A Fair Empirical Risk Minimization with Generalized Entropy

Youngmi Jin\textsuperscript{1}, Jio Gim\textsuperscript{1}, Tae-Jin Lee\textsuperscript{2}, and Young-Joo Suh\textsuperscript{1}

\textsuperscript{1}Pohang University of Science and Technology (POSTECH) \\
\textsuperscript{2}Sungkyunkwan University

February 2, 2023

Abstract

This paper studies a parametric family of algorithmic fairness metrics, called generalized entropy, which originally has been used in public welfare and recently introduced to machine learning community. As a meaningful metric to evaluate algorithmic fairness, it requires that generalized entropy specify fairness requirements of a classification problem and the fairness requirements should be realized with small deviation by an algorithm. We investigate the role of generalized entropy as a design parameter for fair classification algorithm through a fair empirical risk minimization with a constraint specified in terms of generalized entropy. We theoretically and experimentally study learnability of the problem.

1 Introduction

As the use of machine learning algorithms grows in diverse areas such as criminal justice, lending and hiring, the issue of algorithmic fairness gets big attention. In response, a variety of work on algorithmic fairness has been proposed \cite{9,10,13,4,19,30,17,5,18,8}. Regarding algorithmic fairness, the most important questions would be "What is algorithmic fairness?" and "How do we quantify the degree of algorithmic fairness for a given algorithm?" This paper is more concerned with the second question than the first one.

Several works \cite{27,14,6} propose fairness measures to answer the second question. Their approaches are axiomatic ones in that they consider a set of axioms required as a reasonable measure for fairness and then find a function satisfying the set of axioms as a fairness metric. Cardinal social welfare (justice) in economics is applied for the proposal of new algorithmic fairness metrics in the works of \cite{14,6}. The work of \cite{27} propose to use generalized entropy (which are commonly used for evaluation of income inequality) as an algorithmic fairness because of several prominent properties of it; for example, i) high generalized entropy implies high inequality and the value of 0 of generalized inequality means the perfect fairness, ii) generalized entropy can simultaneously represent individual fairness and group fairness and reveals the trade-off between individual (un)fairness and group (un)fairness.
In this paper, we investigate whether the generalized entropy can play well the role of specifying fairness requirement for a classification algorithm. For this objective, first, we examine the role of generalized entropy as group-level and individual-level fairness concepts and second, we analyze the deviation of the empirical fairness degree measured by generalized entropy on some sample data set from the true fairness degree of it on the original space where the sample data set has been drawn. The second task requires us to extend generalized entropy defined on a finite set so that it can work on even an uncountable space while keeping desirable properties of the original one after an extension. Ultimately we are interested in finding an optimal empirical classifier satisfying the fairness requirement specified by the generalized entropy with small deviation. Ultimately we are interested in finding an optimal empirical classifier satisfying the fairness requirement specified by the generalized entropy.

The most related works are the papers of [27, 15, 1, 6]. The starting point of our work is the paper of [27] that proposes to use generalized entropy as an algorithmic fairness metric. The major difference of the paper and ours is the role of generalized entropy. Speicher et al. are interested in how to measure fairness of a given algorithm and examine generalized entropy as a metric to quantifying fairness degree. However, they do not consider an algorithm design seeking fairness requirement specified in terms of generalized entropy. Unlike the paper, our paper focuses on the design and analysis of a fair algorithm where the requirement is given by generalized entropy. In this paper, the role of generalized entropy is not limited to a fairness metric but extends to a design parameter in pursuit of a fair algorithm.

Our algorithm-designing philosophy is similar to that of [15, 1] in that all are to seek a randomized algorithm based on Hedge algorithm and analyze the performance of them with the help of minmax game theory. However, the objectives of the papers are very different from ours. The objective of Agarwal et al.’s work [1] is to provide the unified reduction approach for fair classification where the fairness concepts are represented by a linear function of number of false positive and false negative labels, including demographic parity and equalized opportunity (or odds) fairness. Our work is not related with such reduction approach. The objective of Kearns et al.’s work [15] is how to prevent fairness gerrymandering, the situation where a classifier satisfies some fairness constraint on small number of pre-defined groups but it severely violates the fairness constrain on subgroups which are generated on a more granular level. Our work does not focus on preventing fairness gerrymandering.

Our paper and [6] study PAC learnable fair learning algorithms. The fairness measure in [6] is based on malfare, disadvantage or loss caused by wrong prediction, and our fairness measure is based on generalized entropy.

We briefly summarize our contributions:

- We formulate and theoretically analyze a (randomized) empirical risk minimization with a fairness constraint given by generalized entropy. We show the PAC learnability of the empirical risk minimization (ERM) problem by proving that the true fairness degree does not deviate much from an empirical one with high probability if the sample size is big enough. We believe that our work provides the validity of generalized entropy as an algorithmic fairness metric.
• Our theoretical analysis gives some guidelines how to choose parameters that convert classification results to the fairness metric. We believe that these are useful in practice to ensure small deviation from true fairness degree and finding an optimal empirical risk minimizer.

• We propose new group-level fairness concepts, named equal prediction, equal error and equal benefits, which are closely related with the between-group term, the value of generalized entropy when each individual of a same subgroup is assigned to an identical value. We also identify that equal prediction is equivalent to equalized odds under the equal base rates.

The structure of this paper is as follows. Problem formulation and the definition of generalized entropy are described in Section 2. Section 3 examines generalized entropy in terms of group-level fairness and individual-level fairness concepts. Several group-level fairness concepts including equal prediction are proposed. In Section 4, we prove that the empirical fairness degree does not deviate much from the true one with high probability. We study on the optimal empirical risk minimizer and analyze its performance in Section 5. Experimental results using a real data set are reported in Section 6. The paper concludes with a summary in Section 7.

2 Problem Formulation

In this paper, vectors are denoted in boldface font and scalars (real-valued numbers) in normal font; \( \mathbf{b} \) is a vector while \( b \) is a real value.

2.1 Definition of Generalized Entropy Index

Generalized entropy index is originally developed to measure income inequality over finite population in \([26]\). Consider a population of \( n \) individuals. Each individual, denoted by \( \mathbf{x}_i \), has feature vector \( \mathbf{x}_i \) and benefit value \( b_i \). For a given income vector \( \mathbf{b} = (b_1, \ldots, b_n) \) with \( b_i \geq 0 \) for all \( i \), Shorrock has considered an inequality measure, \( I(\mathbf{b}; n) \), that satisfies the following axioms.

• Axiom 1: \( I(\mathbf{b}; n) \) is continuous and symmetric in \( \mathbf{b} \), i.e., \( I(\mathbf{b}; n) = I(\mathbf{b}'; n) \) where \( \mathbf{b}' \) is a permutation of \( \mathbf{b} \).

• Axiom 2 : \( I(\mathbf{b}; n) \geq 0 \) and \( I(\mathbf{b}; n) = 0 \) if and only if \( b_i = c \) for some constant \( c \) and for all \( i \).

• Axiom 3: \( \frac{\partial^2 I(\mathbf{b}; n)}{\partial b_i \partial b_j} \) is continuous for all \( i, j \).
• Axiom 4 (Additive decomposability): For \( n \geq 2 \) and any partition \( [b_1, b_2, \ldots, b_G] \) of \( b \) with \( b_i = (b_{i,1}, \ldots, b_{i,n_i}) \), there exists a set of \( w_g^G(\mu, n) \) such that

\[
I(b_1, \ldots, b_G; n) = \sum_g w_g^G(\mu, n) I(b^g; n_g) + V
\]

where \( \mu = (\mu_1, \ldots, \mu_G) \) with \( \mu_g = \frac{1}{n_g} \sum_{i=1}^{n_g} b_{i}^g \), \( n = (n_1, \ldots, n_G) \) with \( n_g \) the cardinality of \( b^g \), and \( V = I(\mu_1 1_{n_1}, \ldots, \mu_G 1_{G}; n) \) with \( 1_m = (1,1,\ldots,1) \in \mathbb{R}^m \).

• Axiom 5: \( I(b, b, \ldots, b; kn) = I(b; n) \) for any positive integer \( k \).

• Axiom 6 (Pigou-Dalton principle of transfers): if a transfer \( \delta > 0 \) is made from an individual \( j \) to another \( i \) with \( b_j > b_i \) such that \( b_j - \delta > b_i + \delta \), then the inequality index decreases after the transfer.

• Axiom 7: \( I(rb; n) = I(b; n) \) for any \( r > 0 \).

**Definition 2.1.** For any \( \alpha \in [0, \infty) \) and a given vector \( b = (b_1, \ldots, b_n) \) with \( b_i \geq 0 \) for all \( i \), the generalized entropy index of \( b \), denoted by \( I_{\alpha}(b; n) \), is defined as

\[
I_{\alpha}(b; n) = \begin{cases} 
\frac{1}{n} \sum_{i=1}^{n} \ln \frac{\mu_i}{\mu} & \text{for } \alpha = 0, \\
\frac{1}{n} \sum_{i=1}^{n} \frac{b_i}{\mu} \ln \frac{b_i}{\mu} & \text{for } \alpha = 1, \\
\frac{1}{n\alpha(\alpha-1)} \sum_{i=1}^{n} \left( \frac{b_i}{\mu} \right)^{\alpha} - 1 & \text{for } \alpha \neq 0,1,
\end{cases}
\]

where \( \mu = \frac{1}{n} \sum_{i=1}^{n} b_i \).

When \( \alpha = 1 \), the generalized entropy index is called Theil index \([28, 26]\). The more positive \( \alpha \), the more sensitive \( I_{\alpha}(b; n) \) to the inequalities of high income distribution.

### 2.2 Applying Generalized Entropy Index to Measuring Algorithmic Fairness

Consider a supervised machine learning problem. Each individual is represented by \((x, y) \in X \times Y\) where \( x \) is a feature vector and \( y \in Y \) is a (ground truth) label of \( x \). We assume that \( Y = \{0,1\} \) and there is an unknown distribution \( P \) over \( X \times Y \). Furthermore, the label value 1 corresponds to the desirable case for an individual and the value 0 to the undesirable one. For a credit lending example, acceptance of a loan application corresponds to 1 and rejection of it to 0. The marginal distribution of \( P \) over \( X \) is denoted by \( P_x \). A sample data set (or training data) with size \( n \), \( S = \{(x_i, y_i)\}_{i=1}^{n} \), consists of elements \((x_i, y_i)\) independently and identically distributed (i.i.d) according to the unknown distribution \( P \) over \( X \times Y \). A hypothesis, called

\[1\text{For a given set } X, \text{ a class of subsets } \{X^i\}_{i=1}^{G} \text{ with } X^i \subseteq X \text{ is called a partition of } X \text{ if and only if } \bigcup_i X^i = X \text{ and } X^i \cap X^j = \emptyset \text{ if } i \neq j.\]
also a learning algorithm, is a function \( h : \mathcal{X} \to \mathcal{Y} \) that outputs a predicted label \( h(x) \), either correct or incorrect, for \( x \in \mathcal{X} \). The empirical risk (or error) of hypothesis \( h \) is the error that \( h \) incurs on the sample data \( X = \{x_i\}_{i=1}^n \):

\[
R_S(h) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{h(x_i) \neq y_i\}
\]

where \( \mathbb{1}\{C\} \) is an indicator function that returns 1 if condition \( C \) is satisfied and returns 0 otherwise. The true error of a hypothesis \( h \) is the error that \( h \) generates over the whole domain \( \mathcal{X} \):

\[
R_X(h) = \mathbb{P}_{(x,y) \sim P}[h(x) \neq y] = \mathbb{P}(\{(x,y)\mid h(x) \neq y\})
\]

We assume that \( \mathcal{X} \) and a class of hypothesis \( \mathcal{H} \) are given. The objective of learning is to find a hypothesis \( h \in \mathcal{H} \) for a given hypothesis class \( \mathcal{H} \) that predicts well the label of a new instance \( x \) (i.e., yields small \( R_X(h) \)) with the help of a sample data set, which is usually done by empirical risk minimization (ERM): find \( h^* \)

\[
h^* = \arg \min_{h' \in \mathcal{H}} R_S(h')
\]

and we expect \( R_X(h^*) \) is small too. It is very well-known that the true error of \( h \) is close to the empirical error with high probability if the sample data size is large enough and VC dimension is finite: The standard VC dimension bound is provided in Theorem 2.2 [21].

**Theorem 2.2.** (*Standard VC Dimension Bound*) For any distribution \( P \) over \( \mathcal{X} \times \mathcal{Y} \), let \( S = \{(x_i, y_i)\}_{i=1}^n \) be a sample data set identically and independently drawn according to \( P \). For any \( 0 < \delta < \frac{1}{2} \) and any \( h \in \mathcal{H} \), with probability at least \( 1 - \delta \), it holds that

\[
R_X(h) \leq R_S(h) + \sqrt{\frac{8d_H \ln \left( \frac{2en}{d_H} \right) + 8 \ln \frac{4}{\delta}}{n}}
\]

where \( d_H \) is the VC dimension of \( \mathcal{H} \).

Speicher et al. proposes to use \( I_\alpha(b; n) \) for some \( \alpha > 0, \alpha \neq 0, 1 \) as a metric to measure the fairness degree of \( h \) by converting prediction results of \( h \) into benefit values \( b_i \) in the following way,

\[
b_i = h(x_i) - y_i + 1
\]

The last term, adding one, makes \( b_i \) is non-negative so that \( I_\alpha(b; n) \) can be defined for finite \( n \) and \( \alpha > 0, \alpha \neq 0, 1 \). The philosophy of \( I_\alpha(b; n) \) can be understood by an example of a bank’s lending system where the label of 1 corresponds to the acceptance of a loan application and the label 0 the rejection of it. Each loan applicant is either creditworthy and can pay back the loan, denoted by the label 1, or she is not creditworthy and will default, denoted by the label 0. For an applicant with true label \( y_i = 1 \), she would think the decision is unfair if \( h(x_i) = 0 \). For another applicant with her true label \( y_j = 0 \), she would get more benefit than she deserves if
\[ h(x_j) = 1; \] others would think the decision is unfair. Later this conversion from the predicted and true labels to the benefit is more generalized in [14].

Similarly as in (1), for \( h \in H \), we define \( b_h(x) \) as
\[
b_h(x) = a(h(x) - y) + c
\]
with \( c > a > 0 \) and \( c - a \geq 1 \). By the definition of \( b_h(x) \) in (2), the benefit of an individual is \( c \) for correct prediction, \( c+a \) for false positive prediction, and \( c-a \) for false negative prediction. We will drop the subscript \( h \) if \( h \) is clear in context. We give an example of computation of \( I_\alpha \) for a given hypothesis \( h \) and sample data set \( S = \{x_i, y_i\} \).

Now, for a given hypothesis \( h \in H \) and \( \alpha \in [0, \infty) \), we can measure its empirical algorithmic unfairness from the sample data set \( \{x_i, y_i\}_{i=1}^n \) by \( I_\alpha(b_h; n) \) as in Definition 2.1 where \( b_h = (b_h(x_1), \cdots, b_h(x_n)) \). We consider an ERM with a fairness constraint that is specified by generalized entropy for some \( \alpha \in [0, \infty) \), which we call a fair empirical risk minimization (FERM) with generalized entropy;
\[
\text{FERM-GE:} \quad \min_{h \in H} R_S(h) \quad \text{subject to} \quad I_\alpha(b_h; n) \leq \gamma.
\]
Let \( h^* \) be the optimal solution of FERM-GE (3). We investigate the true error \( R_X \) and the fairness degree of \( h^* \) over \( X \) (where \( x_i \) belongs to) do not deviate much from the empirical ones over a sample data set if the training set is sufficiently large. Ultimately, we are interested in learning a classifier with a small error and small degree of fairness.

For this, we extend the original definition of generalized entropy defined on a finite population \( \{x_i\}_{i=1}^n \) to its extension that is defined on \( X \), an arbitrary space, and check whether the properties, especially including additive decomposability property, still hold for the extended definition. Before we extend the generalized entropy and formulate an FERM, we address the advantages of generalized entropy as a metric for algorithmic fairness.

3 Advantages of Generalized Entropy as a Fairness Metric

Partitioning of a population into several subgroups can be described by using features; for example, using the gender feature, we can partition the whole population into two groups, a group of females and a group of males, if gender has only two components, male and female. We assume that the whole population is partitioned to \( G \) subgroups. One of the most prominent properties of generalized entropy is additive decomposability of Axiom 4,
\[
I(b^1, \ldots, b^G; n) = \sum_{g=1}^G w^G_g(\mu, n)I(b^g; n) + V
\]
where

\[ w^G_g(\mu, n) = \frac{n_g}{n} \left( \frac{\mu_g}{\mu} \right)^\alpha, \]

and

\[ V = I(\mu_1 n_1, \ldots, \mu_G n_G; n) \]

= \begin{cases} 
\sum_{g=1}^{G} \frac{n_g}{n} \ln \frac{\mu_g}{\mu} & \text{for } \alpha = 0, \\
\sum_{g=1}^{G} \frac{n_g\mu_g}{n\mu} \ln \frac{\mu_g}{\mu} & \text{for } \alpha = 1, \\
\frac{1}{\alpha(\alpha-1)} \sum_{g=1}^{G} \frac{n_g}{n} \left( \frac{\mu_g}{\mu} \right)^\alpha - 1 & \text{otherwise.} 
\end{cases} 

The first term in (4), \( \sum_g w^G_g(\mu, n)I(b^g; n_g) \), called by within-group term, is the weighted sum of inequality in subgroup \( b^g \) and the second term \( V \), called by between-group term, is the inequality of the population with size \( n \) where each subgroup consists of \( n_g \) members who have equal income \( \mu_g \), which implies that in each subgroup, perfect equality is achieved, that is every member in the same subgroup have equal income. In general, many group fairness definitions implicitly assume that the individuals in a same subgroup are treated equally (for example, in equal opportunity of [13], any individual of a subgroup \( g \) is implicitly assumed to have the same false negative rate). From such perspective, between-group term can be regarded as a kind of group fairness definition. (The computation of additive decomposability including \( I(b^1, ..., b^G; n) \) and between-group term \( V \) can be found in Appendix B).

