strategies are developed. A simulation analytics example is used to demonstrate how the budget is allocated to different scenarios to further improve the learning of optimal decision functions.

1 INTRODUCTION

Ranking and Selection (R&S) is a well-developed research area that aims at selecting the best of a finite set of alternatives whose performances are uncertain and can only be estimated with noisy simulation (Fu et al. 2015). Various budget allocation strategies are developed to solve different kinds of R&S problems such as single-objective problem (Chen et al. 2000; Glynn and Juneja 2004; Frazier 2014), multi-objective problem (Lee et al. 2010; Feldman and Hunter 2018; Li et al. 2018), constrained optimization (Lee et al. 2012; Hunter and Pasupathy 2013; Healey et al. 2014), feasibility determination (Szechtman and Yucesan 2008; Gao and Chen 2017) and input uncertainty (Lam 2016; Gao et al. 2017; Liu et al. 2017; Song and Nelson 2019).

Besides ranking problems which define a measure function to rank alternatives, few studies have been done for the other evaluation problems with stochastic noises. Ryzhov (2018) developed a local time method for the targeting and selection problem which intends to select the alternative whose mean performance matches a prespecified target as closely as possible. Other new problem classes worth studying for simulation optimization community might include but are not limited to the classification with noisy labels, and clustering with noisy queries. In this study, we investigate the classification with noisy labels or unreliable labels. This type of problem is commonly seen in the real world. For example, people might post their labeling task on the outsourcing platform like Amazon Mechanical Turk (https://www.mturk.com/) due to its cheap cost. However, the annotators might make mistakes and give the wrong labels for certain

978-1-7281-3283-9/19/$31.00 ©2019 IEEE

data records such that the data quality of training datasets for classification can be very low, and thus the classification models can be misled. Bertsimas et al. (2018) applied robust optimization techniques to the formulations of support vector machines, logistic regression, and decision trees when uncertainty exists in data features and labels. Promising out-of-sample accuracy of their method has been shown through high-dimensional and difficult classification problems. Li et al. (2018) proposed a noise-tolerant training algorithm through a meta-learning method that simulates actual training by generating synthetic noisy labels to avoid the overfitting to the specific noise. Natarajan et al. (2018) provided some unbiased estimators of a family of loss measures for cost-sensitive learning considering class-conditional random noise with known noise rates. Fawzi et al. (2016) developed theoretical bounds on the robustness of nonlinear classifiers facing random and adversarial noises. Menon et al. (2015) adopted mutually contaminated distribution framework to estimate class-probability whose labels are corrupted in some ways. Zhang et al. (2015) proposed an active learning framework to deal with imbalanced multiple noisy labeling problem by introducing label integration and instance selection procedures. Cavallanti et al. (2011) introduced efficient margin-based algorithms and showed the bounds on the convergence rate to Bayes risk by adopting Mammen-Tsybakov low noise condition and assuming linear label noise. Frénay and Verleysen (2014) provided a comprehensive survey for the classification in the presence of label noise including different types of label noise, their impacts and the algorithms dealing with label noise.

In this study, we revisit the binary classification with label noise problem from a budget allocation perspective. That is, given chance and resources to recollect the labels of certain corrupted data, how should we allocate the budget to different data points to improve classification accuracy. We proposed two closed-form asymptotic optimal budget allocation strategies that maximize the decay rate of probability of false classification for two problem settings with unknown and known classification structures based on the large deviation principle. With the successful development of budget allocation strategy for classification with noisy labels, it can be applied to many other research areas and practical problems. For example, one might utilize the method developed in this study to learn the optimal decision function for the simulation analytics which intends to identify the optimal decision under each scenario and learn their mapping to support real-time decision when the specific scenario is realized (Jin et al. 2019). A simple simulation analytic example of the M/M/1 queue is demonstrated in Section 4. It may also be extended to identify the Pareto/non-Pareto alternatives, or feasible/infeasible solutions. To some extents, critical scenarios that exert a significant impact on the learning of decision boundary can also be identified through the proposed method.

The remainder of this paper is organized as follows. The budget allocation problem for binary classification with noisy labels is presented in Section 2. Section 3 provides the derivation of asymptotic optimal budget allocation strategies for binary classification with unknown and known structures based on large deviation principle. Section 4 demonstrates how the simulation budget is allocated to different scenarios through a simulation analytics example. Section 5 summarizes this work and points out some future research directions.

2 PROBLEM STATEMENT

In this study, we consider the binary classification with noisy labels. That is to classify a set of data points into two mutually exclusive sets based on noisy observations. The set of data points to be classified are $\{(\boldsymbol{x}_i, y_i) : \forall i \in \mathcal{S}\}$ where the indices set $\mathcal{S} = \{1, 2, \cdots, m\}$. The features of data point i, i.e. $\boldsymbol{x}_i \in \mathbb{R}^d$, is deterministic and given, while the corresponding label $y_i \in \{0, 1\}$ is unknown and can be estimated by random variable Y_i which follows Bernoulli distribution with success probability p_i. Given a known threshold τ, data point i is labeled with $y_i = 0$ (or equivalently, $i \in \mathcal{S}_0$) if $p_i \in (0, \tau]$, otherwise data point i is labeled with $y_i = 1$ (or equivalently, $i \in \mathcal{S}_1$) if $p_i \in (\tau, 1]$. Denote the cardinality of set $\mathcal{S}_0$ as n, then the size of set $\mathcal{S}_1$ is $m - n$.