Let us consider \( \sum_{g=1}^{G} w^G_g(\mu, n) \). Obviously, it holds that \( \sum_{g=1}^{G} w^G_g(\mu, n) = 1 \) only when \( \alpha = 1 \) or 0 [28]. Considering that \( w^G_g(\mu, n) = \frac{n_g}{n} \) for \( \alpha = 0 \) and \( w^G_g(\mu, n) = \frac{n_g\mu_g}{n\mu} \) for \( \alpha = 1 \), we know that the weight for \( \alpha = 0 \) is the ratio between the subgroup size and the total population while the weight for \( \alpha = 1 \) is the ratio between the subgroup’s benefit and the total benefit.

There have been two approaches in the study of fairness in machine learning, group-level fairness and individual-level fairness. Through the lens of additive decomposability, we examine the role of generalized entropy as group-level and individual-level fairness concepts in the following subsections. We summarize several advantages of generalized entropy as an algorithmic fairness.

- As revealed in the equation (4) of additive decomposability, generalized entropy simultaneously represents individual-level fairness and group-level fairness and explains how individual-level fairness and group-level fairness are related.

- Generalized entropy is easy to compute; it only needs the ground-truth labels and a classifier’s predicted labels. In general, individual-level fairness is not easy to compute [9, 16, 17]. One nice property of between-group term is that it is a scalar-valued statistics summarizing the differences between a classifier’s prediction over subgroups. Note that many existing definitions of group-level fairness does not have such a scalar valued statistics over subgroups. When generalized entropy is used as a fairness metric, the auditing becomes drastically easy, just check the value of generalized entropy, since unfairness is expressed as one single scalar value.
• It is well known that group-level fairness such as equal opportunity, equalized odd, or demographic parity, does not guarantee individual-level fairness. In generalized entropy, individual-level fairness is achieved only when \( I_\alpha = 0 \). Non-zero value of generalized entropy ensures that individual fairness is not achieved. More specifically, the additive decomposability, Axiom 4, \( I(b^1, \ldots, b^G; n) = \sum_g w_g^G I(b^g; n_g) + V \) explicitly shows that even though \( V \) is zero (i.e., group-level fairness is achieved), it may hold that generalized entropy is positive, i.e., \( I(b^1, \ldots, b^G; n) > 0 \), if there exists a subgroup such that \( I(b^g; n_g) > 0 \) (unfairness exists in a subgroup) since \( w_g^G > 0 \).

3.1 Generalized Entropy vs. Group-level Fairness

Many group-level fairness definitions partition the whole population into several subgroups and compare some statistical measures over the subgroups. Typical examples are demographic parity, equal opportunity, and equalized odds [10, 13]. Such group-level fairness definitions implicitly assume that the individuals in a same subgroup are treated equally. From this perspective, between-group term, \( V \), in (4) can be regarded as a metric quantifying the degree of a new group-level fairness definition, “equal prediction”. We define “equal prediction” and study its properties and relationship with generalized entropy.

Let \( X = \{x_i \mid 1 \leq i \leq n\} \) and \( \{X^i\}_i^G \) be a partition of \( X \) where \( X^g \) has exactly \( n_g \) elements. Then \( b^g = \{b(x_i) \mid x_i \in X^g\} \). For a given subgroup \( X^g \), we define the false positive fraction, \( \phi_g^{FP} \), and false negative fraction, \( \phi_g^{FN} \), as follows

\[
\phi_g^{FP} = P[h(x_i) = 1, y_i = 0 \mid x_i \in X^g], \quad \phi_g^{FN} = P[h(x_i) = 0, y_i = 1 \mid x_i \in X^g].
\]

**Definition 3.1.** (Equal prediction) We say that a hypothesis \( h \) satisfies equal prediction if \( \phi_{g_1}^{FP} = \phi_{g_2}^{FP} \) and \( \phi_{g_1}^{FN} = \phi_{g_2}^{FN} \) for any \( g_1, g_2 \) with \( 1 \leq g_1, g_2 \leq G \).

The condition that \( \phi_g^{FP} = c_1 \) and \( \phi_g^{FN} = c_2 \) for some constants \( c_1, c_2 \), and any \( g \), implies that the prediction capability of hypothesis \( h \) is the same over subgroups. Under equal prediction, the between-group term becomes 0 as in Proposition 3.2. The converse of Proposition 3.2 does not hold, in general.

**Proposition 3.2.** For a given partition \( \{X^g\}_g^G \) of \( X \), if \( h \) satisfies equal prediction, then the between-group term \( V \) is 0.

**Proof.** If a hypothesis \( h \) satisfies equal prediction, then obviously \( \mu_g = \mu \) for all \( 1 \leq g \leq G \). From (4), it is clear that \( V = 0 \) if \( \mu_g = \mu \) for all \( 1 \leq g \leq G \). \( \square \)

Related to equal prediction, we propose two group-level fairness definitions, equal error and equal benefit.

**Definition 3.3.** (Equal error and equal benefit) We say that a hypothesis \( h \) satisfies equal error if \( \phi_{g_1}^{FP} + \phi_{g_1}^{FN} = \phi_{g_2}^{FP} + \phi_{g_2}^{FN} \) for any \( g_1, g_2 \) with \( 1 \leq g_1, g_2 \leq G \). We say that a hypothesis \( h \) satisfies equal benefit if \( \phi_{g_1}^{FP} - \phi_{g_1}^{FN} = \phi_{g_2}^{FP} - \phi_{g_2}^{FN} \) for any \( g_1, g_2 \) with \( 1 \leq g_1, g_2 \leq G \).
Note that \( \phi_{g_1}^{FP} + \phi_{g_1}^{FN} \) is an error probability (or a risk) of hypothesis \( h \) over the subgroup \( X^{g_1} \). Hence equal equal requires that hypothesis \( h \) have equal error probability (or risk) over the subgroups. If hypothesis \( h \) satisfies equal prediction, then it obviously satisfies equal error and equal benefit. When hypothesis \( h \) satisfies equal benefit, then \( \mu_{g_1} = \mu_{g_2} \) for any \( g_1, g_2 \).

The between-group fairness, \( V \), is closely related with the fairness definition of equalized odds among many definitions of group-level fairness definitions. Recall the fairness definition of equalized odds, false positive rate, \( r_g^{FP} \), and false negative rate, \( r_g^{FN} \), for a given subgroup \( X^g \) in [13]:

\[
\begin{align*}
    r_g^{FP} &= P[h(x_i) = 1 \mid x_i \in X^g, y_i = 0], \\
    r_g^{FN} &= P[h(x_i) = 0 \mid x_i \in X^g, y_i = 1].
\end{align*}
\]

A classifier \( h \) satisfies equalized odds if \( r_{g_1}^{FP} = r_{g_2}^{FP} \) and \( r_{g_1}^{FN} = r_{g_2}^{FN} \) for any \( 1 \leq g_1, g_2 \leq G \).

Note that obviously \( \phi_g^{FP} \neq r_g^{FP} \) and \( \phi_g^{FN} \neq r_g^{FN} \) for a given subgroup \( X^g \). Hence, in general, classifier \( h \) satisfying equalized odds does not satisfy equal prediction and vice versa. However, under some specific condition, satisfying equalized odds is equivalent to satisfying equal prediction, as in Theorem 3.4.

**Theorem 3.4.** Let \( r_g^{POS} = P[y_i = 1 \mid x_i \in X^g] \). If \( r_{g_1}^{POS} = r_{g_2}^{POS} \) for any \( 1 \leq g_1, g_2 \leq G \), then equalized odds is equivalent to equal prediction.

**Proof.** Since \( P[AB] = P[A \mid B] \cdot P[B] \) for any \( A \) and \( B \), we have

\[
\begin{align*}
    \phi_g^{FP} &= P[h(x_i) = 1, y_i = 0 \mid x_i \in X^g] \\
    &= P[h(x_i) = 1 \mid y_i = 0, x_i \in X^g] \cdot P[y_i = 0 \mid x_i \in X^g] \\
    &= r_g^{FP} \cdot (1 - r_g^{POS}) \quad (5)
\end{align*}
\]

for any \( 1 \leq g \leq G \). By (5) and the assumption of \( r_g^{POS} = r_g^{POS} \) for any \( g_1, g_2 \), we can easily know that \( \phi_{g_1}^{FP} = \phi_{g_2}^{FP} \) is equivalent to \( r_{g_1}^{FP} = r_{g_2}^{FP} \) for any \( g_1, g_2 \). Similarly, it holds that \( \phi_{g_1}^{FN} = \phi_{g_2}^{FN} \) is equivalent to \( r_{g_1}^{FN} = r_{g_2}^{FN} \) for any \( g_1, g_2 \).

We remark that under the condition of equal \( r_g^{POS} \) for any \( g \), called by equal base rates, a classifier simultaneously satisfies the notions of fairness in calibration and in equalized odds, in [18]. The following corollary is a direct result of Theorem 3.4.

**Corollary 3.5.** If \( r_{g_1}^{POS} = r_{g_2}^{POS} \) for any \( 1 \leq g_1, g_2 \leq G \) and a hypothesis \( h \) satisfies equalized odds, then \( V = 0 \).

### 3.2 Generalized Entropy vs. Individual-level Fairness

In the previous subsection, we provide the perspective that the between-group term is a new group-level fairness definition. Consider a special case where each subgroup consists of exactly one single element (or individual). In this special case, each \( b^g \) comprises of a single \( b_i \) for some \( i \). Hence \( I(b^1; n_g) = I(b^1; 1) = 0 \), which results in \( \sum_g w^g_y(\mu, n) I(b^g; n_g) = 0 \), that is, within-group term becomes 0. Therefore, we have \( I(b^1, ..., b^G; n) = V \), which implies that
generalized entropy is an extreme case of group-level fairness, V. Kearns et al. propose in [17]
the notion of “average” individual fairness that seeks equal averaged error rates over individuals,
when there are sufficiently many classification tasks so that each individual takes an averaged
error rate over multiple classification tasks. The average individual fairness of [17] can be
also regarded as an extreme case of subgroup-level fairness if we regard each individual as a
subgroup. In [15, 16], they also propose a notion of rich subgroup fairness to bridge the gap
between group-level fairness and individual-level fairness by considering very large number of
subgroups through the combination of setting feature values rather than a small number of
subgroups. In the context of this rich subgroup fairness, the finest level is the special case that
each individual is a subgroup with size one. The viewpoint that individual-level fairness is a
special extreme case of group-level fairness is a natural perspective.

Note that if \( b_i \neq b_j \) for some \( i \neq j \), then \( I_\alpha(b; n) \neq 0 \) by Axiom 2. Also \( I_\alpha(b; n) = 0 \) only
when all \( b_i \)'s are equal. Non-zero value of generalized entropy ensures the existence of (at least
two) individuals whose classification result (predicted value - ground truth value) is different
from the other one’s.

4 From Empirical Fairness to True Fairness

4.1 Extension of Generalized Entropy

The parametric family of generalized entropy in Definition 2.1 is originally defined over a finite
population under the premise that \( n \) individuals are separately identified and each has the same
weight \( \frac{1}{n} \). Hence Definition 2.1 is easily applied to the sample data set \( S = \{x_i, y_i\}_{i=1}^n \) with
finite size but not adequate for some space \( X \), like \( \mathbb{R}^m \), that has uncountably many elements.
We extend the generalized entropy \( I_\alpha \) so that the extended one can work even for a space with
uncountably many elements. Let \( \mathbb{R}^+ \) be the set of non-negative real numbers.

Definition 4.1. (Extension of Generalized Entropy) Let \( b : X \rightarrow \mathbb{R}^+ \) and \( P_x \) be a probability
distribution on \( X \). For a constant \( \alpha \in [0, \infty) \), generalized entropy of \( b(X) \) with respect to the
probability distribution \( P_x \) on \( X \) is defined by

\[
I_\alpha(b, X, P_x) = \int_X f_\alpha \left( \frac{b(x)}{E[b(x)]} \right) dP_x
\]

(6)

where \( dP_x \) is the probability density function of \( P_x \),

\[
f_\alpha(x) = \begin{cases} 
-\ln x & \text{if } \alpha = 0, \\
x \ln x & \text{if } \alpha = 1, \\
\frac{1}{\alpha(\alpha - 1)}(x^{\alpha} - 1) & \text{otherwise}, 
\end{cases}
\]

(7)

and \( E[b(x)] = \int_X b(x) dP_x \).

It is important to check if the above extension of generalized entropy satisfies Axioms 1-7
that define the original one. Especially Axioms 2, 4, and 6 are important because Axiom 2
states about the degree of individual-level fairness, Axiom 4 is the property exhibiting the
relationship between individual-level fairness and group-level fairness, and Axiom 6 is one of the fundamental requirement defining equity measure. It is easy to see that axioms 1, 5, and 7 hold even for the extension. Axioms 3, 4, and 6 should be modified to reflect the change.

- Axiom 3: \( \frac{\partial g}{\partial b} \) is continuous.
- Axiom 4' (Additive decomposability): For any partition \( \{X^g\}_{g=1}^G \) of \( X \) where \( X^g \) has the probability distribution \( P^g_x \) such that \( P^g_x(A) = \frac{P_x(A)}{P_x(X^g)} \) for any \( A \subseteq X^g \), it holds that

\[
I_\alpha(b, X, P_x) = \sum_{g=1}^G w_{\alpha,g} I_\alpha(b|X^g, X^g, P^g_x) + I_\alpha(v, X, P_x)
\]

where \( b|X^g \) is the restricted function of \( b(x) \) on \( X^g \), \( w_{\alpha,g} = m_g \left( \frac{\mu_g}{\mu} \right)^{\alpha} \), \( m_g = P_x(X^g), \mu = \mu_g = \int_{X^g} b(x) \, dP^g_x \), and \( v(x) = \sum_{g=1}^G \mu_g \mathbb{1}_g \{ x \in X^g \} \).

- Axiom 6' (Pigou-Dalton transfer principle): Consider \( M_i, M_j \subset X \) such that \( M_i \cap M_j = \emptyset \) and \( b(x_i) > b(x_j) \) for any \( x_i \in M_i \) and \( x_j \in M_j \). Define \( \tilde{b}(x) \) as below

\[
\tilde{b}(x) = \begin{cases} 
  b(x) - \delta & \text{if } x \in M_i, \\
  b(x) + \delta & \text{if } x \in M_j, \\
  b(x) & \text{otherwise}.
\end{cases}
\]

It holds that \( I_\alpha(b, X, P_x) > I_\alpha(\tilde{b}, X, P_x) \), if \( E[b(x)] = E[\tilde{b}(x)] \) and \( b(x_i) - \delta > b(x_j) + \delta \) for any \( x_i \in M_i, x_j \in M_j \).

We can prove that Axioms 2, 4', and 6' with a mild condition (refer to Appendix \( \Box \) for the proof of Theorem 4.2)

**Theorem 4.2.** The extension of generalized entropy, \( I_\alpha(b, X, P_x) \), satisfies Axiom 2, 4', and 6' for all \( \alpha \in [0, \infty) \), if \( b(x) \) is bounded on \( X \); the extension of generalized entropy satisfies all (modified) Axioms 1-7.

Recall that the benefit function of a given hypothesis \( h \) is \( b_h(x) = a(h(x) - y) + c \), with \( c > a > 0 \), and \( c - a \geq 1 \). It is obvious that \( b_h(x) \) is bounded over \( X \). Hence we can consider \( I_\alpha(b_h, X, P_x) \) with properties of Axiom 2, additive decomposability, and Pigou-Dalton transfer principle for any \( \alpha \in [0, \infty) \). As we observed in Section 3, \( \sum_{g=1}^G w_{\alpha,g} \) is 1 only when \( \alpha \in \{0, 1\} \) for the extension of generalized entropy.

### 4.2 Difference between True Fairness and Empirical Fairness

Now we are ready to define algorithmic fairness using generalized entropy. Recall that for any distribution \( P \) defined on \( X \times Y \), \( P_x \) denotes the marginal distribution of \( P \) over \( X \). To
emphasize a hypothesis \( h \) and a probability distribution \( P \) on \( X \times Y \), we will use \( I_{\alpha}(h, P) \) instead of \( I_{\alpha}(h, X, P_X) \) from now on, even though the generalized entropy definition needs \( P_X \) not \( P \). For the sample data set \( S \), we still use \( I_{\alpha}(h; n) \). Hence for a given \( h \), \( I_{\alpha}(h, X) \) is the true fairness of \( h \) over \( X \times Y \), \( I_{\alpha}(h; n) \) the empirical fairness of \( h \) over \( S \) and \( I_{\alpha}(h, X) \) the true fairness of \( h \) over the whole domain \( X \).

**Definition 4.3.** For a hypothesis \( h : X \to Y \), it is said that \( h \) satisfies \((\alpha, \gamma)\)-generalized entropy fairness if \( I_{\alpha}(h, P) < \gamma \).

Similar to the generalization error of a hypothesis, we can also prove that the true fairness is close enough to the empirical fairness with high probability if the sample data size is large enough, Theorem 4.4 (whose proof can be found in Appendix D).

**Theorem 4.4.** (Generalization Fairness) For any distribution \( P \) over \( X \times Y \), let \( S = \{(x_i, y_i)\}_{i=1}^n \) be a sample data set identically and independently drawn according to \( P \). Let \( r = \frac{\varepsilon}{a} \). For any \( 0 < \delta < 1 \), with probability at least \( 1 - \delta \), for every \( h \in \mathcal{H} \) and \( \alpha \in [0, \infty) \),

\[
|I_{\alpha}(h, P) - I_{\alpha}(h; n)| \leq \psi_{\alpha}(a, r) \sqrt{\frac{1}{2n} \ln \frac{4}{\delta}}
\]

where

\[
\psi_0(a, r) = \frac{2}{r-1} + \ln \left( 1 + \frac{2}{r-1} \right),
\]

\[
\psi_1(a, r) = \frac{4}{r-1} + \frac{4r \ln(ar + a)}{(r - 1)^2},
\]

\[
\psi_{\alpha}(a, r) = \begin{cases} \frac{2}{(1-\alpha)(r-1)} \left( \frac{r+1}{r-1} \right)^{\alpha} + 1 & \text{if } 0 < \alpha < 1, \\ \frac{4r}{(\alpha-1)(r-1)^2} \left( \frac{r+1}{r-1} \right)^{\alpha-1} & \text{if } \alpha > 1. \end{cases}
\]

For fixed \( \alpha \), note that \( \psi_{\alpha}(a, r) \sqrt{\frac{1}{2n} \ln \frac{4}{\delta}} \) gets close to 0 as \( n \) goes \( \infty \). Theorem 4.4 tells us that large \( \alpha \) may have the big difference between true fairness degree and empirical one, which implies that small \( \alpha \) is preferred in practice. It can be easily shown that \( \psi_{\alpha}(a, r) \) is decreasing with \( r \) for fixed \( a \). From this fact, we learn that large \( r \) is preferred for small deviation of empirical fairness from true one. However, if \( r \) is too big, then it yields a very small value of \( I_{\alpha} \), which may cause difficulty in discerning the existence of unfairness.

Using the fact that \( \psi_{\alpha}(a, r) \) is decreasing with \( r \) and \( r \geq 1 + \frac{1}{\alpha} \) when \( a \) is fixed, we have the following corollary.

**Corollary 4.5.** Under the same conditions of Theorem 4.4, it holds that

\[
|I_{\alpha}(h, P) - I_{\alpha}(h; n)| \leq \psi_{\alpha}(a, r) \sqrt{\frac{1}{2n} \ln \frac{4}{\delta}}
\]

where

\[
\psi_0(a, r) \leq 2a + \ln(1 + 2a),
\]

\[
\psi_1(a, r) = 4a + 4a \ln(1 + 2a),
\]

\[
\psi_{\alpha}(a, r) = \begin{cases} \frac{2a}{(1-\alpha)}(1 + 2a)^{\alpha} + 1 & \text{if } 0 < \alpha < 1, \\ \frac{4a(a+1)}{(\alpha-1)}(1 + 2a)^{\alpha-1} & \text{if } \alpha > 1. \end{cases}
\]
In Theorem 4.4, the deviation bound of the empirical fairness from the true fairness degree is independent of classifier’s accuracy. In most cases, we are not interested in inaccurate hypotheses but in accurate ones. Since $I(b_h; n) = 0$ if $h$ has no error, we conjecture that the deviation of empirical fairness of an accurate classifier is smaller than that of an inaccurate one. Theorem 4.6 tells that our conjecture is true; When the empirical error itself is small (and is close enough to the true error), we can significantly improve the bound between empirical and true fairness degrees. Recall that $R_S(h)$ is the empirical error for a hypothesis $h$.

**Theorem 4.6.** For any distribution $P$ over $X \times Y$, let $S = \{(x_i, y_i)\}_{i=1}^n$ be a sample data set identically and independently drawn according to $P$. For any $0 < \delta < 1$ and $h \in \mathcal{H}$, with probability at least $1 - \delta$, for every $h \in \mathcal{H}$, it holds that for $\alpha \in [0, \infty)$,

$$|I_{\alpha}(h, P) - I_{\alpha}(b_h; n)| \leq \tilde{\psi}_\alpha(r, \varepsilon_2) \sqrt{\frac{8d_H \ln \left( \frac{2en}{d_H} \right) + 8 \ln \frac{8}{\delta}}{n}}$$

where $\varepsilon_2 = \sqrt{\frac{8d_H \ln \left( \frac{2en}{d_H} \right) + 8 \ln \frac{8}{\delta}}{n}}$ and

$$\tilde{\psi}_0(r, \varepsilon_2) = \frac{1}{r - R_S(h) - \varepsilon_2} + \ln \left( 1 + \frac{1}{r - 1} \right),$$

$$\tilde{\psi}_1(r, \varepsilon_2) = \frac{1}{r - R_S(h) - \varepsilon_2} \left[ 1 + r \left( 1 + 2 \ln(ar + a) \right) \right],$$

$$\tilde{\psi}_\alpha(r, \varepsilon_2) = \begin{cases} \frac{1}{\alpha} \left( \frac{r}{r - R_S(h)} \right)^{\alpha} \left[ 1 + \frac{C_\tilde{\psi}}{r - R_S(h) - \varepsilon_2} \right] & \text{if } 0 < \alpha < 1, \\ \frac{1}{(\alpha - 1)} \left( \frac{r}{r - R_S(h)} \right)^{\alpha} \left[ \frac{1}{r} \left( 1 + \frac{1}{r} \right)^{\alpha} + \frac{1 + \varepsilon_2}{r - R_S(h)} \right] & \text{if } \alpha > 1. \end{cases}$$

and $C_\tilde{\psi} = 1 + (1 + \frac{1}{r})^{\alpha} - 1) R_S(h) + \varepsilon_2 ((1 + \frac{1}{r})^{\alpha} + 1)$.

## 5 A Learning Algorithm Achieving Fair Empirical Risk Minimization

This section studies how to find an optimal (randomized) hypothesis satisfying $(\alpha, \gamma)$-fairness for given $\alpha$ and $\gamma$. By randomizing hypotheses, we can achieve better accuracy-fairness tradeoffs than using only the pure hypotheses. A randomized hypothesis $D$ is a probability distribution on $\mathcal{H}$, that is $D = \sum_{h \in \mathcal{H}} D_h h$ with $\sum_{h \in \mathcal{H}} D_h = 1$. The sample error of $D$ is given by

$$R_S(D) = \sum_{h \in \mathcal{H}} D_h R_S(h)$$

and for some $\alpha \in [0, \infty)$, the corresponding generalized entropy is

$$I_{\alpha}(D; n) = \sum_{h \in \mathcal{H}} D_h I_{\alpha}(b_h; n).$$
Let $\Delta \mathcal{H}$ be the set of all probability distributions over $\mathcal{H}$. We consider the following (randomized) fair empirical risk minimization problem:

$$
\min_{D \in \Delta \mathcal{H}} R_S(D)
$$

subject to $I_\alpha(D; n) \leq \gamma$. 

The above problem (8) is a convex optimization: the objective function is linear in $D_h$, $\sum_{h \in \mathcal{H}} D_h R_S(h)$ and the linear constraint is also linear in $D_h$, $\sum_{h \in \mathcal{H}} D_h I_\alpha(b_h; n) \leq \gamma$, and we want to find $\{D_h\}_{h \in \mathcal{H}}$ with $\sum D_h = 1$.

Consider the Lagrangian $L(D, \lambda)$ of (8), which is

$$
L(D, \lambda) = R_S(D) + \lambda \left( I_\alpha(D; n) - \gamma \right).
$$

Since (8) is a convex problem, the duality gap is zero:

$$
L^\star = \max_{\lambda \in \mathbb{R}^+} \min_{D \in \Delta \mathcal{H}} L(D, \lambda) = \min_{D \in \Delta \mathcal{H}} \max_{\lambda \in \mathbb{R}^+} L(D, \lambda). \tag{9}
$$

Note that $\lambda \in [0, \infty)$ in (9). To ensure the convergence to an equilibrium, we bound the range of $\lambda$ so that $\lambda \in \Lambda = [0, \lambda_{\max}]$. Since $\Lambda$ is compact and convex, the following still holds:

$$
L^\star_\Lambda = \max_{\lambda \in \Lambda} \min_{D \in \Delta \mathcal{H}} L(D, \lambda) = \min_{D \in \Delta \mathcal{H}} \max_{\lambda \in \Lambda} L(D, \lambda). \tag{10}
$$

The optimal solution $(D^\star_\Lambda, \lambda^\star_\Lambda)$ of (10), usually called the saddle point of $L^\star_\Lambda$, can be found as the equilibrium of a repeated zero sum game of two players, the learner seeking $D$ that minimizes $L(D, \lambda)$ and Nature seeking $\lambda$ that maximizes $L(D, \lambda)$ (see Chapter 5 of [3]). In the game, the set of pure strategies of Nature is $\{0, \lambda_{\max}\}$ and the set of pure strategies of the learner is $\mathcal{H}$. In finding $L^\star$ (or $L^\star_\Lambda$) in (9) (in [10]), it is sufficient to only one of the players play mixed strategies (hence the other players use pure strategies, instead of mixed ones) [11, 22].

Freund and Schapire has proposed a provable method to find an approximate solution of (10) that includes Hedge algorithm in [11, 12]. Using Hedge algorithm,\(^3\) we will find an approximate equilibrium $(\bar{D}, \bar{\lambda})$ of $L^\star_\Lambda$: for given $\nu > 0$, it holds that

$$
L^\star_\Lambda - \nu \leq \min_{\lambda \in \Lambda} L(D, \lambda), \tag{11}
$$

$$
\max_{\lambda \in \Lambda} L(\bar{D}, \lambda) \leq L^\star_\Lambda + \nu. \tag{12}
$$

One basic assumption of Hedge algorithm is the existence of an oracle yielding the best response of the learner: in our case, such an oracle corresponds to the process of finding $\arg \min_{h \in \mathcal{H}} L(h, \lambda)$ for $\lambda \in [0, \lambda_{\max}]$, the item 2 in the for loop of Algorithm 1. Another assumption of it is that the amount of gain of a strategy, which corresponds to $L(h, \lambda)$ for a

\(^2\)Even though $\mathcal{H}$ has infinitely many hypotheses, since the hypotheses in $\mathcal{H}$ are applied to the finite sample space $S$, the number of different labelings of $h \in \mathcal{H}$ is finite, i.e., cardinality of $\{h(S) \mid h \in \mathcal{H}\}$ is finite by Sauer’s Lemma [21] (i.e., count only once if $h_1(S) = h_2(S)$ for $h_1 \neq h_2$) when VC dimension of is finite.

\(^3\)We modify the original Hedge algorithm for our objective that Nature finds $\lambda \in [0, \lambda_{\max}]$ maximizing $L(D, \lambda)$, since the original Hedge algorithm is to find a randomized hypothesis minimizing loss.
Algorithm 1 Hedge algorithm for empirical generalized entropy \((\alpha, \gamma)\)-fairness

**Input:** A sample data set with size \(n\), \(\lambda_{\text{max}}\), \(\nu\), and an oracle finding \(\arg \min_{h \in \mathcal{H}} L(h, \lambda)\).

Initialize \(w_0^{(0)} = w_1^{(0)} = 1\).

Set \(\lambda_0 = 0\), \(\lambda_1 = \lambda_{\text{max}}\), \(T = \frac{4A_{\alpha}^2 \ln 2}{\nu \alpha^2}\), \(\kappa = \frac{\nu}{2A_{\alpha}}\).

**for** \(t = 1\) to \(T\) **do**

1. Nature chooses \(\hat{\lambda}^{(t)}\):

\[
\hat{\lambda}^{(t)} = \begin{cases} 
\lambda_0 & \text{with prob. } \frac{w_0^{(t-1)}}{w_0^{(t-1)} + w_1^{(t-1)}}, \\
\lambda_1 & \text{with prob. } \frac{w_1^{(t-1)}}{w_0^{(t-1)} + w_1^{(t-1)}},
\end{cases}
\]

where \(\lambda_0 = 0\) and \(\lambda_1 = \lambda_{\text{max}}\).

2. The learner chooses a hypothesis

\[
\hat{h}^{(t)} = \arg \min_{h \in \mathcal{H}} L(h, \hat{\lambda}^{(t)})
\]

3. Nature updates the weight vector: for \(j \in \{0, 1\}\),

\[
w_j^{(t)} = w_j^{(t-1)} \cdot (1 + \kappa)^{l(h^{(t)}, \lambda_j)}
\]

where \(l(h, \lambda) = \frac{L(h, \lambda) + B}{A_{\alpha}}\).

**end for**

**Output:** \(\hat{D} = \frac{1}{T} \sum_{t=1}^{T} \hat{h}^{(t)}\) and \(\bar{\lambda} = \frac{1}{T} \sum_{t=1}^{T} \hat{\lambda}^{(t)}\)

hypothesis \(h\), is always between 0 and 1. In our case, \(L(h, \lambda) \notin [0, 1]\) for a pure hypothesis \(h \in \mathcal{H}\): since \(L(h, \lambda) = R_S(h) + \lambda(I_\alpha(b; n) - \gamma)\), we have

\[
\lambda(I_\alpha^{\text{min}} - \gamma) \leq L(h, \lambda) \leq 1 + \lambda(I_\alpha^{\text{max}} - \gamma)
\]

where \(I_\alpha^{\text{min}} = \min_b I_\alpha(b; n)\) and \(I_\alpha^{\text{max}} = \max_b I_\alpha(b; n)\). Since \(I_\alpha(b; n) \geq 0\), we have \(I_\alpha^{\text{min}} = 0\). It can be easily checked that \(I_\alpha^{\text{max}} \leq I_\alpha^{UP}(r)\) where

\[
I_\alpha^{UP}(r) := \begin{cases} 
\ln \left(\frac{r+1}{r-1}\right) & \text{for } \alpha = 0, \\
\frac{r+1}{r-1} \ln \left(\frac{r+1}{r-1}\right) & \text{for } \alpha = 1, \\
\frac{1}{|\alpha(\alpha-1)|} \left[\left(\frac{r+1}{r-1}\right)^\alpha - 1\right] & \text{for } \alpha \neq 0, 1.
\end{cases}
\]

Let \(A_\alpha = 1 + \lambda_{\text{max}}(\gamma + I_\alpha^{UP}(r))\) and \(B = \gamma \lambda_{\text{max}}\). If we define \(l(h, \lambda) = \frac{L(h, \lambda) + B}{A_\alpha}\), then \(l(h, \lambda) \in [0, 1]\). We will use \(l(h, \lambda)\) instead of \(L(h, \lambda)\) in updating the weight vector, \((w_0^{(t)}, w_1^{(t)})\), in Hedge algorithm. Note that \(\arg \min_{h \in \mathcal{H}} L(h, \lambda) = \arg \min_{h \in \mathcal{H}} l(h, \lambda)\) and \(\arg \max_{\lambda \in \Lambda} L(D, \lambda) = \arg \max_{\lambda \in \Lambda} l(D, \lambda)\) since adding and multiplying a positive constant has no effect on optimization.

Algorithm 1 is the algorithm finding an approximate solution of (10).

**Theorem 5.1.** After \(T = \frac{4A_{\alpha}^2 \ln 2}{\nu \alpha^2}\) iterations, the output of the algorithm, \((\hat{D}, \bar{\lambda})\), satisfies (11) and (12).
Theorem 5.2. Suppose that $\gamma > \inf_{h \in H} I_\alpha(b_h; n)$. For any given $\nu$, the randomized hypothesis $\bar{D}$ satisfies

$$R_S(\bar{D}) \leq L^* + 2\nu,$$
$$I_\alpha(\bar{D}; n) \leq \gamma + \frac{1 + 2\nu}{\lambda_{\text{max}}}$$

after $T = \frac{4A^2 \ln 2}{\nu^2}$ iterations.

6 Experiments

This section shows the behaviors and performance of Algorithm 1 on real data sets, “Adult income data set” \cite{20} and “COMPAS” recidivism data set \cite{2} (more experiments on other data sets can be found in Appendix A). The task of Adult income data set is to predict if a person’s income is no less than $50k$ per year. For COMPAS data set, we use data samples whose race is either Caucasian or African-American. The task of COMPAS data set is to predict if an individual is rearrested within two years after the first arrest. In our setting, the case of recidivism within two years corresponds to label “0”, since the label “0” indicates an undesirable decision to an individual and no body wants re-arrest. Each data set is split into training examples (70%) and test examples (30%). We use $\lambda_{\text{max}} = 20$, $\nu = 0.005$. Regarding to $b_i = a(h(x_i) - y_i) + c$, we fix the value of $a$ as 5, and change the values of $c \in \{8, 9, 10\}$ to investigate the effect of $r = \frac{c}{a}$ on the performance. All figures are obtained after $10^4$ times running $\bar{D}$, the output of Algorithm 1 on the test data set.

Algorithm 1 assumes the existence of an oracle $W(\lambda)$. For the implementation of such an oracle, we use thresholding \cite{25}, a simple technique to directly find the best decision threshold for a given objective from the training data and use this threshold to predict the class label for the test data. Logistic regression has been used as a base classifier for thresholding.

6.1 Trade-off between Fairness and Accuracy

Figure 1 shows how the averaged values of test error of $\bar{D}$ behave as fairness constraint $\gamma$ (representing the $x$ axis) changes for each $\alpha = 0, 1, 2$. Each graph is the averaged value of test error of the randomized hypothesis $\bar{D}$ and the shaded regions over the graphs represent 95% confidence intervals. The dotted lines represent the test error of the empirical risk minimizer, $h_0 = \arg \min_{h \in H} L(h, 0)$. Recall that the fairness constraint is $I_\alpha \leq \gamma$ and that high $\gamma$ implies loose fairness constraint. In Figure 2 all the graphs of test error exhibit decreasing behavior as $\gamma$ increases. Based on this observation, in general, test error gets low (i.e., accuracy is enhanced) as fairness constraint gets loose.

Figure 2 illustrate how $I_\alpha$ of $\bar{D}$ varies as $\gamma$ changes for each $\alpha = 0, 1, 2$. Each graph is the average value of $I_\alpha$ and the shaded regions over the graphs represent 95% confidence intervals. The dotted lines represent $I_\alpha$ of the (empirical) risk minimizer. Every graph in Figure 2 shows that $I_\alpha$ increases as $\gamma$ increases: unfairness increases as the fairness constraint gets loose. For
each fixed $\alpha$, as the value of $c$ increases, i.e., $r = \frac{c-a}{c}$ increases, we observe that the fairness degree $I_\alpha$ decreases since the quantity $\frac{c-a}{c}$, which is the relative difference among benefits, $(c-a, c, c+a)$, gets diminishing.

From Figures 1 and 2 we assure that fairness is achieved at the cost of accuracy. Trade-off between accuracy and fairness is more sensitive to low $\alpha$ than to high $\alpha$.

Regarding to the choice of $a$ and $c$, small $\frac{c-a}{a}$ may not be a good choice for stringent $\gamma$ when $\alpha = 0$, since the test error for $\frac{c-a}{a} = \frac{8}{5}$ is relatively high compared with other values of it as in Figure 1(a). But small $\frac{c-a}{a}$ seems not to have such a problem when $\alpha = 2$ as in Figure 1(c).

### 6.2 Generalized Entropy, Between-group Term, and Equalized Odds

This subsection investigates how the between-group term $V_\alpha$ and equalized odds change as $\gamma$ increases. We partition the whole population into two groups, majority group ($g = 0$) and minority group ($g = 1$); for Adult income data set, $g = 0$ corresponds to male and
g = 1 to female, and for COMPAS data set, g = 0 corresponds to Caucasian and g = 1 to African-American. The values of a = 5, c = 8, and α = 1 are used.