To reduce the label uncertainty of each data point, we can collect more information by observing n_i independent and identically distributed copies of Y_i which are denoted as $Y_{i1}, Y_{i2}, \cdots, Y_{in_i}$. Based on the law of large numbers, the sample average of success probability for data point i, i.e. $\hat{p}_i = \frac{1}{n_i}\sum_{k=1}^{n_i} Y_{ik}$, converges to true success probability p_i with probability one when the number of observations n_i goes to positive infinity. Then the estimated label $\hat{y}_i = 0$ if $\hat{p}_i \in (0,\tau]$; otherwise the estimated label $\hat{y}_i = 1$ if $\hat{p}_i \in (\tau,1]$. In this study, we consider two problem settings with known/unknown structures about the classification. The first is that there is no clear structure about the problem, and we can only determine data point i is estimated to be included in $\mathcal{S}_0$ if $\hat{y}_i = 0$ (or equivalently, $\hat{p}_i \in (0,\tau]$) and data point j is estimated to be included in $\mathcal{S}_1$ if $\hat{y}_i = 1$ (or equivalently, $\hat{p}_i \in (\tau,1)$). The second is that there exists a linear decision boundary $\hat{\boldsymbol{\beta}}^T\boldsymbol{x} = 0$ with intercept term that can separate data points with different labels into two half-planes, and relying on logistic regression, data point i is estimated to be in set $\mathcal{S}_0$ if $1/(1+\exp(-\hat{\boldsymbol{\beta}}^T\boldsymbol{x}_i)) \in (0,\tau]$; otherwise data point i is estimated to be in set $\mathcal{S}_1$ if $1/(1+\exp(-\hat{\boldsymbol{\beta}}^T\boldsymbol{x}_i)) \in (\tau,1]$.

However, it is usually the case that simulation budget is limited, and we cannot observe infinitely many trials of each data point to estimate its success probability as accurate as possible. The question worth asking is how to determine budget $n_i = \alpha_i n$ (α_i is the budget allocation proportion) allocated to data point $i, \forall i \in \mathcal{S}$ given limited budget n such that the probability of false classification (PFC) can converge to zero at the fastest decay rate. The PFC is the probability of the event that there exists a data point falsely classified into the wrong set. Based on the problem settings, we consider two types of false classification event defined in Definitions 1-2 which will be further used in Sections 3.3-3.4. Throughout this paper, we make several common assumptions which can be found in the previous papers (Glynn and Juneja 2004; Szechtman and Yucesan 2008; Li et al. 2018).

Definition 1 (False Classification Event with Unknown Structures) False Classification Event happens if and only if: 1) there exists a data point $i \in \mathcal{S}_0$ which is estimated to be falsely classified into set $\mathcal{S}_1$ (denoted as $i \hat{\in} \mathcal{S}_1$) due to $\hat{p}_i \in (\tau,1]$; 2) or there exists a data point $j \in \mathcal{S}_1$ which is estimated to be falsely classified into set $\mathcal{S}_0$ (denoted as $j \hat{\in} \mathcal{S}_0$) due to $\hat{p}_i \in (0,\tau]$.

Definition 2 (False Classification Event with Known Structures) Given a linear decision boundary $\hat{\boldsymbol{\beta}}^T\boldsymbol{x} = 0$ trained by logistic regression and observed datasets $\{(\boldsymbol{x}_i, \hat{y}_i) : \forall i \in \mathcal{S}\}$, False Classification Event happens if and only if: 1) there exists a data point $i \in \mathcal{S}_0$ which is estimated to be falsely classified into set $\mathcal{S}_1$ (denoted as $i \hat{\in} \mathcal{S}_1$) due to $1/(1+\exp(-\hat{\boldsymbol{\beta}}^T\boldsymbol{x}_i)) \in (\tau,1]$; 2) or there exists a data point $j \in \mathcal{S}_1$ which is estimated to be falsely classified into set $\mathcal{S}_0$ (denoted as $j \hat{\in} \mathcal{S}_0$) due to $1/(1+\exp(-\hat{\boldsymbol{\beta}}^T\boldsymbol{x}_j)) \in (0,\tau]$.

3 ASYMPTOTIC OPTIMAL BUDGET ALLOCATION STRATEGY

3.1 Decay Rate of Probability of False Classification

Based on Definitions 1-2, False Classification Event happens if and only if there exists a data point $i \in \mathcal{S}_0$ which is estimated to be falsely classified into set $\mathcal{S}_1$ or there exists a data point $j \in \mathcal{S}_1$ which is estimated to be falsely classified into set $\mathcal{S}_0$. Hence, the general PFC can be mathematically formulated as (1)

$$\text{PFC} = P(\cup_{i\in\mathcal{S}_0} i \hat{\in} \mathcal{S}_1 \bigcup \cup_{j\in\mathcal{S}_1} j \hat{\in} \mathcal{S}_0). \tag{1}$$

It is noteworthy that the closed-form expression of PFC is hard to derive, and we present its upper and lower bound instead in Theorem 1 based on Bonferroni inequality.

Theorem 1 (Bounds of PFC) Let $C = \max\{\max_{i\in\mathcal{S}_0} P(i \hat{\in} \mathcal{S}_1), \max_{j\in\mathcal{S}_1} P(j \hat{\in} \mathcal{S}_0)\}$, then $C \leq \text{PFC} \leq mC$ where m is the cardinality of set $\mathcal{S}_0 \cup \mathcal{S}_1$.

PFC converges to zero when the number of the budget allocated to each data point, i.e. n_i, goes to positive infinity. However, to achieve a promising finite time performance, we want to maximize the decay rate of PFC which evaluates how fast PFC decays to zero assuming the number of total simulation budget goes to positive infinity. Definition 3 gives the formal definition of decay rate of general PFC.

Definition 3 (Decay Rate of PFC) The decay rate of PFC is

$$\lim_{n\to\infty} -\frac{1}{n}\log \text{PFC} = \min\{\min_{i\in\mathcal{S}_0}\lim_{n\to\infty} -\frac{1}{n}\log P(i\hat{\in}\mathcal{S}_1), \min_{j\in\mathcal{S}_1}\lim_{n\to\infty} -\frac{1}{n}\log P(j\hat{\in}\mathcal{S}_0)\}.$$

Note that $\lim_{n\to\infty} -\frac{1}{n}\log P(i\hat{\in}\mathcal{S}_1)$ can be a function of not only α_i but also some $\alpha_j, j\in\mathcal{S}/\{i\}$ in some cases. However, in a myopic perspective, let us write $\lim\limits_{n\to\infty} -\dfrac{1}{n}\log P(i\hat{\in}\mathcal{S}_1) = R_i(\alpha_i)$ and also $\lim\limits_{n\to\infty} -\dfrac{1}{n}\log P(j\hat{\in}\mathcal{S}_0) = R_j(\alpha_j)$.

3.2 Optimality Conditions for Budget Allocation Problem

To derive the optimal budget allocation to each data points such that the decay rate of PFC can be maximized, the optimal budget allocation problem can be formulated as (2). That is to find the optimal budget allocation $\alpha_i, \forall i\in\mathcal{S}$ and maximum z that is upper bounded with the minimum of rate functions $R_i(\alpha_i), \forall i\in\mathcal{S}$.

$$\begin{aligned} \max_{z,\alpha_i} \quad & z \\ \text{s.t.} \quad & z \le R_i(\alpha_i), \forall i\in\mathcal{S}_0, \\ & z \le R_j(\alpha_j), \forall j\in\mathcal{S}_1, \\ & \sum_{i=1}^{m}\alpha_i = 1, \alpha_i \ge 0, \forall i\in\mathcal{S}. \end{aligned} \tag{2}$$

Depending on the structure of $R_i(\alpha_i), \forall i\in\mathcal{S}_0$ and $R_j(\alpha_j), \forall j\in\mathcal{S}_1$, budget allocation problem in (2) can be a convex optimization problem with unique optimal solutions. In this case, the Karush-Kuhn-Tucker condition (Karush 1939; Kuhn and Tucker 1951) is sufficient and necessary for the optimality of (2).

Let $\gamma_i \ge 0, \forall i\in\mathcal{S}_0$; $\eta_j \ge 0, \forall j\in\mathcal{S}_1$; $\lambda\in\mathbb{R}$ be the dual variables for the three types of constraints in (2). Then optimal budget allocation problem is equivalent to

$$\min_{z\in\mathbb{R},\alpha_i\ge 0} \quad \max_{\gamma_i\ge 0,\eta_j\ge 0,\lambda\in\mathbb{R}} -z + \sum_{i\in\mathcal{S}_0}\gamma_i(z - R_i(\alpha_i)) + \sum_{j\in\mathcal{S}_1}\eta_j(z - R_j(\alpha_j)) + \lambda(\sum_{i=1}^{m}\alpha_i - 1).$$

Based on the complementary slackness and the value of dual variables, KKT conditions of (2) can be simplified, and the optimality conditions of (2) are shown in Theorem 2.

Theorem 2 (Optimality Conditions) The optimal allocation solution to budget allocation problem in (2) are

1. $R_i(\alpha_i) = R_j(\alpha_j), \forall i,j\in\mathcal{S}$;
2. $\sum_{i=1}^{m}\alpha_i = 1$;
3. $\alpha_i \ge 0, \forall i\in\mathcal{S}$.

In the next two subsections, we will discuss two problem settings with unknown/known structures of the classification problem. The closed-form optimality conditions for both cases will be presented based on Theorem 2.

3.3 Binary Classification with Unknown Structures

In the first case, we consider the false classification event with unknown structures which is defined formally in Definition 1. The label of each $\boldsymbol{x}_i$ is determined by comparing its sample average of success probability $\hat{p}_i = \frac{1}{n_i}\sum_{k=1}^{n_i} Y_{ik}$ with the known threshold τ. $Y_{i1}, Y_{i2}, \cdots, Y_{in_i}$ are the independent and identically distributed copies of Y_i which are assumed to follow Bernoulli distribution with unknown success probability p_i.