Figure 3 shows how between-group term $V$ changes with $γ$. Figure 3(a) and Figure 3(b) show the opposite behaviors of $V$ as $γ$ increases: in Figure 3(a) between group $V$ decreases while in Figure 3(b) $V$ increases, as $γ$ increases. This observation shows that regulation of $I_α$ does not have much impact on the between-group term $V$. Note that $V$ is very small compared to $I_α$, which is observed in experiments of the other data sets. We have found that $V$ gets large as the number of groups, $G$, gets large.

Recall that equal opportunity is the case $r^F_{g} = r^F_{1}$ and equalized odds the case that $r^F_{0} = r^F_{1}$ and $r^P_{0} = r^P_{1}$. Figure 4 exhibits $r^F_{g}$ and Figure 5 $r^P_{g}$ for $g \in \{0, 1\}$. We define $r^F_{tot}$ and $r^P_{tot}$ as $r^F_{tot} = \mathbb{P}[h(x) = 1 | y = 0]$ and $r^P_{tot} = \mathbb{P}[h(x) = 0 | y = 1]$. In Figure 4 the minority group of $g = 1$ has higher false negative rate than the majority group of $g = 0$. Also in Figure 5, the group of $g = 1$ has low false positive rate than the group with $g = 1$. These phenomena are observed in all of the data sets we used. Since we set $h(x) = 1$ is a favorable decision to an individual both two figures show that more favorable decisions are made for group $g = 0$ than for group $g = 1$.

To examine how the group-level fairness definition of generalized odds (or equal opportunity) is related to $I_α$ we consider the quantities of $|r^F_{0} - r^F_{1}|$ and $|r^P_{0} - r^P_{1}|$ in Figures 4 and 5. Note that $|r^F_{0} - r^F_{1}|$ is the gap between the graph of $r^F_{0}$ and the graph of $r^F_{1}$. In order to showing $|r^F_{0} - r^F_{1}|$ and $|r^P_{0} - r^P_{1}|$, we plot $r^F_{g} - r^F_{tot}$ and $r^P_{g} - r^P_{tot}$ in Figures 6 and 7. As $γ$ increases (i.e., fairness constraint gets loose), the behavior of $|r^F_{0} - r^F_{1}|$ of Adult income data set is opposite to that of COMPAS data set; the quantity of $|r^F_{0} - r^F_{1}|$ decreases in Figure 6(a) but it increases in Figure 6(b). For the quantity $|r^P_{0} - r^P_{1}|$, we also observe similar

\begin{figure}[h]
\centering
\begin{subfigure}{0.4\textwidth}
\centering
\includegraphics[width=\textwidth]{fn_a.png}
\caption{(a) Adult income}
\end{subfigure}
\hfill
\begin{subfigure}{0.4\textwidth}
\centering
\includegraphics[width=\textwidth]{fn_b.png}
\caption{(b) COMPAS}
\end{subfigure}
\caption{$r^F_g$ for $g \in \{0, 1\}$ when $α = 1$ (x axis is $γ$.)}
\end{figure}

\begin{figure}[h]
\centering
\begin{subfigure}{0.4\textwidth}
\centering
\includegraphics[width=\textwidth]{fp_a.png}
\caption{(a) Adult income}
\end{subfigure}
\hfill
\begin{subfigure}{0.4\textwidth}
\centering
\includegraphics[width=\textwidth]{fp_b.png}
\caption{(b) COMPAS}
\end{subfigure}
\caption{$r^P_g$ for $g \in \{0, 1\}$ when $α = 1$ (x axis is $γ$.)}
\end{figure}
behaviors as in false negative rate case. Based on these observations, controlling fairness degree by generalized entropy has no specific impact on equalized odds (or equal opportunity). The experimental results of Figures 4-7 do not violate our analysis of FERM-GE since FERM-GE controls only $I_{\alpha}$.

6.3 Convergence of Empirical Errors and Generalized Entropy for Adult Income Data Set

One important performance issue is the convergence behavior of the empirical errors and values of empirical generalized entropy. We plotted the trajectories of empirical errors and empirical generalized entropy values for the adult income data set.

To investigate the convergence behaviors of time averaged empirical error and empirical $I_{\alpha}$, we plot the graphs of time averaged empirical error with time $t$ in Figure 8 and Figure 9 for given $t$, time averaged error is $\frac{1}{t} \sum_{\tau=1}^{t} \hat{h}(\tau)$ where $\hat{h}^\tau = \arg \min_{h \in H} L(h, \hat{\lambda}(\tau))$ in Algorithm 1. All figures have time $t$ as $x$ axis and use $a = 5$, and $c \in \{8, 9, 10\}$. Figure 9 is a closer look of Figure 8 for $1 \leq t \leq 5500$. One important performance issue is the convergence behavior of the empirical errors and values of empirical generalized entropy. We plotted the trajectories of empirical errors and empirical generalized entropy values for the data sets we used. Figures 8 and 9 provide convergence behaviors of empirical errors. Figures 10 and 11 provide convergence behaviors of empirical $I_{\alpha}$ of Adult income data set as $\gamma$ changes. Appendix A.2 provides figures showing how test error and test $I_{\alpha}$ changes with $\gamma$ as well as convergence behavior of empirical error and empirical $I_{\alpha}$. Appendix A.3 and Appendix A.4 provide all related figures on Law school data set and Dutch census data set, respectively.
Figure 8: Adult income: Empirical $I_\alpha$ trace when $a = 5$ for adult income data set ($x$ axis is $\gamma$)

Figure 9: Adult income: Close look of Figure 8 for $1 \leq t \leq 5500$ ($x$ axis is $\gamma$)

Figure 10: Adult income: Empirical error trace when $a = 5$ ($x$ axis is $\gamma$)

Figure 11: Adult income: Close look of Figure 10 for $1 \leq t \leq 5500$ ($x$ axis is $\gamma$)
7 Summary

We theoretically and experimentally analyze a (randomized) empirical risk minimization with a fairness constraint given by generalized entropy and prove that the randomized FERM-GE is PAC learnable by proving that the true fairness degree does not deviate much from an empirical one with high probability if sample size is big enough. Regarding to the choice of $\alpha$, $a$ and $c$, based on our analysis, a low value of $\alpha$ and a high value of $r = \frac{c}{a} (> 1)$ are preferred. But if $r$ takes a too high value, it may cause a difficulty in discovering the existence of unfairness. Our experiments do not clearly show how fairness constraints affect on between-group term $V$ even though there exists a clear trade-off between $I_\alpha$ and accuracy. We believe it is an interesting future work to identify when $V$ has similar behaviors as $I_\alpha$ when $\gamma$ changes.

Acknowledgements

This work is supported by the National Research Foundation of Korea (NRF) funded by Korean Government under NRF-2019R1I1A1A01059863 and NRF-2022R1A6A1A03052954.

References

[1] A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach. A reductions approach to fair classification. In Proc. of the 35th International Conference on Machine Learning, 2018.

[2] J. Angwin, J. Larson, S. Mattu, and L. Kircher. Machine bias. ProPublica, 2016.

[3] S. Boyd and L. Vandenberghe. Convex optimization. Cambridge, 2004.

[4] A. Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data, Special Issue on Social and Technical Trade-Offs, 5(2), 2017.

[5] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq. Algorithmic decision making and the cost of fairness. In Proc. of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, page 797–806, 2017.

[6] C. Cousins. An axiomatic theory of provably-fair welfare-centric machine learning. In Proc. of the 35th Conference on Neural Information Processing Systems, 2021.

[7] P. Van der Laan. The 2001 census in the netherlands. In Proc. of the Census of Population, 2017.

[8] M. Donini, L. Oneto, S. Ben-David, J. Shawe-Taylor, and M. Pontil. Empirical risk minimization under fairness constraints. In Proc. of the 32nd Neural Information Processing Systems, pages 2796–2806, 2018.
[9] C. Dwork, M. Hardt, T. Pitaasi, O. Reingold, and R. Zemel. Fairness through awareness. In Proc. of the 3rd Innovations in Theoretical Computer Science Conference, pages 214–226, 2012.

[10] M. Feldman, S. A. Fiedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. In Proc. of the 21th ACM SIGKDD of International Conference on Knowledge Discovery and Data Mining, pages 259–268, 2015.

[11] Y. Freund and R. E. Schapire. Game theory, online prediction and boosting. In Proc. of the 9th Annual Conference on Computational Learning Theory, pages 325–332, 1996.

[12] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and application to boosting. Journal of Computer and System Sciences, 55(1):119–139, 1997.

[13] M. Hardt, E. Price, and N. Srebor. Equality of opportunity in supervised learning. In Proc. of the 30th Conference on Neural Information Processing Systems, 2016.

[14] H. Heidari, C. Ferrari, K. P. Gummandi, and A. Krause. Fairness behind a veil of ignorance: a welfare analysis for automated decision making. In Proc. of the 32nd Conference on Neural Information Processing Systems, pages 1273–1283, 2018.

[15] M. Kearns, S. Neel, A. Roth, and Z. S. Wu. Preventing fairness gerrymandering: auditing and learning for subgroup fairness. In Proc. of the 35th International Conference on Machine Learning, 2018.

[16] M. Kearns, S. Neel, A. Roth, and Z. S. Wu. An empirical study of rich subgroup fairness for machine learning. In Proc. of the Conference on Fairness, Accountability, and Transparency, 2019.

[17] M. Kearns, A. Roth, and S. Sharifi-Malvajerdi. Average individual fairness: Algorithms, generalization and experiments. In Proc. of the 33rd Conference on Neural Information Processing Systems, 2019.

[18] J. Kleinberg, S. Mullainathan, and M. Raghavan. Inherent trade-offs in the fair determination of risk scores. In Proc. of the 8th Innovations in Theoretical Computer Science Conference, 2017.

[19] M. Kusner, J. Loftus, C. Russel, and R. Silva. Counterfactual fairness. In Proc. of the 31st Conference on Neural Information Processing Systems, 2017.

[20] M. Lichman. UCI machine learning repository. http://archive.ics.uci.edu/ml 2013.

[21] M. Mohr, A. Rostamizadeh, and A. Talwalkar. Foundations of Machine Learning. The MIT Press, 2nd edition, 2018.

[22] R. Motwani and P. Raghavan. Randomized algorithms. Cambridge University Press, 1995.

[23] W. Rudin. Principles of mathematical analysis. McGraw-Hill, 3rd edition, 1976.

[24] W. Rudin. Real and complex analysis. McGraw-Hill, 3rd edition, 1987.
Appendix

A Experimental Setup and Supplementary Experiments

A.1 Data and Implementation

The data sets we used are

- Adult income data set [20]: The task is to predict if a person’s income is no less than $50K per year. The population is partitioned to male, corresponding to $g = 0$, and female, corresponding to $g = 1$. The label value 1 indicates that the income of an individual is greater than or equal to $50K per year.

- COMPAS recidivism data set [2]: Our experiments have used data samples whose race attribute is either Caucasian or African American. The task is to predict if an individual is rearrested within two years after the first arrest. The population is partitioned to Caucasian, corresponding to $g = 0$, and African American attributes, corresponding to $g = 1$. The label value 1 indicates no re-arrest within two years and the label value 0 indicates re-arrest within two years.

- Law school data set [29]: The task is to predict if a student passes the bar exam. The population is partitioned to male, corresponding to $g = 0$, and female, corresponding to $g = 1$. The label value 1 indicates that a student passes the bar exam.
• Dutch census data set [7]: The task is to predict if an individual has a prestigious
occupation. The population is partitioned to male, corresponding to \( g = 0 \), and female,
corresponding to \( g = 1 \). The label value 1 indicates that an individual has a prestigious
job.

Algorithm 1 assumes the existence of an oracle \( W(\lambda) \). For the implementation of an
oracle, finding \( \hat{h}(t) = \min_{h \in H} L(h, \lambda) \) for \( \lambda \in \{0, \lambda_{\text{max}}\} \), in item 2 of Algorithm 1, we use
thresholding [25], a simple technique to directly find the best decision threshold for a given
objective from the training data and use this predict the class label for test data. A smaller
probability than this threshold then it is classified as 0, otherwise as 1. Logistic regression has
been used as a base classifier for thresholding. We describe below the thresholding technique
used for the implementation of oracle finding \( \hat{h}(t) = \min_{h \in H} L(h, \lambda) \) for \( \lambda \in \{0, \lambda_{\text{max}}\} \), in
item 2 of Algorithm 1. First, we train logistic regression with the training data set of adult.
During the training, 0.5 is the typical threshold value of logistic regression for decision: the
label of an instance is predicted as 1 if its predicted probability for the positive class is higher
than or equal to \( \frac{1}{2} \), otherwise, its label is decided as 0. Second, we divide the interval \([0, 1]\) into
201 points with step size \( \frac{5}{1000} \). Each point of 201 points is used a threshold and plays the role
of a hypothesis. Third, for each threshold, we predict the labels of instances by comparing the
threshold value and the probability of the positive class: for a give instance, if its probability
of the positive class is higher or equal to the threshold value, then the label of the instance is
1, otherwise, the label is 0. Finally, to find the oracle for a given \( \lambda \), we examine the value of
\( L(h, \lambda) \) for every \( h \in H \) and take \( \hat{h} = \arg \min_{h \in H} L(h, \lambda) \).

We use \( \lambda_{\text{max}} = 20, \nu = 0.005 \). Regarding to the value of \( b_i = a(h(x_i) - y_i) + c \), we fix the
value of \( a \) as 5, and change the values of \( c \in \{8, 9, 10\} \) to investigate the effect of \( r = \frac{c}{a} \) on
the performance for all data sets except Dutch census data set. For Dutch data set, we use
\( \{9, 9.5, 10\} \) for \( c \) values.

Computational Resources All experiments were run on a server with about 250GB
RAM. The server is not equipped with GPU acceleration. About 10 minutes is the run-time of
an experiment to generate all related graphs for a fixed \( \alpha \) and the set of \( c \) values.
A.2 Supplementary Experiments on COMPAS data set

This subsection provides the plots illustrating the change of test error and test $I_{\alpha}$ obtained by the hypothesis $\bar{D}$ of Algorithm 1 in Figures 12 and 13. The behaviors of them are similar to those of adult income data set. The convergence behaviors of empirical error and $I_{\alpha}$ are shown in Figures 14-15.

Figure 12: COMPAS: Averaged test error when $a = 5$ ($x$ axis is $\gamma$)

Figure 13: COMPAS: Averaged test $I_{\alpha}$ when $a = 5$ ($x$ axis is $\gamma$)

\[\text{(a) } \alpha = 0 \quad \text{(b) } \alpha = 1 \quad \text{(c) } \alpha = 2\]
Figure 14: COMPAS: Empirical error trace when $a = 5$ ($x$ axis is $\gamma$)

Figure 15: COMPAS: Close look of Figure 14 for $1 \leq t \leq 5500$ ($x$ axis is $\gamma$)

Figure 16: COMPAS: Empirical $I_\alpha$ trace when $a = 5$ ($x$ axis is $\gamma$)

Figure 17: COMPAS: Close look of Figure 16 for $1 \leq t \leq 5500$ ($x$ axis is $\gamma$)
A.3 Experiments on The Dutch Census Data set

Figure 18: Dutch census: Averaged test error when $\alpha = 5$ (x axis is $\gamma$)

Figure 19: Dutch census: Averaged test $I_\alpha$ when $a = 5$ (x axis is $\gamma$)

Figure 20: Dutch Census: $I_\alpha$ and $V$ when $\alpha = 1$

Figure 21: Dutch Census: $r_g^{FN}$ and $r_g^{FP}$ when $\alpha = 1$
Figure 22: Dutch census: Empirical error trace when $a = 5$ ($x$ axis is $\gamma$)

Figure 23: Dutch census: Close look of Figure 22 for $1 \leq t \leq 5500$ ($x$ axis is $\gamma$)

Figure 24: Dutch census: Empirical $I_\alpha$ trace when $a = 5$ ($x$ axis is $\gamma$)

Figure 25: Dutch census: Close look of Figure 24 for $1 \leq t \leq 5500$ ($x$ axis is $\gamma$)
A.4 Experiments on Law School Admissions Data Set

Figure 26: Law school: Averaged test error when $a = 5$ (x axis is $\gamma$)

Figure 27: Law school: Averaged test $I_\alpha$ when $a = 5$ (x axis is $\gamma$)

Figure 28: Law school: $I_\alpha$ and $V$ when $\alpha = 1$

Figure 29: Law school: $r_{gF}^{FN}$ and $r_{gF}^{FP}$ when $\alpha = 1$
Figure 30: Law school: Empirical $I_\alpha$ trace when $a = 5$ ($x$ axis is $\gamma$)

Figure 31: Law school: Close look of Figure 30 for $1 \leq t \leq 5500$ ($x$ axis is $\gamma$)

Figure 32: Law school: Empirical $I_\alpha$ trace when $a = 5$ ($x$ axis is $\gamma$)

Figure 33: Law school: Close look of Figure 32 for $1 \leq t \leq 5500$ ($x$ axis is $\gamma$)
Table 1: Classifier $h_1$

| Group | $X^1$ | $X^2$ |
|-------|-------|-------|
| Individuals $x_i$ | $x_1$ $x_2$ $x_3$ $x_4$ $x_5$ | $x_6$ $x_7$ $x_8$ $x_9$ $x_{10}$ |
| $y_i$: true label | 1 1 1 1 0 | 1 0 1 1 1 |
| $h_1(x_i)$: predicted label | 1 0 1 0 1 | 1 1 1 1 0 |
| C, FP, FN | C FN C FN FP | C FP C C FN |
| $b_i$ | 3 2 3 2 4 | 3 4 3 3 2 |

B A Computation Example of Additive Decomposability

We explain with an example how to compute generalized entropy $I_\alpha(b; n)$ and check additive decomposability, \( [4] \),

\[
I(b^1, \ldots, b^G; n) = \sum_g w^G_g(\mu, n) I(b^g; n_g) + V
\]

We set $\alpha = 1$ in this example and use $I$ instead of $I_\alpha$. Consider 10 individuals \( \{x_1, x_2, \ldots, x_{10}\} \). There are two groups $X^1$ and $X^2$ such that $X^1 = \{x_1, \ldots, x_5\}$ and $X^2 = \{x_6, \ldots, x_{10}\}$. The true label is denoted by $y_i$ and predicted label by $h(x_i)$. The benefit value $b_i$ is defined by $b_i = h(x_i) - y_i + 3$ for a given classifier $h$. Then $b = (b^1, b^2) = (b_{x_1}, \ldots, b_{x_{10}})$. For groups $X^1$ and $X^2$, we have $b^1 = (b_1, \ldots, b_5)$ and $b^2 = (b_6, \ldots, b_{10})$. For this special case of groups $X^1$ and $X^2$, the Axiom 4 (the additive decomposability) is written by

\[
I(b_1, \ldots, b_{10}; 10) = w_1 I(b_1, \ldots, b_5; 5) + w_2 I(b_6, \ldots, b_{10}; 5) + V
\]

if we use a simple notation $w_i$ instead of $w^G_i(\mu, n)$.

Note that $I(b_1, \ldots, b_5; 5)$ is the generalized entropy for group $X^1$ whose members have benefits $b_1, b_2, \ldots, b_5$. Similarly $I(b_6, \ldots, b_{10})$ is the generalized entropy for group $X^2$ with the benefits, $b_6, b_7, \ldots, b_{10}$. Consider a classifier $h_1$ whose prediction $h_1(x_i)$ is given in Table 1. We will use superscript $h_1$ to make clear which classifier is used, if necessary.

Let $n_C$ be the number of correct labels, $n_{FP}$ the number of false positive labels, and $n_{FN}$ the number of false negative labels. Then the average of the value $b_i$’s and generalized entropy $I^{h_1}(b_{h_1}; 10)$ for the whole population are

\[
\mu^{h_1} = \frac{1}{10} \sum_{i=1}^{10} b_i = \frac{1}{10} \left( n_C * 3 + n_{FP} * 4 + n_{FN} * 2 \right) = \frac{29}{10},
\]

\[
I^{h_1}(b_{h_1}; 10) = \frac{1}{10} \left( n_C * \frac{3}{\mu^{h_1}} \ln \frac{3}{\mu^{h_1}} + n_{FP} * \frac{4}{\mu^{h_1}} \ln \frac{4}{\mu^{h_1}} + n_{FN} * \frac{2}{\mu^{h_1}} \ln \frac{2}{\mu^{h_1}} \right)
\]

\[
= 0.0848.
\]
Let’s compute the generalized entropy of group $X^1$ that has five individuals $x_1, ..., x_5$. Consider the average of $b_i$ for the individuals of $X^1$, i.e.,

$$
\mu^{h_1}(X^1) = \frac{1}{5} \sum_{i=1}^{5} b_i = \frac{2 \cdot 3 + 1 \cdot 4 + 2 \cdot 2}{5} = \frac{14}{5}.
$$

The generalized entropy for $X^1$ is

$$
I^{h_1}(b_{x_1}, ..., b_{x_5}; 5) = \frac{1}{5} \left( 2 \cdot \frac{3}{\mu^{h_1}(X^1)} \ln \frac{3}{\mu^{h_1}(X^1)} + 1 \cdot \frac{4}{\mu^{h_1}(X^1)} \ln \frac{4}{\mu^{h_1}(X^1)} + 2 \cdot \frac{2}{\mu^{h_1}(X^1)} \ln \frac{2}{\mu^{h_1}(X^1)} \right)
= 0.0990.
$$

Similarly, for group $X^2 = \{x_6, ..., x_{10}\}$, we have $\mu^{h_1}(X^2) = 3$ and $I^{h_1}(b_{x_6}, ..., b_{x_{10}}; 5) = 0.0680$ Now we can find $w_1^{h_1} = \frac{n_1}{\mu^{h_1}(X^1)} = \frac{5}{10} \frac{14}{29} = \frac{14}{55}$ and $w_2^{h_1} = \frac{5}{10} \frac{3}{29} = \frac{15}{29}$.

Now consider $V$. To compute $V$, we consider two groups, $\{j_1, j_2, ..., j_5\}$ and $\{j_6, j_7, ..., j_{10}\}$, and the benefit of the individuals which is such that $b_{j_1} = b_{j_2} = ... = b_{j_5} = \mu(b_{j_1}, ..., b_{j_5}) = \frac{14}{5}$ and $b_{j_6} = b_{j_7} = ... = b_{j_{10}} = \mu(b_{j_6}, ..., b_{j_{10}}) = 3$. Note that every member in $\{j_1, ..., j_5\}$ have the identical benefit $\frac{14}{5}$ and every member in $\{j_6, ..., j_{10}\}$ have identical benefit 3. Definitely the average value of the benefit for the individuals $j_1, ..., j_{10}$ is $\mu^{h_1}$. Hence

$$
V^{h_1} = I(b_{j_1}, ..., b_{j_{10}}; 10) = \frac{1}{10} \left( 5 \cdot \frac{\mu(b_{j_1}, ..., b_{j_5})}{\mu^{h_1}} \ln \frac{\mu(b_{j_1}, ..., b_{j_5})}{\mu^{h_1}} + 5 \cdot \frac{\mu(b_{j_6}, ..., b_{j_{10}})}{\mu^{h_1}} \ln \frac{\mu(b_{j_6}, ..., b_{j_{10}})}{\mu^{h_1}} \right)
= \frac{1}{10} \left( 5 \cdot \frac{14/5}{29/10} \ln \frac{14/5}{29/10} + 5 \cdot \frac{3}{29/10} \ln \frac{3}{29/10} \right)
= 0.0018.
$$

Indeed we can check that $I(b_{x_1}, ..., b_{x_{10}}; 10) = w_1^{h_1} I^{h_1}(X^1; 5) + w_2^{h_1} I^{h_1}(X^2; 5) + V^{h_1}$ with the values of $w_1^{h_1} I^{h_1}(X^1; 5) = 0.0478$, $w_2^{h_1} I^{h_1}(X^2; 5) = 0.0352$, and $V^{h_1} = 0.0018$.

### C Proof of Theorem 4.2

Obviously Axioms 1, 3’, and 7 hold.

**Theorem C.1** (Jensen’s Inequality, [24]). Let $Q$ be a probability distribution on $\mathcal{X}$ and $dQ$ the probability density function of $Q$. If $b$ is a real function with $\int_{\mathcal{X}} |b(x)|dQ$, $a_1 < b(x) < a_2$ for all $x \in \mathcal{X}$, and $\phi$ is convex on $(a_1, a_2)$, then

$$
\phi\left( \int_{\mathcal{X}} b(x)dQ \right) \leq \int_{\mathcal{X}} \phi(b(x))dQ.
$$

32
C.1 Proof for Axiom 2′

Note that $f_{\alpha}(x)$ is convex over $\mathbb{R}$ for all $\alpha \in [0, \infty)$. We will show that $I_1(b, \mathcal{X}, P_x) \geq 0$ for $\alpha = 1$: recall that $b(x) = a(h(x) - y) + c$.

\[
I_1(b, \mathcal{X}, P_x) = \int_{\mathcal{X}} \frac{b(x)}{\mu} \ln \left( \frac{b(x)}{\mu} \right) dP_x
\]

\[
= \frac{1}{\mu} \int_{\mathcal{X}} b(x) \ln b(x) dP_x - \frac{1}{\mu} \int_{\mathcal{X}} (b(x) \ln \mu) dP_x
\]

\[
= \frac{1}{\mu} \left( \int_{\mathcal{X}} b(x) \ln b(x) dP_x - \mu \ln \mu \right)
\]

\[
\geq \frac{1}{\mu} \left[ f_1 \left( \int_{\mathcal{X}} b(x) \ln b(x) dP_x \right) - \mu \ln \mu \right]
\]

\[
\geq 0.
\]

The first inequality holds by Jensen’s inequality since $b(x)$ is bounded and $f_1(x)$ is convex. In a similar way, we can show that $I_\alpha(b, \mathcal{X}, P_x) \geq 0$ for $\alpha \neq 1$.

C.2 Proof for Axiom 4′ (Additive decomposability)

Recall that $v(x) = \sum_{g=1}^{G} \mu_g \mathbbm{1}\{x \in \mathcal{X}^g\}$ and $m_g = P_x(\mathcal{X}^g)$. For any $A \subset \mathcal{X}^g$, we have $P^g_x(A) = \frac{P_x(A)}{m_g} = \frac{P_x(\mathcal{X}^g)}{m_g}$. Hence $P_x = P^g_x$ and $dP_x = m_g dP_x^g$ for all $g$.

Lemma C.2.

\[
\mathbb{E}[v(x)] = \mu = \sum_{g=1}^{G} m_g \mu_g.
\]

Proof. Since $\mathcal{X}$ is a disjoint union of $\mathcal{X}^g$ for $1 \leq g \leq G$, we have

\[
\mu = \int_{\mathcal{X}} b(x) dP_x = \int_{\bigcup_{g=1}^{G} \mathcal{X}^g} b(x) dP_x = \sum_{g=1}^{G} \int_{\mathcal{X}^g} b(x) dP_x = \mu
\]

(13)

where $\bigcup$ means a disjoint union. Then

\[
\sum_{g=1}^{G} \int_{\mathcal{X}^g} b(x) dP_x = \sum_{g=1}^{G} m_g \int_{\mathcal{X}^g} b(x) dP_x = \sum_{g=1}^{G} m_g \mu_g = \mu.
\]

(14)

Combining (13) and (14), we have $\mu = \sum_{g=1}^{G} m_g \mu_g$. Consider $\mathbb{E}[v(x)]$:

\[
\mathbb{E}[v(x)] = \mathbb{E} \left[ \sum_{g=1}^{G} \mu_g \mathbbm{1}\{x \in \mathcal{X}^g\} \right] = \sum_{g=1}^{G} \mu_g \mathbb{E} \left[ \mathbbm{1}\{x \in \mathcal{X}^g\} \right] = \sum_{g=1}^{G} m_g \mu_g = \mu.
\]

\[\square\]
Lemma C.3. For \( \alpha \in [0, \infty) \), it holds that

\[
I_\alpha(b, \mathcal{X}, P_x) = \sum_{g=1}^{G} m_g f_\alpha \left( \frac{\mu_g}{\mu} \right).
\]

Proof. Since \( \mathcal{X} = \bigcup \mathcal{X}^g \), for each \( x \in \mathcal{X} \), there is a unique \( g \) such that \( x \in \mathcal{X}^g \). Therefore

\[
v(x) = \sum_{i=1}^{G} \mu_i \mathbb{1}\{x \in \mathcal{X}^i\} = \mu_g \quad \text{for} \quad x \in \mathcal{X}^g.
\]

(15)

Consider \( I_\alpha(v, \mathcal{X}, P_x) \);

\[
I_\alpha(v, \mathcal{X}, P_x) = \int_{\mathcal{X}} f_\alpha \left( \frac{v(x)}{E[v(x)]} \right) dP_x
= \int_{\mathcal{X}} f_\alpha \left( \frac{v(x)}{\mu} \right) dP_x \quad \text{(by Lemma C.2)}
= \sum_{g=1}^{G} \int_{\mathcal{X}^g} f_\alpha \left( \frac{\mu_g}{\mu} \right) dP_x \quad \text{(by (15))}
= \sum_{g=1}^{G} m_g f_\alpha \left( \frac{\mu_g}{\mu} \right).
\]

\( \square \)

i) \( \alpha = 1 \)

Using \( \mathcal{X} = \bigcup \mathcal{X}^g \), we have

\[
I_1(b, \mathcal{X}, P_x) = \int_{\bigcup \mathcal{X}^g} \frac{b(x)}{\mu} \ln \frac{b(x)}{\mu} dP_x
= \sum_{g=1}^{G} \int_{\mathcal{X}^g} \frac{b(x)}{\mu} \ln \frac{b(x)}{\mu} dP_x
= \sum_{g=1}^{G} m_g \int_{\mathcal{X}^g} \frac{b(x)}{\mu} \ln \frac{b(x)}{\mu} dP^g_x.
\]

(16)

Since \( \frac{b(x)}{\mu} \ln \frac{b(x)}{\mu} = \frac{\mu_g b(x)}{\mu_g} \left( \ln \frac{b(x)}{\mu_g} + \ln \frac{\mu_g}{\mu} \right) \), we have

\[
\int_{\mathcal{X}^g} \frac{b(x)}{\mu} \ln \frac{b(x)}{\mu} dP^g_x = \frac{\mu_g}{\mu} \left( \int_{\mathcal{X}^g} \frac{b(x)}{\mu_g} \ln \frac{b(x)}{\mu_g} dP^g_x + \ln \frac{\mu_g}{\mu} \int_{\mathcal{X}^g} \frac{b(x)}{\mu_g} dP^g_x \right)
= \frac{\mu_g}{\mu} \left( I_1(b|\mathcal{X}^g, X^g, P^g_x) + \ln \frac{\mu_g}{\mu} \right) \quad (\because \mu_g = \int_{\mathcal{X}^g} b(x) dP^g_x) \quad (17)
\]

34
By \( (17) \), equation \( (16) \) becomes

\[
I_{\alpha}(b, X, P_x) = \sum_{g=1}^{G} m_g \frac{\mu_g}{\mu} \left( I_{\alpha}(b|X^g, X_g, P^g_x) + \ln \frac{\mu_g}{\mu} \right)
\]

\[
= \sum_{g=1}^{G} m_g \frac{\mu_g}{\mu} I_{\alpha}(b|X^g, X_g, P^g_x) + \sum_{g=1}^{G} m_g \frac{\mu_g}{\mu} \ln \frac{\mu_g}{\mu}
\]

By Lemma \( C.3 \) we have

\[
I_1(\upsilon, X, P_x) = \sum_{g=1}^{G} m_g f_1 \left( \frac{\mu_g}{\mu} \right) = \sum_{g=1}^{G} m_g \frac{\mu_g}{\mu} \ln \frac{\mu_g}{\mu}
\]

Hence we have proved that the extended generalized entropy has the additive decomposable property when \( \alpha = 1 \).

ii) \( \alpha = 0 \)

This case can be shown similarly as in the case \( \alpha = 1 \). So we omit the detail of it.

iii) \( \alpha \neq 0, 1 \) (and \( 0 < \alpha < \infty \))

\[
I_{\alpha}(b, X, P_x) = \frac{1}{\alpha(\alpha-1)} \int_{X} \left[ \left( \frac{b(x)}{\mu} \right)^\alpha - 1 \right] dP_x
\]

\[
= \frac{1}{\alpha(\alpha-1)} \sum_{g=1}^{G} \int_{X^g} \left[ \left( \frac{b(x)}{\mu} \right)^\alpha - 1 \right] dP_x
\]

\[
= \frac{1}{\alpha(\alpha-1)} \sum_{g=1}^{G} m_g \int_{X^g} \left[ \left( \frac{b(x)}{\mu} \right)^\alpha - 1 \right] dP^g_x
\]

Since \( \left( \frac{b(x)}{\mu} \right)^\alpha - 1 = \left( \frac{\mu_g}{\mu} \right)^\alpha \left[ \left( \frac{b(x)}{\mu} \right)^\alpha - 1 \right] + \left( \frac{\mu_g}{\mu} \right)^\alpha - 1 \), we have

\[
\int_{X^g} \left[ \left( \frac{b(x)}{\mu} \right)^\alpha - 1 \right] dP^g_x = \int_{X^g} \left( \frac{\mu_g}{\mu} \right)^\alpha \left[ \left( \frac{b(x)}{\mu} \right)^\alpha - 1 \right] dP^g_x + \int_{X^g} \left( \frac{\mu_g}{\mu} \right)^\alpha - 1 dP^g_x
\]

\[
= \left( \frac{\mu_g}{\mu} \right)^\alpha (\alpha-1) I_{\alpha}(b|X^g, X_g^g, P^g_x) + \left( \frac{\mu_g}{\mu} \right)^\alpha - 1
\]

(because \( \mu_g = \int_{X^g} b(x) dP^g_x \)).

By \( (18) \), we have

\[
I_{\alpha}(b, X, P_x) = \sum_{g=1}^{G} m_g \left( \frac{\mu_g}{\mu} \right)^\alpha I_{\alpha}(b|X^g, X_g^g, P^g_x) + \frac{1}{\alpha(\alpha-1)} \sum_{g=1}^{G} m_g \left[ \left( \frac{\mu_g}{\mu} \right)^\alpha - 1 \right]
\]

\[
= \sum_{g=1}^{G} m_g \left( \frac{\mu_g}{\mu} \right)^\alpha I_{\alpha}(b|X^g, X_g^g, P^g_x) + \sum_{g=1}^{G} m_g f_\alpha \left( \frac{\mu_g}{\mu} \right)
\]

\[
= \sum_{g=1}^{G} m_g \left( \frac{\mu_g}{\mu} \right)^\alpha I_{\alpha}(b|X^g, X_g^g, P^g_x) + I_{\alpha}(b, X, P_x)
\]

(By Lemma \( C.3 \)).
C.3  Proof for Axiom 6’

It can be easily checked that \( P_x(M_i) = P_x(M_j) \) from the assumption that \( \mathbb{E}[b(x)] = \mathbb{E}[\tilde{b}(x)] = \mu \).

We need a simple property of a real-valued differentiable convex function, which is stated in (McDiarmid’s Inequality).

**Theorem D.1**

Theorem D.1

**Proof of Theorem 4.4**

**Lemma C.4.** Suppose that \( f : \mathbb{R} \to \mathbb{R} \) is a differentiable real-valued function. Then \( f \) is convex if and only if

\[
f(x_2) \geq f(x_1) + f'(x_1)(x_2 - x_1)
\]

Note that \( f_\alpha(x) \) in (7) is a real-valued convex function defined on \( \mathbb{R} \) for all \( \alpha \in [0, \infty) \). Therefore Lemma C.4 holds and moreover \( f_\alpha'(x_2) > f_\alpha'(x_1) \) if \( x_2 > x_1 \).