In this case, the PFC in (1) can be rewritten as (3)

$$\text{PFC} = P(\cup_{i\in\mathcal{S}_0}\{\frac{1}{n_i}\sum_{k=1}^{n_i} Y_{ik} > \tau\}\bigcup\cup_{j\in\mathcal{S}_1}\{\frac{1}{n_j}\sum_{k=1}^{n_j} Y_{jk} \le \tau\}). \tag{3}$$

And the decay rate functions in Theorem 2 can be written as

$$R_i(\alpha_i) = \lim_{n\to\infty} -\frac{1}{n}\log P(\frac{1}{n_i}\sum_{k=1}^{n_i} Y_{ik} > \tau) = \alpha_i[\tau\log\frac{\tau}{p_i} + (1-\tau)\log\frac{1-\tau}{1-p_i}], \forall i \in \mathcal{S}_0,$$
$$R_j(\alpha_j) = \lim_{n\to\infty} -\frac{1}{n}\log P(\frac{1}{n_j}\sum_{k=1}^{n_j} Y_{jk} \le \tau) = \alpha_j[\tau\log\frac{\tau}{p_j} + (1-\tau)\log\frac{1-\tau}{1-p_j}], \forall j \in \mathcal{S}_1.$$

Therefore, based on Theorem 2, the optimality conditions for the binary classification with unknown structures are presented in Corollary 1. These results are the same with Szechtman and Yucesan (2008) which considers the simulation budget allocation for feasibility determination problem.

Corollary 1 (Optimality Conditions for Binary Classification with Unknown Structures) The optimal allocation solution for binary classification with unknown structures that maximize the decay rate of PFC is

$$\alpha_i = \frac{[\tau\log\frac{\tau}{p_i} + (1-\tau)\log\frac{1-\tau}{1-p_i}]^{-1}}{\sum_{j=1}^m[\tau\log\frac{\tau}{p_j} + (1-\tau)\log\frac{1-\tau}{1-p_j}]^{-1}}, \quad \forall i \in \mathcal{S}.$$

It is noteworthy that $f(p_i) = \tau\log\frac{\tau}{p_i} + (1-\tau)\log\frac{1-\tau}{1-p_i}$ is a convex function with global minima zero when $p_i = \tau$. Therefore, intuitively speaking, Corollary 1 suggests that more budget should be allocated to those data points that have success probability very close to the threshold τ.

3.4 Binary Classification with Known Structures

In the second case, we consider the false classification event with known structures which is defined formally in Definition 2. Specifically, known structures mean that there exists a linear hyperplane that can separate the data points with different labels into two disjoint half-planes. In this case, we utilize the logistic regression model to help classifications. The label of each $\boldsymbol{x}_i$ is determined by comparing it samples average of success probability with a known threshold. Given the dataset $\{(\boldsymbol{x}_i, \hat{y}_i)\}$, we train the logistic models with parameter $\boldsymbol{\beta}$ by maximizing its logarithm of posterior probability. Lemma 1 states the posterior distribution of $\boldsymbol{\beta}$ given a Gaussian prior by investigating Taylor series of the logarithm of $\boldsymbol{\beta}$'s posterior distribution evaluated at its maximum a posteriori probability (MAP) estimation and Laplace approximation method.

Lemma 1 (Posterior Distribution of Logistic Regression Parameters) Given a logistic regression model $y = 1/(1+\exp(-\boldsymbol{\beta}^T\boldsymbol{x}))$, suppose parameter $\boldsymbol{\beta}$ has Gaussian prior distribution $N(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0)$, then the logarithm of its posterior distribution can be written as (4)

$$\Psi(\boldsymbol{\beta}) = \log \prod_{i=1}^{m} [(\frac{1}{1+\exp(-\boldsymbol{\beta}^T \boldsymbol{x}_i)})^{\hat{y}_i} (\frac{1}{1+\exp(\boldsymbol{\beta}^T \boldsymbol{x}_i)})^{1-\hat{y}_i}] + \log \frac{\exp(-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0))}{\sqrt{(2\pi)^m \det(\boldsymbol{\Sigma}_0)}}. \tag{4}$$

And the posterior distribution of $\boldsymbol{\beta}$ can be approximated by Gaussian distribution $N(\boldsymbol{\beta}^*, \boldsymbol{\Sigma}_m)$ by Laplace Approximation where $\boldsymbol{\beta}^*$ and $\boldsymbol{\Sigma}_m$ can be derived according to (5).

$$\begin{aligned} \boldsymbol{\beta}^* &= \arg\max_{\boldsymbol{\beta}} \Psi(\boldsymbol{\beta}), \\ \boldsymbol{\Sigma}_m &= [-\frac{\partial^2 \Psi(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}]^{-1} = [\sum_{i=1}^{m} \boldsymbol{x}_i \boldsymbol{x}_i^T \frac{1}{(1+\exp(-\boldsymbol{\beta}^{*T}\boldsymbol{x}_i))(1+\exp(\boldsymbol{\beta}^{*T}\boldsymbol{x}_i))} + \boldsymbol{\Sigma}_0^{-1}]^{-1}. \end{aligned} \tag{5}$$

A data point i is classified into set $\mathcal{S}_0$ if $\frac{1}{1+\exp(-\frac{1}{n_i}\sum_{k=1}^{n_i} \boldsymbol{\beta}_k^T \boldsymbol{x}_i)} \in (0, \tau)$ where $\{\boldsymbol{\beta}_k : k = 1, 2, \cdots\}$ are the independent copies from Gaussian distribution $N(\boldsymbol{\beta}^*, \boldsymbol{\Sigma}_m)$ in Lemma 1, otherwise it is classified into set $\mathcal{S}_1$. In this case, the PFC in (1) can be rewritten as (6)

$$\text{PFC} = P(\cup_{i\in\mathcal{S}_0} \{\frac{1}{1+\exp(-\frac{1}{n_i}\sum_{k=1}^{n_i} \boldsymbol{\beta}_k^T \boldsymbol{x}_i)} > \tau\} \bigcup \cup_{j\in\mathcal{S}_1} \{\frac{1}{1+\exp(-\frac{1}{n_j}\sum_{k=1}^{n_j} \boldsymbol{\beta}_k^T \boldsymbol{x}_j)} \le \tau\}). \tag{6}$$