We will show that \( I_\alpha(b, \mathcal{X}, P_x) - I_\alpha(b, \mathcal{X}, P_x) \geq 0 \).

\[
I_\alpha(b, \mathcal{X}, P_x) - I_\alpha(b, \mathcal{X}, P_x)
= \int_{\mathcal{X}} f_\alpha\left(\frac{b(x)}{\mu}\right) dP_x - \int_{\mathcal{X}} f_\alpha\left(\frac{\tilde{b}(x)}{\mu}\right) dP_x
= \int_{M_i \cup M_j} f_\alpha\left(\frac{b(x)}{\mu}\right) dP_x - \int_{M_i \cup M_j} f_\alpha\left(\frac{\tilde{b}(x)}{\mu}\right) dP_x
= \int_{M_i} f_\alpha\left(\frac{b(x)}{\mu}\right) dP_x + \int_{M_j} f_\alpha\left(\frac{b(x)}{\mu}\right) dP_x - \int_{M_i} f_\alpha\left(\frac{\tilde{b}(x)}{\mu}\right) dP_x + \int_{M_j} f_\alpha\left(\frac{\tilde{b}(x)}{\mu}\right) dP_x
= \int_{M_i} f_\alpha\left(\frac{b(x)}{\mu}\right) dP_x + \int_{M_j} f_\alpha\left(\frac{b(x)}{\mu}\right) dP_x - \int_{M_i} f_\alpha\left(\frac{b(x) - \delta}{\mu}\right) dP_x + \int_{M_j} f_\alpha\left(\frac{b(x) + \delta}{\mu}\right) dP_x
= \int_{M_i} \left[f_\alpha\left(\frac{b(x)}{\mu}\right) - f_\alpha\left(\frac{b(x) - \delta}{\mu}\right)\right] dP_x - \int_{M_j} \left[f_\alpha\left(\frac{b(x) + \delta}{\mu}\right) - f_\alpha\left(\frac{b(x)}{\mu}\right)\right] dP_x
= \int_{M_i} f_\alpha'\left(\frac{b(x) - \delta}{\mu}\right) \delta dP_x - \int_{M_j} f_\alpha'\left(\frac{b(x)}{\mu}\right) \delta dP_x \quad \text{(by Lemma C.4)}
= \delta P_x(M_i) \left[f_\alpha\left(\frac{b(x) - \delta}{\mu}\right) - f_\alpha\left(\frac{b(x)}{\mu}\right)\right]
\geq 0 \quad \text{(because } P_x(M_i) = P_x(M_j) \text{ and the convexity of } f_\alpha).
\( x \), the function satisfies
\[
\sup_{x \in \mathcal{X}} \phi(x_1, \ldots, x_{i-1}, x, x_{i+1}, \ldots, x_n) - \inf_{x \in \mathcal{X}} \phi(x_1, \ldots, x_{i-1}, x, x_{i+1}, \ldots, x_n) \leq q_i. \tag{19}
\]
with \( q_i > 0 \), then
\[
P\left[ \left| \phi(X_1, \ldots, X_n) - \mathbb{E}[\phi(X_1, \ldots, X_n)] \right| \geq \varepsilon \right] \leq 2 \exp\left( \frac{-2 \varepsilon^2}{\sum_{i=1}^{n} q_i^2} \right).
\]

**Lemma D.3.** The function \( f(x) = x \ln x \) is non-negative, increasing and convex for \( x \geq \frac{1}{e} \).

**Proof.** Obviously \( f(x) \geq 0 \) for \( x \geq 1 \). The first derivative is \( f'(x) = \ln x + 1 > 0 \) for \( x \geq 1 \); \( f(x) \) is increasing over \( x \geq \frac{1}{e} \). The second derivative is \( f''(x) = \frac{1}{x} > 0 \) for \( x > 0 \); \( f(x) \) is convex over \( x \geq \frac{1}{e} \). \( \square \)

For the purpose of distinguishing \( b_i \) and \( b(x) \) and simple notation, we use \( \beta \) for \( b(x) \) instead of \( b(x) \): for example, \( \mathbb{E}[\beta] \) stands for \( \mathbb{E}[b(x)] \) and \( \mathbb{E}[\beta \ln(\beta)] \) for \( \mathbb{E}[b(x) \ln b(x)] \). Moreover, we let \( \omega = \mathbb{E}[b(x)] = \mathbb{E}[\beta] \).

**D.1 When \( \alpha = 0 \)**

Recall that \( \mu = \frac{1}{n} \sum_{i=1}^{n} b_i \) where \( b_i = a(h_{(i)} - y_i) + c \) with \( c > a > 0 \) and \( c - a \geq 1 \). Consider \( |I_0(h, P) - I_0(b; n)| \):
\[
|I_0(h, P) - I_0(b; n)| = \left| \int_{\mathcal{X}} \ln \left( \frac{\mathbb{E}[b(x)]}{b(x)} \right) dP_x - \frac{1}{n} \sum_{i=1}^{n} \ln \frac{\mu}{b_i} \right|
\]
\[
= \left| \ln \mathbb{E}[\beta] - \ln \mathbb{E}[\beta] - \left( \ln \mu - \frac{1}{n} \sum_{i=1}^{n} \ln b_i \right) \right|
\]
\[
= \left| \ln \mathbb{E}[\beta] - \ln \mu - \left( \mathbb{E}[\ln \beta] - \frac{1}{n} \sum_{i=1}^{n} \ln b_i \right) \right|
\]
\[
\leq \left| \ln \mathbb{E}[\beta] - \ln \mu \right| + \left| \mathbb{E}[\ln \beta] - \frac{1}{n} \sum_{i=1}^{n} \ln b_i \right| \tag{20}
\]
Let \( M = \max(\mu, \mathbb{E}[\beta]) \) and \( m = \min(\mu, \mathbb{E}[\beta]) \). Then
\[
\left| \ln \mathbb{E}[\beta] - \ln \mu \right| = \ln \left( 1 + \frac{M - m}{m} \right)
\]
\[
\leq \frac{M - m}{m} \left( = \frac{\mathbb{E}[\beta] - \mu}{m} \right) \quad \text{(since } \ln(1 + x) < x \text{)} \tag{21}
\]
\[
\leq \frac{\mathbb{E}[\beta] - \mu}{c - a} \quad \text{(since } \mu \geq c - a, \mathbb{E}[\beta] \geq c - a \text{).} \tag{22}
\]
Then, \( \text{(20)} \) becomes
\[
\left| I_0(h, P) - I_0(b; n) \right| \leq \frac{|\mathbb{E}[\beta] - \mu|}{c - a} + \left| \mathbb{E}[\ln \beta] - \frac{1}{n} \sum_{i=1}^{n} \ln b_i \right|. \tag{23}
\]
Lemma D.4. With probability at least $1 - \frac{\delta}{2}$, we have

$$\left| \mathbb{E}[\beta] - \mu \right| < a \sqrt{\frac{2}{n} \ln \frac{4}{\delta}}. \quad (24)$$

Proof. For $|\mathbb{E}[\beta] - \mu|$, we will apply McDiarmid’s Inequality to the function $\phi(x_1, \ldots, x_n) = \frac{1}{n} \sum_{i=1}^{n} b(x_i)$. Note that $\phi(x_1, \ldots, x_n)$ satisfies (19) with $q_i = \frac{2a}{n}$ since $b(x) : \mathcal{X} \rightarrow \mathbb{R}$ takes one of $\{c - a, c, c + a\}$ for all $i$. By McDiarmid’s Inequality, we have

$$P \left[ \left| \frac{1}{n} \sum_{i=1}^{n} b_i - \mathbb{E}[\beta] \right| > \varepsilon \right] \leq 2 \exp \left( \frac{-2n\varepsilon^2}{4a^2} \right).$$

If we take $2 \exp \left( \frac{-2n\varepsilon^2}{4a^2} \right) = \frac{\delta}{2}$, then with probability at least $1 - \frac{\delta}{2}$, (24) holds.

Lemma D.5. With probability at least $1 - \frac{\delta}{2}$, we have

$$\left| \mathbb{E}[\ln \beta] - \frac{1}{n} \sum_{i=1}^{n} \ln b_i \right| \leq \ln \left( \frac{c + a}{c - a} \right) \sqrt{\frac{1}{2n} \ln \frac{4}{\delta}}. \quad (25)$$

Proof. For $|\mathbb{E}[\ln \beta] - \frac{1}{n} \sum_{i=1}^{n} \ln b_i|$, we set $\phi(x_1, \ldots, x_n) = \frac{1}{n} \sum_{i=1}^{n} \ln b(x_i)$. Then, $\phi(x_1, \ldots, x_n)$ satisfies (19) with $q_i = \ln \left( \frac{c + a}{c - a} \right)$ since $b(x) : \mathcal{X} \rightarrow \mathbb{R}$ takes one of $\{c - a, c, c + a\}$ for all $i$. Moreover $\mathbb{E}[\phi(x_1, \ldots, x_n)] = \mathbb{E}\left[ \frac{1}{n} \sum_{i=1}^{n} \ln b(x_i) \right] = \mathbb{E}[\ln b(x)] = \mathbb{E}[\ln \beta]$. By McDiarmid’s Inequality, we have

$$P \left[ \left| \frac{1}{n} \sum_{i=1}^{n} \ln b_i - \mathbb{E}[\ln \beta] \right| > \varepsilon \right] \leq 2 \exp \left( \frac{-2n\varepsilon^2}{\ln^2 \left( \frac{c + a}{c - a} \right)} \right).$$

If we take $2 \exp \left( \frac{-2n\varepsilon^2}{\ln^2 \left( \frac{c + a}{c - a} \right)} \right) = \frac{\delta}{2}$, then with probability at least $1 - \frac{\delta}{2}$, we can check that (25) holds.

By (24), (25), and union bounds, (23) becomes

$$\left| I_0(h, P) - I_0(b; n) \right| \leq \left( \frac{2a}{c - a} + \ln \left( \frac{c + a}{c - a} \right) \right) \sqrt{\frac{1}{2n} \ln \frac{4}{\delta}} \leq \left[ \frac{2}{r - 1} + \ln \left( \frac{r + 1}{r - 1} \right) \right] \sqrt{\frac{1}{2n} \ln \frac{4}{\delta}}. $$

with probability at least $1 - \delta$. 38
D.2 When $\alpha = 1$

Recall $\omega = \mathbb{E}[b(x)] = \int_X b(x) dP_x$. Then, we have

\[
I_1(h, P) = \frac{1}{\omega} \int_X \left( \frac{b(x)}{\mathbb{E}[b(x)]} \ln \left( \frac{b(x)}{\mathbb{E}[b(x)]} \right) \right) dP_x
\]

\[
= \frac{1}{\omega} \int_X b(x) \ln \left( \frac{b(x)}{\omega} \right) dP_x
\]

\[
= \frac{1}{\omega} \mathbb{E}[b(x) \ln b(x)] - \ln \omega
\]

(26)

Recall that $I_1(b; n) = \frac{1}{n} \sum_{i=1}^{n} b_i \ln \frac{b_i}{\mu}$;

\[
I_1(b; n) = \frac{1}{n\mu} \left( \sum_{i=1}^{n} b_i \ln b_i - \sum_{i=1}^{n} b_i \ln \mu \right) = \left( \frac{1}{n\mu} \sum_{i=1}^{n} b_i \ln b_i \right) - \ln \mu.
\]

(27)

Then,

\[
|I_1(h, P) - I_1(b; n)| = \left| \frac{\mathbb{E}[\beta \ln \beta]}{\omega} - \frac{1}{n\mu} \sum_{i=1}^{n} b_i \ln b_i - \left( \ln \omega - \ln \mu \right) \right|
\]

\[
\leq \left| \frac{\mathbb{E}[\beta \ln \beta]}{\omega} - \frac{1}{n\mu} \sum_{i=1}^{n} b_i \ln b_i \right| + \left| \ln \omega - \ln \mu \right|
\]

\[
\leq \left| \frac{\mathbb{E}[\beta \ln \beta]}{\mathbb{E}[\beta]} - \frac{1}{n\mu} \sum_{i=1}^{n} b_i \ln b_i \right| + \left| \mathbb{E}[\beta] - \mu \right| \left( \frac{1}{c-a} \right)
\]

(28)

(by (22), $\omega = \mathbb{E}[\beta]$.)

Consider $\left| \frac{\mathbb{E}[\beta \ln \beta]}{\mathbb{E}[\beta]} - \frac{1}{n\mu} \sum_{i=1}^{n} b_i \ln b_i \right|$

\[
\left| \frac{\mathbb{E}[\beta \ln \beta]}{\mathbb{E}[\beta]} - \frac{1}{n\mu} \sum_{i=1}^{n} b_i \ln b_i \right| = \left| \frac{1}{\mu \mathbb{E}[\beta]} \left( \mu \mathbb{E}[\beta \ln \beta] - \frac{\mathbb{E}[\beta]}{n} \sum_{i=1}^{n} b_i \ln b_i \right) \right|
\]

\[
= \left| \frac{1}{\mu \mathbb{E}[\beta]} \left( \mu \mathbb{E}[\beta \ln \beta] - \mathbb{E}[\beta] \mathbb{E}[\beta \ln \beta] + \mathbb{E}[\beta] \mathbb{E}[\beta \ln \beta] - \frac{\mathbb{E}[\beta]}{n} \sum_{i=1}^{n} b_i \ln b_i \right) \right|
\]

\[
= \left| \mu - \mathbb{E}[\beta] \right| \left| \frac{\mathbb{E}[\beta \ln \beta]}{\mu \mathbb{E}[\beta]} \right| + \frac{1}{\mu} \left| \mathbb{E}[\beta \ln \beta] - \frac{1}{n} \sum_{i=1}^{n} b_i \ln b_i \right|.
\]

Then, (28) becomes

\[
|I_1(h, P) - I_1(b; n)|
\]

\[
\leq \left| \mu - \mathbb{E}[\beta] \right| \left| \frac{\mathbb{E}[\beta \ln \beta]}{\mu \mathbb{E}[\beta]} \right| + \frac{1}{\mu} \left| \mathbb{E}[\beta \ln \beta] - \frac{1}{n} \sum_{i=1}^{n} b_i \ln b_i \right| + \left| \mathbb{E}[\beta] - \mu \right| \left( \frac{1}{c-a} \right)
\]

\[
= \left| \mu - \mathbb{E}[\beta] \right| \left( \left| \frac{\mathbb{E}[\beta \ln \beta]}{\mu \mathbb{E}[\beta]} \right| + \frac{1}{c-a} \right) + \frac{1}{\mu} \left| \mathbb{E}[\beta \ln \beta] - \frac{1}{n} \sum_{i=1}^{n} b_i \ln b_i \right|.
\]

(29)
Regarding to $|E[\beta] - \mu|$, by (24), we know that $|E[\beta] - \mu| \leq a\sqrt{\frac{2}{n}\ln \frac{4}{\delta}}$ holds with probability at least $1 - \frac{\delta}{2}$.

Consider $\frac{|E[\beta \ln \beta]|}{\mu E[\beta]}$. Recall that $b(x) = a(h(x) - y) + c$ with $c > a > 0$ and $c - a \geq 1$; $b(x) \geq 1$ for any $x$. Note that $f_1(x) = x \ln \frac{x}{a}$ is a non-negative increasing function for $x \geq \frac{1}{2}$. Hence, since $c - a \geq 1$, it holds that $0 \leq (c - a) \ln(c - a) \leq b(x) \ln b(x) \leq (c + a) \ln(c + a)$ for all $x \in X$. From this observation, we have $(c - a) \ln(c - a) \leq E[\beta \ln \beta] \leq (c + a) \ln(c + a)$. Moreover $\min(\mu, E[\beta]) \geq c - a$, since $b(x) \geq c - a$. Therefore, it holds that

$$\left| \frac{E[\beta \ln \beta]}{\mu E[\beta]} \right| \leq \frac{(c + a) \ln(c + a)}{(c - a)^2}$$

(30)

Consider $\left| E[\beta \ln \beta] - \frac{1}{n} \sum_{i=1}^{n} b_i \ln b_i \right|$. Let $\phi(x_1, \ldots, x_n) = \frac{1}{n} \sum_{i=1}^{n} b(x_i) \ln b(x_i)$. Then $\phi(x_1, \ldots, x_n)$ satisfies (19) with $q_i = \Delta_1/n$ for all $i$, where

$$\Delta_1 = (c + a) \ln(c + a) - (c - a) \ln(c - a).$$

By McDiarmid’s Inequality

$$P \left[ \left| \frac{1}{n} \sum_{i=1}^{n} b_i \ln b_i - E[\beta \ln \beta] \right| > \epsilon \right] \leq 2 \exp \left( -\frac{2n \epsilon^2}{\Delta_1^2} \right).$$

If we take $2 \exp \left( -\frac{2n \epsilon^2}{\Delta_1^2} \right) = \frac{\delta}{2}$, then with probability at least $1 - \frac{\delta}{2}$, we have

$$\left| \frac{1}{n} \sum_{i=1}^{n} b_i \ln b_i - E[\beta \ln \beta] \right| \leq \Delta_1 \sqrt{\frac{2}{n} \ln \frac{4}{\delta}}.$$  (31)

By (30), (31) and union bounds, with probability at least $1 - \delta$, (29) becomes

$$|I_1(h, P) - I_1(b; n)| \leq \left( \frac{(c + a) \ln(c + a)}{(c - a)^2} + \frac{1}{c - a} \right) a \sqrt{\frac{2}{n} \ln \frac{4}{\delta}} + \frac{\Delta_1}{c - a} \sqrt{\frac{2}{n} \ln \frac{4}{\delta}}$$

$$= \left[ \left( \frac{c + a}{c - a} \right)^2 \ln(c + a) + \frac{2a}{c - a} - \ln(c - a) \right] \sqrt{\frac{2}{n} \ln \frac{4}{\delta}}.$$  (32)

We compute $\left( \frac{c + a}{c - a} \right)^2 \ln(c + a) + \frac{2a}{c - a} - \ln(c - a)$. Let $r = \frac{c}{a}$. Then we have

$$\left( \frac{c + a}{c - a} \right)^2 \ln(c + a) + \frac{2a}{c - a} - \ln(c - a)$$

$$= \left[ \left( \frac{r + 1}{r - 1} \right)^2 \ln \left( ar + 1 \right) + \frac{2}{r - 1} - \ln \left( ar - 1 \right) \right]$$

$$= \left[ \left( 1 + \frac{2}{r - 1} \right)^2 \ln(a(r + 1)) + \frac{2}{r - 1} - \ln(a(r - 1)) \right]$$

$$= \frac{2}{r - 1} + \ln \left( 1 + \frac{2}{r - 1} \right) + \frac{4r \ln(ar + a)}{(r - 1)^2}$$

$$\leq \frac{4}{r - 1} + \frac{4r \ln(ar + a)}{(r - 1)^2} \quad (\because \ln(1 + x) < x).$$

40
Therefore, it holds that
\[ |I_1(h, P) - I_1(b; n)| \leq \left( \frac{4}{r - 1} + \frac{4r \ln(ar + a)}{(r - 1)^2} \right) \sqrt{\frac{1}{2n} \ln \frac{4}{\delta}} \]

**D.3 When \( \alpha \neq 0, 1 \)**

Recall that we assume \( \alpha \in [0, \infty) \). Hence \( \alpha \neq 0, 1 \) means that \( \alpha \in [0, 1) \cup (1, \infty) \). When \( \alpha \neq 0, 1 \), we have

\[
|I_\alpha(h, P) - I_\alpha(b; n)| = \left| \int_X \frac{1}{\alpha(\alpha - 1)} \left[ \left( \frac{b(x)}{\mathbb{E}[b(x)]} \right)^\alpha - 1 \right] dP_x - \frac{1}{n\alpha(\alpha - 1)} \sum_{i=1}^n \left( \frac{b_i}{\mu} \right)^\alpha \right|
\]

\[
= \left| \frac{1}{\alpha(\alpha - 1)} \left( \frac{\mathbb{E}[\beta^\alpha]}{(\mathbb{E}[\beta])^\alpha} - \frac{1}{n\mu^\alpha} \sum_{i=1}^n (b_i)^\alpha \right) \right|
\]

\[
\leq \frac{1}{n(\alpha(\alpha - 1))} \left( \left| \frac{\mathbb{E}[\beta^\alpha]}{(\mathbb{E}[\beta])^\alpha} \right| \cdot |\mu^\alpha - \mathbb{E}[\beta]| + \frac{1}{\mu^\alpha} \left| \mathbb{E}[\beta] - \frac{1}{n} \sum_{i=1}^n b_i^\alpha \right| \right). \tag{32}
\]

Recall that \( M = \max(\mu, \mathbb{E}[\beta]) \) and \( m = \min(\mu, \mathbb{E}[\beta]) \). By Mean Value Theorem, there exists \( z_0 \in (m, M) \) at which

\[
|\mu^\alpha - \mathbb{E}[\beta]| = |M^\alpha - m^\alpha| = |\alpha(M - m)z_0^{\alpha - 1}| = \alpha|\mathbb{E}[\beta] - \mu| \cdot |z_0^{\alpha - 1}| \tag{33}
\]

By (24), with probability at least \( 1 - \frac{\delta}{2} \), (33) reduces

\[
|\mu^\alpha - \mathbb{E}[\beta]| \leq 2\alpha|z_0^{\alpha - 1}| \cdot 2\sqrt{\frac{1}{2n} \ln \frac{4}{\delta}} \quad \text{(by (24))} \tag{34}
\]

Since \( c - a \geq 1 \) and \( 1 \leq (c - a) \leq b(x) \leq c + a \), we have \( m \in [c - a, c + a] \) and \( M \in [c - a, c + a] \). Note that \( z_0 \in (m, M) \), hence \( z_0 \in [c - a, c + a] \). Recall that for \( 0 < \alpha < 1 \), \( f(x) = x^{\alpha - 1} \) is decreasing with \( x \geq 0 \) and that for \( \alpha > 1 \), it is increasing with \( x \geq 0 \). Hence, (34) becomes

\[
|\mu^\alpha - \mathbb{E}[\beta]| \leq 2\alpha|z_0^{\alpha - 1}| \cdot \sqrt{\frac{1}{2n} \ln \frac{4}{\delta}}
\]

\[
\leq \begin{cases} 
2\alpha m^{\alpha - 1} \sqrt{\frac{1}{2n} \ln \frac{4}{\delta}} & \text{for } 0 < \alpha < 1, \\
2\alpha M^{\alpha - 1} \sqrt{\frac{1}{2n} \ln \frac{4}{\delta}} & \text{for } \alpha > 1.
\end{cases} \tag{35}
\]

By (35) and \( \left| \frac{\mathbb{E}[\beta^\alpha]}{(\mathbb{E}[\beta])^\alpha} \right| \leq \frac{(c + a)^\alpha}{m^\alpha M^\alpha} \),

\[
\left| \frac{\mathbb{E}[\beta^\alpha]}{(\mathbb{E}[\beta])^\alpha} \right| \cdot |\mu^\alpha - \mathbb{E}[\beta]| \leq \begin{cases} 
\frac{(c + a)^\alpha}{m^\alpha} 2\alpha m^{\alpha - 1} \sqrt{\frac{1}{2n} \ln \frac{4}{\delta}} & \text{for } 0 < \alpha < 1, \\
\frac{(c + a)^\alpha}{M^\alpha} 2\alpha M^{\alpha - 1} \sqrt{\frac{1}{2n} \ln \frac{4}{\delta}} & \text{for } \alpha > 1.
\end{cases} \tag{36}
\]
Consider \( \left| E[\beta^\alpha] - \frac{1}{n} \sum_i b_i^\alpha \right| \). We will apply McDiarmid’s inequality to it. Let \( \phi(x_1, \ldots, x_n) = \frac{1}{n} \sum_{i=1}^n b(x_i)^\alpha \). Then \( \phi(x_1, \ldots, x_n) \) satisfy (19) with \( q_i = \frac{\Delta_\alpha}{n} \) for all \( i \). where \( \Delta_\alpha = (c+a)\alpha - (c-a)\alpha \). Using McDiarmid’s Inequality, with probability at least 1 \(-\delta\), we have

\[
\left| E[\beta^\alpha] - \frac{1}{n} \sum_i b_i^\alpha \right| \leq \Delta_\alpha \sqrt{\frac{1}{2n} \ln \frac{4}{\delta}}.
\]  

(37)

Using (37), we have \( \frac{1}{\mu^\alpha} \left| E[\beta^\alpha] - \frac{1}{n} \sum_i b_i^\alpha \right| \) as follows;

\[
\frac{1}{\mu^\alpha} \cdot \left| E[\beta^\alpha] - \frac{1}{n} \sum_i b_i^\alpha \right| \leq \frac{\Delta_\alpha}{\mu^\alpha} \cdot \sqrt{\frac{1}{2n} \ln \frac{4}{\delta}} \\
\leq \frac{(c+a)^\alpha - (c-a)^\alpha}{(c-a)^\alpha} \cdot \sqrt{\frac{1}{2n} \ln \frac{4}{\delta}} \\
= \left[ \frac{(c+a)^\alpha}{c-a} - 1 \right] \sqrt{\frac{1}{2n} \ln \frac{4}{\delta}}.
\]

(38)

By (36), (38) and union bounds, with probability at least 1 \(-\delta\), (32) becomes

\[
\left| I_\alpha(h, P) - I_\alpha(b; n) \right| \leq \frac{1}{\alpha(\alpha-1)} \left[ 2\alpha a \left( \frac{c+a}{c-a} \right)^\alpha + 1 + \left( \frac{c+a}{c-a} \right)^\alpha - 1 \right] \sqrt{\frac{1}{2n} \ln \frac{4}{\delta}}.
\]

Note that by Mean Value Theorem, there exists \( z_0 \in (1, \frac{c+a}{c-a}) \) such that

\[
\left( \frac{c+a}{c-a} \right)^\alpha - 1 = \alpha \left( \frac{c+a}{c-a} - 1 \right) z_0^{\alpha-1}.
\]

Similarly as we do in (38), we have

\[
\left( \frac{c+a}{c-a} \right)^\alpha - 1 = \alpha \left( \frac{c+a}{c-a} - 1 \right) z_0^{\alpha-1} \\
\leq \begin{cases} 
\frac{2\alpha a}{c-a} \left( \frac{c+a}{c-a} \right)^\alpha + 1 & \text{if } 0 < \alpha < 1, \\
\frac{2\alpha a}{c-a} \left( \frac{c+a}{c-a} \right)^{\alpha-1} + 1 & \text{if } \alpha > 1.
\end{cases}
\]

Hence

\[
\left| I_\alpha(h, P) - I_\alpha(b; n) \right| \leq \begin{cases} 
\frac{2\alpha a}{|\alpha-1|(c-a)} \left[ \left( \frac{c+a}{c-a} \right)^\alpha + 1 \right] \sqrt{\frac{1}{2n} \ln \frac{4}{\delta}} & \text{if } 0 < \alpha < 1, \\
\frac{4\alpha a}{|\alpha-1|(c-a)^2} \left[ \frac{c+a}{c-a} \right]^{\alpha-1} + 1 \sqrt{\frac{1}{2n} \ln \frac{4}{\delta}} & \text{if } \alpha > 1.
\end{cases}
\]

By letting \( r = \frac{c}{a} \), we have

\[
\left| I_\alpha(h, P) - I_\alpha(b; n) \right| \leq \psi_\alpha(a, r) \sqrt{\frac{1}{2n} \ln \frac{4}{\delta}}
\]

where

\[
\psi_\alpha(a, r) = \begin{cases} 
\frac{2}{|\alpha-1|(r-1)} \left[ \left( 1 + \frac{2}{r-1} \right)^\alpha + 1 \right] & \text{if } 0 < \alpha < 1, \\
\frac{4}{|\alpha-1|(r-1)^2} \left( 1 + \frac{2}{r-1} \right)^\alpha & \text{if } \alpha > 1.
\end{cases}
\]

42
E Proof of Theorem 4.6

Lemma E.1. (See Problem 23 on page 101 of [23]) Suppose that \( f(x) \) is a convex function over \((a_0, b_0)\) and \( a_0 < s < t < u < b_0 \). Then, it holds that

\[
\frac{f(t) - f(s)}{t - s} \leq \frac{f(u) - f(s)}{u - s} \leq \frac{f(u) - f(t)}{u - t}.
\]

We denote by \( n_{FP} \) the number of false positive labels and by \( n_{FN} \) the number of false negative labels and let

\[
q_{FP} = \frac{n_{FP}}{n}, \quad q_{FN} = \frac{n_{FN}}{n}.
\]

The empirical average \( \mu \) becomes as below;

\[
\mu = \frac{1}{n} \sum_{i=1}^{n} b_i = c + a \frac{n_{FP} - n_{FN}}{n} = c + a(q_{FP} - q_{FN}).
\]

Let

\[
m_C = \Pr_{x \sim P_x} \left[ x : h(x) = y \right], \\
m_{FP} = \Pr_{x \sim P_x} \left[ x : h(x) = 1, y = 0 \right], \\
m_{FN} = \Pr_{x \sim P_x} \left[ x : h(x) = 0, y = 1 \right],
\]

that is, \( m_C \) is the probability (measure) of the set of individuals \( x \) with correct labels, \( m_{FP} \) is the probability (measure) of the set of \( x \) with false positive labels and \( m_{FN} \) the the probability (measure) of the set of \( x \) with false negative labels. Using these quantities, we express \( \mathbb{E}[b(x)] \) as below;

\[
\mathbb{E}[b(x)] = c + a(m_{FP} - m_{FN}).
\]

Let \( \varepsilon_2 = \sqrt{\frac{8d_H \ln \left( \frac{2d_H}{\pi n} \right)}{n} + 8 \ln \frac{8}{\delta}} \) and

\[
S_p = q_{FP} + q_{FN}, \quad S_m = q_{FP} - q_{FN}, \\
T_p = m_{FP} + m_{FN}, \quad T_m = m_{FP} - m_{FN}.
\]

Lemma E.2. With probability at least \( 1 - \delta \), the following two inequalities “simultaneously” hold

\[
|T_m - S_m| \leq \varepsilon_2, \quad |T_p - S_p| \leq \varepsilon_2.
\]

Proof. From the definitions of \( S_p, S_m, T_p, \) and \( T_m \), the followings are hold;

\[
q_{FP} = \frac{S_p + S_m}{2}, \quad q_{FN} = \frac{S_p - S_m}{2}, \\
m_{FP} = \frac{T_p + T_m}{2}, \quad m_{FN} = \frac{T_p - T_m}{2}.
\]
Let $\varepsilon_1 = \sqrt{\frac{2}{n} \ln \frac{4}{\delta}}$ and $\varepsilon_2 = \sqrt{\frac{8d_H \ln \left( \frac{2en}{n} + 8 \ln \frac{4}{\delta} \right)}{n}}$. From Lemma (E.4), with probability at least $1 - \frac{\delta}{2}$, it holds that $|E[b(x)] - \mu| \leq a\varepsilon_1$. By Theorem 2.2, with probability at least $1 - \frac{\delta}{2}$, we have $|R_X(h) - R_S(h)| \leq \varepsilon_2$. By using the union bound for the above two inequalities and the fact that $\varepsilon_1 < \varepsilon_2$, the two inequalities “simultaneously” hold

\begin{align}
|E[b(x)] - \mu| & \leq a\varepsilon_2, \quad (42) \\
|R_X(h) - R_S(h)| & \leq \varepsilon_2. \quad (43)
\end{align}

Lemma (E.2) follows from the observations;

\begin{align*}
m_{FP} - m_{FN} - (q_{FP} - q_{FN}) &= T_m - S_m, \\
m_{FP} + m_{FN} - (q_{FP} + q_{FN}) &= T_p - S_p.
\end{align*}

and

\begin{align}
|E[b(x)] - \mu| &= a|m_{FP} - m_{FN} - (q_{FP} - q_{FN})| = a|T_m - S_m|, \quad (44) \\
|R_X(h) - R_S(h)| &= |m_{FP} + m_{FN} - (q_{FP} + q_{FN})| = |T_p - S_p|. \quad (45)
\end{align}

We have proved Lemma (E.2) \hfill \square

**E.1 When $\alpha = 0$**

Consider $I_0(h, P) - I_0(b; n)$. From (20), we have

$$|I_0(h, P) - I_0(b; n)| \leq \left| \ln E[\beta] - \ln \mu \right| + \left| E[\ln \beta] - \frac{1}{n} \sum_i \ln b_i \right|.$$  

We find each upper bound, which is expressed in terms of $|T_m - S_m|$ and $|T_p - S_p|$, of $\left| \ln E[\beta] - \ln \mu \right|$ and $\left| E[\ln \beta] - \frac{1}{n} \sum_i \ln b_i \right|$.

- **Upper bound of $\left| \ln E[\beta] - \ln \mu \right|$**:

  Let $\varepsilon_0 = |E[\beta] - \mu|$. By (21), it holds that

  $$\left| \ln E[\beta] - \ln \mu \right| < \frac{|E[\beta] - \mu|}{m} = \frac{\varepsilon_0}{m}. \quad (46)$$

  where $m = \min(E[\beta], \mu)$ and $M = \max(E[\beta], \mu)$. If $m = \mu$, then $m = \mu = c + a(q_{FP} - q_{FN}) > c - aR_S(h)$. If $m = E[\beta]$, then $m = \mu - \varepsilon_0$; summarizing, it holds that

  $$m = \min(E[\beta], \mu) \geq \begin{cases} 
  c - aR_S(h) & \text{if } \mu \leq E[\beta] \\
  c - aR_S(h) - \varepsilon_0 & \text{otherwise.}
  \end{cases} \quad (47)$$

  In either case, we have $m > c - aR_S(h) - \varepsilon_0$. By (42), (46) becomes

  $$\left| \ln E[\beta] - \ln \mu \right| < \frac{\varepsilon_0}{c - aR_S(h) - \varepsilon_0} = \frac{a|T_m - S_m|}{c - aR_S(h) - a|T_m - S_m|}. \quad (48)$$
• Upper bound of $|E[\ln \beta] - \frac{1}{n} \sum_i \ln b_i|$: Using (40) and (41), we can write $\left|\frac{1}{n} \sum_i \ln b_i - E[\ln \beta]\right|$ as follows;

$$\left|\frac{1}{n} \sum_i \ln b_i - E[\ln \beta]\right|$$

$$= \left|\left( m_{FP} + m_{FN} - \left(q_{FP} + q_{FN}\right)\right) \ln c + \left(q_{FP} - m_{FP}\right) \ln(c + a) + \left(q_{FN} - m_{FN}\right) \ln(c - a)\right|$$

$$= |(T_p - S_p) \ln c + \left(\frac{S_p + S_m - T_p + T_m}{2}\right) \ln(c + a) + \left(\frac{S_p - S_m - T_p + T_m}{2}\right) \ln(c - a)|$$

$$\leq \left|\frac{\ln(c + a) + \ln(c - a) - 2 \ln c}{2}\right| \cdot \left(\frac{S_p - T_p}{}\right) + \left(\frac{\ln(c + a) - \ln(c - a)}{2}\right) \cdot |S_m - T_m|.$$ 

Since $\ln(c + a) + \ln(c - a) - 2 \ln c = \ln(\frac{c^2 - a^2}{c^2}) < 0$, we have $\frac{\ln(c + a) + \ln(c - a) - 2 \ln c}{2} = \frac{2 \ln c - \ln(c + a) - \ln(c - a)}{2}$.

Therefore, if we let $c_p = \frac{2 \ln c - \ln(c + a) - \ln(c - a)}{2}$ and $c_m = \frac{\ln(c + a) - \ln(c - a)}{2}$, it holds that

$$\left|E[\ln \beta] - \frac{1}{n} \sum_i \ln b_i\right| \leq c_p |S_p - T_p| + c_m |S_m - T_m|.$$ (49)

Combining (48), (49), and Lemma E.2, with probability at least $1 - \delta$, we have

$$|I_0(h, P) - I_0(b; n)|$$

$$\leq \frac{a|T_m - S_m|}{c - aR_s(h)} + c_p |S_p - T_p| + c_m |S_m - T_m|$$

$$\leq \varepsilon_2 \left[\frac{a}{c - aR_s(h) - a\varepsilon_2} + \ln \left(\frac{c}{c - a}\right)\right] \quad \text{(by Lemma E.2)}$$

$$= \varepsilon_2 \left[\frac{1}{r - R_s(h) - \varepsilon_2} + \ln \left(\frac{1 + \frac{1}{r - 1}}{r - 1}\right)\right] \quad (\therefore r = \frac{c}{a}).$$

Note that there is no need to take union bound when combining (48) and (49), because the two inequalities in Lemma E.2 simultaneously hold.