And the decay rate functions in Theorem 2 can be written as

$$\begin{aligned} R_i(\alpha_i) &= \lim_{n\to\infty} -\frac{1}{n} \log P(\frac{1}{1+\exp(-\frac{1}{n_i}\sum_{k=1}^{n_i} \boldsymbol{\beta}_k^T \boldsymbol{x}_i)} > \tau) \\ &= \alpha_i \frac{(\boldsymbol{\beta}^{*T}\boldsymbol{x}_i - \log\frac{\tau}{1-\tau})^2}{2\boldsymbol{x}_i^T [\sum_{l=1}^{m} \boldsymbol{x}_l \boldsymbol{x}_l^T \frac{1}{(1+\exp(-\boldsymbol{\beta}^{*T}\boldsymbol{x}_l))(1+\exp(\boldsymbol{\beta}^{*T}\boldsymbol{x}_l))} + \boldsymbol{\Sigma}_0^{-1}]^{-1} \boldsymbol{x}_i}, \quad \forall i \in \mathcal{S}_0, \\ R_j(\alpha_j) &= \lim_{n\to\infty} -\frac{1}{n} \log P(\frac{1}{1+\exp(-\frac{1}{n_j}\sum_{k=1}^{n_j} \boldsymbol{\beta}_k^T \boldsymbol{x}_j)} \le \tau) \\ &= \alpha_j \frac{(\boldsymbol{\beta}^{*T}\boldsymbol{x}_j - \log\frac{\tau}{1-\tau})^2}{2\boldsymbol{x}_j^T [\sum_{l=1}^{m} \boldsymbol{x}_l \boldsymbol{x}_l^T \frac{1}{(1+\exp(-\boldsymbol{\beta}^{*T}\boldsymbol{x}_l))(1+\exp(\boldsymbol{\beta}^{*T}\boldsymbol{x}_l))} + \boldsymbol{\Sigma}_0^{-1}]^{-1} \boldsymbol{x}_j}, \quad \forall j \in \mathcal{S}_1. \end{aligned}$$

Therefore, based on Theorem 2, the optimality conditions for the binary classification with known structures are presented in Corollary 2. Intuitively speaking, Corollary 2 suggests that more budget should be allocated to those data points $\boldsymbol{x}_i$ that are close to the decision boundary $\boldsymbol{\beta}^{*T}\boldsymbol{x} = \log\frac{\tau}{1-\tau}$ and with high variance $\boldsymbol{x}_i^T \boldsymbol{\Sigma}_m \boldsymbol{x}_i$.

Corollary 2 (Optimality Conditions for Binary Classification with Known Structures) The optimal allocation solution for binary classification with known structures that maximize the decay rate of PFC is

$$\alpha_i = \frac{2\boldsymbol{x}_i^T \boldsymbol{\Sigma}_m \boldsymbol{x}_i}{(\boldsymbol{\beta}^{*T}\boldsymbol{x}_i - \log\frac{\tau}{1-\tau})^2} / \sum_{j=1}^{m} \frac{2\boldsymbol{x}_j^T \boldsymbol{\Sigma}_m \boldsymbol{x}_j}{(\boldsymbol{\beta}^{*T}\boldsymbol{x}_j - \log\frac{\tau}{1-\tau})^2}, \quad \forall i \in \mathcal{S}.$$

where $\boldsymbol{\beta}^*$ and $\boldsymbol{\Sigma}_m$ can be calculated based on (5)

4 NUMERICAL EXPERIMENTS

In this section, we demonstrate how the simulation budget is allocated through a simulation analytics example of M/M/1 queueing system which intends to learn the best decision of customer given what she observed to facilitate future decision-making in real time. There are two types of customers with service time $T_1 \sim \exp(\mu_1)$ and $T_2 \sim \exp(\mu_2)$ respectively. Denote (z_1, z_2) as the number of type 1 and type 2 customers already in the system. Suppose a customer now arrives at the system and determines whether she should enter or leave the system by considering her maximum expected waiting time T and decision function. This decision function is defined as $P(z_1T_1 + z_2T_2 < T)$ for the case with unknown structures in Section 4.1, and $E[z_1T_1 + z_2T_2]$ for the case with known structures in Section 4.2. Let $\mu_1 = 1, \mu_2 = 2, T = 5$ and $\tau = 0.5$.

4.1 Binary Classification with Unknown Structures

In this setting, we do not assume the underlying classification structures. The customer will enter the system if $P(z_1T_1 + z_2T_2 < T) > \tau = 0.5$; otherwise she will leave the system. Instead of running simulation, the true Bernoulli success rate p (the probability that customer will enter the system based on simulations) for scenario (z_1, z_2) is

$$p = \begin{cases} 1 & z_1 = 0 \text{ and } z_2 = 0, \\ 1 - \exp(-\frac{\mu_1}{z_1}T) & z_1 \neq 0 \text{ and } z_2 = 0, \\ 1 - \exp(-\frac{\mu_2}{z_2}T) & z_1 = 0 \text{ and } z_2 \neq 0, \\ 1 - \exp(-\frac{\mu_1}{z_1}T) - \frac{\mu_1}{z_1}T\exp(-\frac{\mu_1}{z_1}T) & z_1 \neq 0 \text{ and } z_2 \neq 0 \text{ and } \frac{\mu_1}{z_1} = \frac{\mu_2}{z_2}, \\ 1 - \frac{\frac{\mu_1}{z_1}\exp(-\frac{\mu_2}{z_2}T) - \frac{\mu_2}{z_2}\exp(-\frac{\mu_1}{z_1}T)}{\frac{\mu_1}{z_1} - \frac{\mu_2}{z_2}} & z_1 \neq 0 \text{ and } z_2 \neq 0 \text{ and } \frac{\mu_1}{z_1} \neq \frac{\mu_2}{z_2}. \end{cases}$$

Figure 1 shows the contour plot of Bernoulli success probability and its variance, and how the simulation budget is allocated to different scenarios (z_1, z_2). The yellow regions have larger values compared to blue regions. Based on the "Contour of Rank of Simulation Budget Allocation", it can be concluded that more simulation budget are allocated to those scenarios that are close to the decision boundary (i.e. $P(z_1T_1 + z_2T_2 < T) = 0.5$). The reason is that the Bernoulli success probability of those scenarios is quite close to the threshold $\tau = 0.5$ and the corresponding variance is also quite large, and thus more budget is needed to avoid the false classification event for those scenarios.

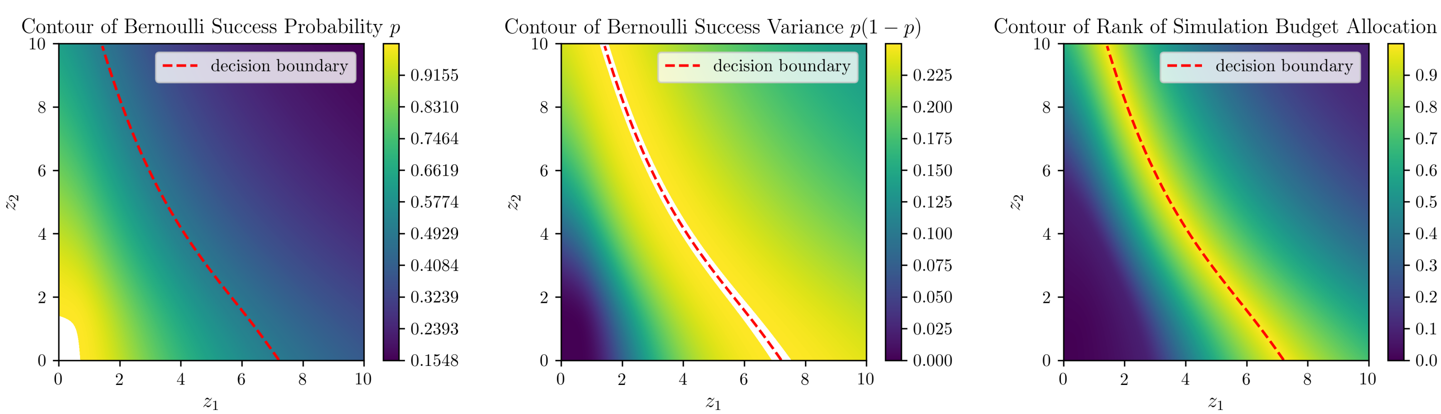

Figure 1: Demonstration of OCBA for Binary Classification with Unknown Structures.

4.2 Binary Classification with Known Structures

In this setting, we assume the scenarios (z_1, z_2) with different labels can be separated by a linear hyperplane $\boldsymbol{\beta}^T \boldsymbol{z} = \log \frac{\tau}{1-\tau} = 0$. The customer will enter the system if $E[z_1 T_1 + z_2 T_2] < T$; otherwise she will leave the system. Since we assume $\mu_1 = 1$, $\mu_2 = 2$ and $T = 5$, then the true decision boundary is $z_1 + 0.5 z_2 = 5$. Hence, $\boldsymbol{\beta} = (1, 0.5, -5)$ and $\boldsymbol{z} = (z_1, z_2, 1)$. Let the prior of covariance of $\boldsymbol{\beta}$ to be an identity matrix $\boldsymbol{I}_3$. Based on Corollary 2, Figure 2 shows the contour plot of $\boldsymbol{\beta}^T \boldsymbol{z}$ and $\boldsymbol{z}^T \boldsymbol{\Sigma}_m \boldsymbol{z}$, and how the simulation budget is allocated to different scenarios (z_1, z_2). Similar to the case with unknown structures, more simulation budget is allocated to the scenarios that are close to the linear decision boundary. More interestingly, the scenarios that lie in the two sides of the decision boundary receive more budget than those at the center of the decision boundary.

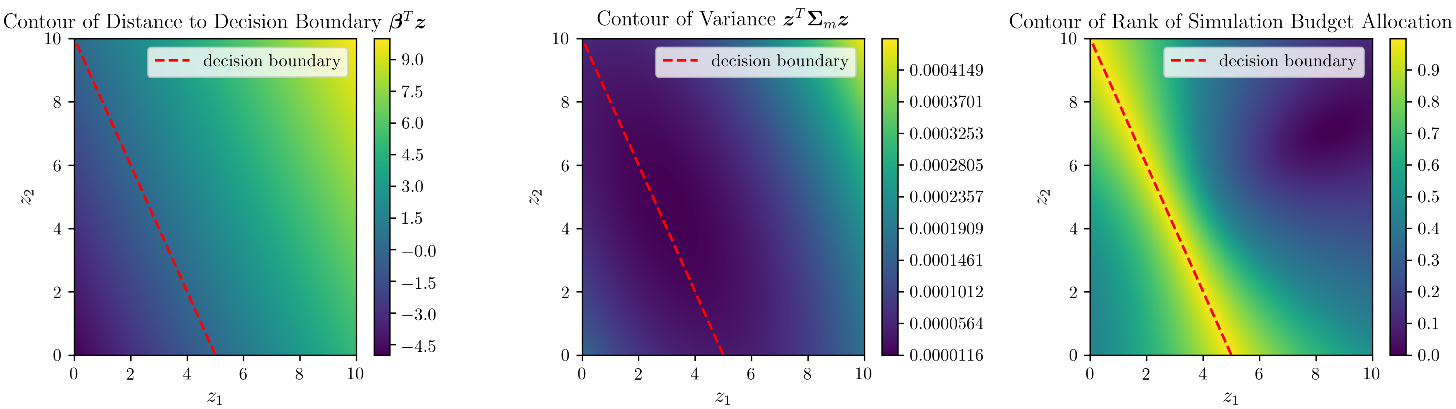

Figure 2: Demonstration of OCBA for Binary Classification with Known Structures.

5 CONCLUSIONS

In this study, we consider the budget allocation problem for the binary classification with noisy labels. Two types of problem settings are investigated with unknown and known classification structures. Two closed-form budget allocation strategies are proposed for both cases. Simulation analytics examples are conducted to demonstrate how the budget is allocated to different scenarios to decrease the probability of false classification at the fastest decay rate.

For future research, kernel tricks can be used to deal with nonlinear separable classification problems and generalized linear models can be used to extend the current work. This study can also be easily applied to multi-class classification with noisy labels by extending binary logistic regression to multinomial version, or use "One vs One" / "One vs Rest" / "Many vs Many" techniques. An online version of the budget allocation strategy integrated with classification methods will also be studied. Identification of support points of classification models is critical to maintain a minimum number of data storage. Model error caused by the incorrect selection of classifiers should be considered besides the heterogeneous noise of labels. More efficient sampling techniques for imbalanced datasets are also quite interesting and valued. Besides classification, the discussed methodology can also be applied to other machine learning tasks like regression and clustering.

ACKNOWLEDGMENTS

This paper is grateful for the support given by the Ministry of Education, Singapore [Grant MOE2015-T2-2-005] and the Centre of Excellence in Modelling and Simulation for Next Generation Ports (C4NGP), funded under the Singapore Maritime Institute (SMI) [Grant SMI-2017-SP-02].

REFERENCES

Bertsimas, D., J. Dunn, C. Pawlowski, and Y. D. Zhuo. 2018. "Robust Classification". *INFORMS Journal on Optimization* 1(1):2–34.

Cavallanti, G., N. Cesa-Bianchi, and C. Gentile. 2011. "Learning Noisy Linear Classifiers via Adaptive and Selective Sampling". *Machine Learning* 83(1):71–102.

Chen, C.-H., J. Lin, E. Yücesan, and S. E. Chick. 2000. "Simulation Budget Allocation for Further Enhancing the Efficiency of Ordinal Optimization". *Discrete Event Dynamic Systems* 10(3):251–270.

Fawzi, A., S.-M. Moosavi-Dezfooli, and P. Frossard. 2016. "Robustness of Classifiers: from Adversarial to Random Noise". In *Advances in Neural Information Processing Systems 29*, edited by D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, 1632–1640. Red Hook, New York: Curran Associates, Inc.

Feldman, G., and S. R. Hunter. 2018. "SCORE Allocations for Bi-Objective Ranking and Selection". *ACM Transactions on Modeling and Computer Simulation* 28(1):7:1–7:28.

Frazier, P. I. 2014. "A Fully Sequential Elimination Procedure for Indifference-Zone Ranking and Selection with Tight Bounds on Probability of Correct Selection". *Operations Research* 62(4):926–942.

Frénay, B., and M. Verleysen. 2014. "Classification in the Presence of Label Noise: A Survey". *IEEE Transactions on Neural Networks and Learning Systems* 25(5):845–869.

Fu, M., J. Hong, B. Nelson, J. Xu, C.-H. Chen, S. Chick, L. H. Lee, N. Pujowidianto, J. Kleijnen, M. Chau, S. Ghadimi, G. Lan, S. Kim, R. Pasupathy, S. Henderson, T. Homem-de Mello, G. Bayraksan, S. Andradóttir, Z. Zabinsky, J. Hu, and A. Gosavi. 2015. *Handbook of Simulation Optimization*, Volume 216. New York: Springer.

Gao, S., and W. Chen. 2017. "Efficient Feasibility Determination with Multiple Performance Measure Constraints". *IEEE Transactions on Automatic Control* 62(1):113–122.

Gao, S., H. Xiao, E. Zhou, and W. Chen. 2017. "Robust Ranking and Selection with Optimal Computing Budget Allocation". *Automatica* 81:30–36.

Glynn, P., and S. Juneja. 2004. "A Large Deviations Perspective on Ordinal Optimization". In *Proceedings of the 2004 Winter Simulation Conference*, edited by R. G. Ingalls, M. D. Rossetti, J. S. Smith, and B. A. Peters, 577–585. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Healey, C., S. Andradóttir, and S.-H. Kim. 2014. "Selection Procedures for Simulations with Multiple Constraints under Independent and Correlated Sampling". *ACM Transactions on Modeling and Computer Simulation* 24(3):14:1–14:25.

Hunter, S. R., and R. Pasupathy. 2013. "Optimal Sampling Laws for Stochastically Constrained Simulation Optimization on Finite Sets". *INFORMS Journal on Computing* 25(3):527–542.

Jin, X., H. Li, and L. H. Lee. 2019. "Optimal Budget Allocation in Simulation Analytics". In *15th IEEE Conference on Automation Science and Engineering*. August 22^{nd}–26^{th}, Vancouver, Canada.

Karush, W. 1939. *Minima of Functions of Several Variables with Inequalities as Side Condition*. Ph.D. thesis, Department of Mathematics, University of Chicago, Chicago, Illinois.

Kuhn, H. W., and A. W. Tucker. 1951. "Nonlinear Programming". In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, 481–492. Berkeley, California: University of California Press.

Lam, H. 2016. "Advanced Tutorial: Input Uncertainty and Robust Analysis in Stochastic Simulation". In *Proceedings of the 2016 Winter Simulation Conference*, edited by T. M. K. Roeder, P. I. Frazier, R. Szechtman, E. Zhou, T. Huschka, and S. E. Chick, 178–192. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Lee, L. H., E. P. Chew, S. Teng, and D. Goldsman. 2010. "Finding the Non-Dominated Pareto Set for Multi-Objective Simulation Models". *IIE Transactions* 42(9):656–674.

Lee, L. H., N. A. Pujowidianto, L.-W. Li, C.-H. Chen, and C. M. Yap. 2012. "Approximate Simulation Budget Allocation for Selecting the Best Design in the Presence of Stochastic Constraints". *IEEE Transactions on Automatic Control* 57(11):2940–2945.

Li, J., W. Liu, G. Pedrielli, L. H. Lee, and E. P. Chew. 2018. "Optimal Computing Budget Allocation to Select the Nondominated Systems–A Large Deviations Perspective". *IEEE Transactions on Automatic Control* 63(9):2913–2927.

Li, J., Y. Wong, Q. Zhao, and M. S. Kankanhalli. 2018. "Learning to Learn from Noisy Labeled Data". *CoRR* abs/1812.05214.

Liu, W., S. Gao, and L. H. Lee. 2017. "A Multi-Objective Perspective on Robust Ranking and Selection". In *Proceedings of the 2017 Winter Simulation Conference*, edited by W. K. V. Chan, A. D'Ambrogio, G. Zacharewicz, N. Mustafee, G. Wainer, and E. Page, 2116–2127. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Menon, A., B. Van Rooyen, C. S. Ong, and B. Williamson. 2015. "Learning from Corrupted Binary Labels via Class-Probability Estimation". In *Proceedings of the 32nd International Conference on Machine Learning*, edited by F. Bach and D. Blei, 125–134. July 7^{th}–9^{th}, Lille, France.

Natarajan, N., I. S. Dhillon, P. Ravikumar, and A. Tewari. 2018. "Cost-Sensitive Learning with Noisy Labels". *Journal of Machine Learning Research* 18(155):1–33.

Ryzhov, I. O. 2018. "The Local Time Method for Targeting and Selection". *Operations Research* 66(5):1406–1422.

Song, E., and B. L. Nelson. 2019. "Input–Output Uncertainty Comparisons for Discrete Optimization via Simulation". *Operations Research* 67(2):562–576.

Szechtman, R., and E. Yucesan. 2008. "A New Perspective on Feasibility Determination". In *Proceedings of the 2008 Winter Simulation Conference*, edited by S. J. Mason, R. R. Hill, L. Mönch, O. Rose, T. Jefferson, and J. W. Fowler, 273–280. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Zhang, J., X. Wu, and V. S. Shengs. 2015. "Active Learning with Imbalanced Multiple Noisy Labeling". *IEEE Transactions on Cybernetics* 45(5):1095–1107.

AUTHOR BIOGRAPHIES

WEIZHI LIU is a Research Fellow in the Centre of Excellence in Modelling and Simulation for Next Generation Ports (C4NGP). He received his Ph.D. degree in 2018 from the Department of Industrial Systems Engineering and Management (ISEM), National University of Singapore. His research interest includes simulation optimization and artificial intelligence with applications on logistics and maritime studies. His email address is weizhiliu@nus.edu.sg.

HAOBIN LI is an Assistant Professor (Research) at C4NGP and ISEM, NUS. He received his Ph.D. degree from the same department in 2014. He has research interests in operations research and simulation optimization with the application on logistics and maritime studies. His email address is li_haobin@nus.edu.sg.

HUI XIAO is an Associate Professor in the Department of Management Science, School of Statistics at Southwestern University of Finance and Economics, China. He holds a Ph.D. in ISEM from the National University of Singapore. His research interests include stochastic optimization, decision-making under uncertainty and reliability engineering. He is a member of IEEE and ORSC. His email address is msxh@swufe.edu.cn.

LOO HAY LEE is an Associate Professor at C4NGP and ISEM, NUS. He received Ph.D. degrees from Harvard University. His research interests include logistics, vehicle routing, supply chain modeling, simulation-based optimization, and evolutionary computation. His email address isiseleelh@nus.edu.sg.

EK PENG CHEW is an Associate Professor at C4NGP and ISEM, NUS. He received his Ph.D. degree from the Georgia Institute of Technology. His research interests include logistics and inventory management, system modeling and simulation, and system optimization. His email address is isecep@nus.edu.sg.