**E.2 When $\alpha = 1$**

Recall that $\omega = E[b(x)]$. From (26) and (27), $I_1(h, P) - I_1(b; n)$ is written as

$$I_1(h, P) - I_1(b; n) = \ln \mu - \ln \omega + c \ln c \left(\frac{1}{\omega} - \frac{1}{\mu} + F_1\right) + (c + a) \ln(c + a)F_2 + (c - a) \ln(c - a)F_3.$$
where
\[ F_1 = \frac{q_{FP} + q_{FN}}{\mu} - \frac{m_{FP} + m_{FN}}{\omega}, \]
\[ F_2 = \frac{m_{FP} - q_{FP}}{\omega} - \frac{q_{FN}}{\mu}, \]
\[ F_3 = \frac{m_{FN} - q_{FN}}{\omega} - \frac{q_{FN}}{\mu}. \]

After rewriting \( F_1, F_2, \) and \( F_3 \) as follows,
\[ F_1 = -\left(1 - \frac{\omega}{\mu}\right)(q_{FP} + q_{FN}) + \frac{1}{\omega} \left[q_{FP} + q_{FN} - (m_{FP} + m_{FN})\right], \]
\[ F_2 = \left(1 - \frac{\omega}{\mu}\right)q_{FP} + \frac{m_{FP} - q_{FP}}{\omega}, \]
\[ F_3 = \left(1 - \frac{\omega}{\mu}\right)q_{FN} + \frac{m_{FN} - q_{FN}}{\omega}, \]
we have
\[ I_1(h, P) - I_1(b; n) = \ln \mu - \ln \omega + \frac{1}{\omega} \left[H_1 + \frac{1}{\omega} H_2\right], \]
where
\[ H_1 = c \ln c(1 - (q_{FP} + q_{FN})) + (c + a) \ln(c + a)q_{FP} + (c - a) \ln(c - a)q_{FN}, \]
\[ H_2 = c \ln c(q_{FP} + q_{FN} - (m_{FP} + m_{FN})) + (c + a) \ln(c + a)(m_{FP} - q_{FP}) + (c - a) \ln(c - a)(m_{FN} - q_{FN}). \]

Therefore, we have
\[ |I_1(h, P) - I_1(b; n)| \leq |\ln \mu - \ln \omega| + \frac{1}{\omega} \left|H_1\right| + \frac{1}{\omega} \left|H_2\right|. \]

We find each upper bound of \(|\ln \mu - \ln \omega|, \left|\frac{1}{\omega} - \frac{1}{\mu}\right|, |H_1| \) and \(|H_2|\).

- An upper bound of \(|\ln \mu - \ln \omega|
  \]
  By (48), we have
  \[ |\ln \mu - \ln \omega| \leq \frac{a|T_m - S_m|}{c - aR_S(h) - a|T_m - S - M|}. \]

- An upper bound of \(\left|\frac{1}{\mu} - \frac{1}{\omega}\right|
  \]
  By Lemma E.2 with probability at least \(1 - \delta\), it holds that
  \[ \left|\frac{1}{\mu} - \frac{1}{\omega}\right| = \left|\frac{\mu - \omega}{\mu \omega}\right| \leq \frac{a|T_m - S_m|}{mM}. \]  \hspace{1cm} (50)
  where \(m = \min(\mu, E[\beta])\) and \(M = \max(\mu, E[\beta])\). By (47), \(m > c - aR_S(h) - |E[\beta] - \mu| = c - aR_S(h) - a|T_m - S_m|\). Note that \(M \geq \mu \geq c - aR_S(h)\). Therefore (50) becomes
  \[ \left|\frac{1}{\mu} - \frac{1}{\omega}\right| \leq \frac{a|T_m - S_m|}{(c - aR_S(h))(c - aR_S(h) - a|T_m - S_m|)}. \]  \hspace{1cm} (51)
• An upper bound of $H_1$

Let $\gamma_p = \frac{(c+a) \ln(c+a) + (c-a) \ln(c-a)}{2}$ and $\gamma_m = \frac{(c+a) \ln(c+a) - (c-a) \ln(c-a)}{2}$. Using (40) and (41), we compute an upper bound of $|H_1|$ as below:

$$
|H_1| = |(1 - S_p) c \ln c + \gamma_p S_p + \gamma_m S_m| \\
= |c \ln c + (\gamma_p - c \ln c) S_p + \gamma_m S_m| \\
= |c \ln c + (\gamma_p - c \ln c) S_p + \gamma_m S_m| \\
\leq c \ln c + |\gamma_p - c \ln c| \cdot |S_p| + |\gamma_m| \cdot |S_m|.
$$

We will show that $\gamma_p - c \ln c \geq 0$. For $\gamma_p - c \ln c \geq 0$, it is enough to show that

$$
\frac{(c + a) \ln(c + a) - c \ln c}{a} \geq \frac{c \ln c - (c - a) \ln(c - a)}{a}.
$$

By Lemma D.3, $f(x) = x \ln x$ is a convex function over $x \geq \frac{1}{e}$. It can be checked that (53) holds by applying Lemma E.1 for $f(x) = x \ln x$. It obviously holds that $\gamma_m \geq 0$. Hence (52) becomes

$$
|H_1| \leq c \ln c + (\gamma_p - c \ln c) |S_p| + \gamma_m |S_m| \\
\leq c \ln c + (\gamma_p - c \ln c) R_S(h) + \gamma_m R_S(h) \quad (\because S_p = R_S(h), |S_m| < S_p) \\
\leq c \ln c + (\gamma_p + \gamma_m - c \ln c) R_S(h).
$$

• An upper bound of $\frac{1}{\omega}$:

Recall that $\omega = \mathbb{E}[b(x)]$ and $m = \min[\mu, \mathbb{E}[b(x)]]$.

$$
\frac{1}{\omega} < \frac{1}{m} < \frac{1}{c - a R_S(h) - a |T_m - S_m|}.
$$

• An upper bound of $|H_2|$

Using (40) and (41), we rewrite $H_2$ as below

$$
H_2 = c \ln c (S_p - T_p) + \gamma_p (T_p - S_p) + \gamma_m (T_m - S_m).
$$

Recall that $\gamma_p = \frac{(c+a) \ln(c+a) + (c-a) \ln(c-a)}{2}$ and $\gamma_m = \frac{(c+a) \ln(c+a) - (c-a) \ln(c-a)}{2}$. Similarly as we have done in finding an upper bound of $|H_1|$, with probability at least $1 - \delta$, we have

$$
|H_2| = \left| c \ln c (S_p - T_p) + \gamma_p (T_p - S_p) + \gamma_m (T_m - S_m) \right| \\
= \left| (\gamma_p - c \ln c) (T_p - S_p) + \gamma_m (S_m - T_m) \right| \\
\leq (\gamma_p - c \ln c) |T_p - S_p| + \gamma_m |S_m - T_m| \quad (\because \gamma_p - c \ln c \geq 0, \gamma_m \geq 0).
$$

By (48), (51), (54), (55), and (53), an upper bound of $|I_1(h, P) - I_1(b; n)|$ is

$$
|I_1(h, P) - I_1(b; n)| \leq S_1 + S_2 + S_3.
$$
where
\[
S_1 = \frac{a|T_m - S_m|}{c - aR_S(h) - a|T_m - S_m|},
\]
\[
S_2 = \frac{a|T_m - S_m|(c \ln c + (\gamma_p + \gamma_m - c \ln c)R_S(h))}{(c - aR_S(h))(c - aR_S(h) - a|T_m - S_m|)},
\]
\[
S_3 = \frac{(\gamma_p - c \ln c)|T_p - S_p| + \gamma_m |T_m - S_m|}{c - aR_S(h) - a|T_m - S_m|}.
\]

Applying Lemma E.2 to \( S_1 + S_2 + S_3 \), we have
\[
S_1 + S_2 + S_3 \leq \frac{a \varepsilon_2}{c - aR_S(h) - a \varepsilon_2} \left[ 1 + \frac{\ln c + (\gamma_p + \gamma_m - c \ln c)R_S(h)}{c - aR_S(h)} + \frac{\gamma_p + \gamma_m - c \ln c}{a} \right]
\]
\[
\leq \frac{a \varepsilon_2}{c - aR_S(h) - a \varepsilon_2} \left[ 1 + \frac{\ln c + (\gamma_p + \gamma_m - c \ln c)}{c - aR_S(h)} + \frac{1}{a} \right] (\gamma_p + \gamma_m - c \ln c)
\]
\[
= \frac{\varepsilon_2}{r - R_S(h) - \varepsilon_2} \left[ 1 + \frac{r \ln(\alpha)}{r - R_S(h)} + \left( \frac{R_S(h)}{r - R_S(h)} + 1 \right) \left( r \ln \left( 1 + \frac{1}{r} \right) + \ln (\alpha(r + 1)) \right) \right]
\]
\[
\text{(since } r = \frac{\varepsilon}{a} \text{ and } \gamma_p + \gamma_m - c \ln c = c \ln \left( \frac{\varepsilon + a}{c} \right) + a \ln(c + a)).
\]

Note that
\[
1 + \frac{r \ln(\alpha)}{r - R_S(h)} + \left( \frac{R_S(h)}{r - R_S(h)} + 1 \right) \left( r \ln \left( 1 + \frac{1}{r} \right) + \ln (\alpha(r + 1)) \right)
\]
\[
= 1 + \frac{r(2 \ln a + \ln r + \ln(r + 1) + \ln \left( 1 + \frac{1}{r} \right))}{r - R_S(h)}
\]
\[
= 1 + \frac{r(1 + 2 \ln a + 2 \ln(r + 1))}{r - R_S(h)} \quad \text{(since } \ln(1 + x) \leq x). \quad (56)
\]

Using (56), we have
\[
S_1 + S_2 + S_3 \leq \frac{\varepsilon_2}{r - R_S(h) - \varepsilon_2} \left[ 1 + \frac{r}{r - R_S(h)} \left( 1 + 2 \ln a + 2 \ln(r + 1) \right) \right].
\]

Recalling \( \varepsilon_2 = \sqrt{\frac{8d_m \ln \left( \frac{2\varepsilon \alpha + 8 \ln \frac{2}{\alpha}}{n} \right)}{a}} \), we have proved the case of \( \alpha = 1 \) of Theorem 4.6.

E.3  When 0 < \( \alpha < 1 \)

After simple algebra, we have
\[
I_{\alpha}(h, P) - I_{\alpha}(b; n) = \frac{1}{\alpha(\alpha - 1)} \left( \frac{J_n}{J^\alpha} - \frac{K_n}{K^\alpha} \right).
\]

where \( \theta = \frac{a}{c} = \frac{1}{\alpha} \) and
\[
J = 1 + \theta(m_{FP} - m_{FN}), \quad J_n = 1 - (m_{FP} + m_{FN}) + (1 + \theta)^{\alpha} m_{FP} + (1 - \theta)^{\alpha} m_{FN},
\]
\[
K = 1 + \theta(q_{FP} - q_{FN}), \quad K_n = 1 - (q_{FP} + q_{FN}) + (1 + \theta)^{\alpha} q_{FP} + (1 - \theta)^{\alpha} q_{FN}.
\]

48
Note that all of $J$, $J_n$, $K$, and $K_n$ are positive since $\theta = \frac{a}{c}$ with $c > a > 0$ and $c - a \geq 1$, $0 \leq m_{FP} + m_{FN} \leq 1$, and $0 \leq q_{FP} + q_{FN} \leq 1$. Moreover,

\[
\frac{J_n}{J^\alpha} - \frac{K_n}{K^\alpha} = T_1 + T_2 + T_3 + T_4
\]

where

\[
T_1 = \frac{1}{J^\alpha} - \frac{1}{K^\alpha},
\]
\[
T_2 = \frac{q_{FP} + q_{FN} - (m_{FP} + m_{FN})}{K^\alpha} - \frac{1}{J^\alpha} \frac{m_{FP} + m_{FN}}{K^\alpha},
\]
\[
T_3 = (1 + \theta)^\alpha \left[ \frac{m_{FP}}{J^\alpha} - \frac{q_{FP}}{K^\alpha} \right],
\]
\[
T_4 = (1 - \theta)^\alpha \left[ \frac{m_{FN}}{J^\alpha} - \frac{q_{FN}}{K^\alpha} \right].
\]

After rewriting $T_2$, $T_3$, and $T_4$ as follows,

\[
T_2 = \frac{q_{FP} + q_{FN} - (m_{FP} + m_{FN})}{K^\alpha} - \frac{1}{J^\alpha} \left( \frac{m_{FP} + m_{FN}}{K^\alpha} \right),
\]
\[
T_3 = (1 + \theta)^\alpha \left( \frac{1}{J^\alpha} - \frac{1}{K^\alpha} \right) m_{FP} + (1 + \theta)^\alpha \frac{m_{FP} - q_{FP}}{K^\alpha},
\]
\[
T_4 = (1 - \theta)^\alpha \left( \frac{1}{J^\alpha} - \frac{1}{K^\alpha} \right) m_{FN} + (1 - \theta)^\alpha \frac{m_{FN} - q_{FN}}{K^\alpha},
\]

we have

\[
\frac{J_n}{J^\alpha} - \frac{K_n}{K^\alpha} = \left( \frac{1}{J^\alpha} - \frac{1}{K^\alpha} \right) (Z_1 + Z_2) + \frac{1}{K^\alpha} (Z_3 + Z_4)
\]

where

\[
Z_1 = 1 - (m_{FP} + m_{FN}),
\]
\[
Z_2 = (1 + \theta)^\alpha m_{FP} + (1 - \theta)^\alpha m_{FN},
\]
\[
Z_3 = q_{FP} + q_{FN} - (m_{FP} + m_{FN}),
\]
\[
Z_4 = (1 + \theta)^\alpha (m_{FP} - q_{FP}) + (1 - \theta)^\alpha (m_{FN} - q_{FN}).
\]

Therefore, it holds that

\[
|I_\alpha(h, P) - I_\alpha(b; n)| \leq \left| \frac{1}{J^\alpha} - \frac{1}{K^\alpha} \right| \cdot \frac{|Z_1| + |Z_2|}{|\alpha(\alpha - 1)|} + \frac{1}{|K^\alpha|} \cdot \frac{|Z_3 + Z_4|}{\alpha(\alpha - 1)}. \tag{57}
\]

We will find each upper bound of $\left| \frac{1}{J^\alpha} - \frac{1}{K^\alpha} \right|$, $|Z_1|$, $|Z_2|$, $\frac{1}{|K^\alpha|}$, and $\frac{|Z_3 + Z_4|}{\alpha(\alpha - 1)}$. Recall that $K = 1 + \theta(q_{FP} - q_{FN})$. Obviously $1 - \theta R_S(h) \leq K \leq 1 + \theta R_S(h)$. Hence

\[
\frac{1}{|K^\alpha|} \leq \frac{1}{(1 - \theta R_S(h))^{\alpha}}.
\]
• An upper bound of \(\frac{1}{J^{\alpha}} - \frac{1}{K^{\alpha}}\)

We compute \(\frac{1}{J^{\alpha}} - \frac{1}{K^{\alpha}}\):

\[
\left|\frac{1}{J^{\alpha}} - \frac{1}{K^{\alpha}}\right| = \frac{|K^\alpha - J^\alpha|}{J^\alpha K^\alpha}.
\]

By Mean Value Theorem, there exists \(x_0 \in (m_0, M_0)\) such that \(K^\alpha - J^\alpha = (K - J) \cdot \alpha \bar{x}_0^{\alpha - 1}\) where \(m_0 = \min(K, J)\), and \(M_0 = \max(K, J)\). Since \(0 < \alpha < 1\), the function \(f(x) = x^{\alpha - 1}\) is a decreasing function of \(x\). Hence \(\bar{x}_0^{\alpha - 1} < m_0^{\alpha - 1}\). Recall that \(J = 1 + \theta(m_{FP} - m_{FN})\) and \(K = 1 + \theta(q_{FP} - q_{FN})\). Therefore \(|K - J| = \theta|m_{FP} - m_{FN} - (q_{FP} - q_{FN})| = \theta|T_m - S_m|\).

Summarizing these, we have:

\[
\left|\frac{1}{J^{\alpha}} - \frac{1}{K^{\alpha}}\right| = \frac{|K - J| \cdot \alpha \bar{x}_0^{\alpha - 1}}{J^\alpha K^\alpha} \leq \frac{\theta|T_m - S_m| \cdot \alpha m_0^{\alpha - 1}}{m_0^\alpha M_0^\alpha} = \frac{\alpha \theta|T_m - S_m|}{m_0^\alpha M_0^\alpha}.
\]

Consider the case that \(J \geq K\). Note that \(K = 1 + \theta(q_{FP} - q_{FN}) > 1 - \theta R_S(h)\) and \(J = K + |K - J| = K + \theta|T_m - S_m| > 1 + \theta(|T_m - S_m| - R_S(h))\). Hence \(m_0 M_0^\alpha = K J^\alpha > (1 - \theta R_S(h))(1 - \theta R_S(h) + \theta|T_m - S_m|)^\alpha\).

Consider the case that \(J < K\). In this case, we have \(J = K - |K - J| = K - \theta|T_m - S_m| > 1 - \theta R_S(h) - \theta|T_m - S_m|\) and \(m_0 M_0^\alpha = J K^\alpha > (1 - \theta R_S(h) - \theta|T_m - S_m|)(1 - \theta R_S(h))^\alpha\).

In either case, we have \(m_0 M_0^\alpha > (1 - \theta R_S(h) - \theta|T_m - S_m|)(1 - \theta R_S(h))^\alpha\). Therefore, \(|\frac{1}{J^{\alpha}} - \frac{1}{K^{\alpha}}|\) becomes:

\[
\left|\frac{1}{J^{\alpha}} - \frac{1}{K^{\alpha}}\right| < \frac{\alpha \theta|T_m - S_m|}{(1 - \theta R_S(h) - \theta|T_m - S_m|)(1 - \theta R_S(h))^\alpha}.
\]

\[
\text{(58)}
\]

• An upper bound of \(|Z_1|\)

Recall that \(|Z_1| = \left|1 - (m_{FP} + m_{FN})\right|\).

\[
|Z_1| = |1 - (m_{FP} + m_{FN})| \\
\leq |1 - (q_{FP} + q_{FN})| + |q_{FP} + q_{FN} - (m_{FP} + m_{FN})| \\
= 1 - R_S(h) + |T_p - S_p|
\]

\[
\text{(59)}
\]

• An upper bound of \(|Z_2|\)

Recall that \(Z_2 = (1 + \theta)\alpha m_{FP} + (1 - \theta)\alpha m_{FN}\). Using (41), we rewrite \(Z_2\) as follows:

\[
Z_2 = \frac{(1 + \theta)\alpha + (1 - \theta)\alpha}{2} T_p + \frac{(1 + \theta)\alpha - (1 - \theta)\alpha}{2} T_m.
\]

Hence

\[
|Z_2| \leq \frac{(1 + \theta)\alpha + (1 - \theta)\alpha}{2} \cdot |T_p| + \frac{(1 + \theta)\alpha - (1 - \theta)\alpha}{2} \cdot |T_m| \\
\leq \frac{(1 + \theta)\alpha + (1 - \theta)\alpha}{2} \cdot R_X(h) + \frac{(1 + \theta)\alpha - (1 - \theta)\alpha}{2} \cdot R_X(h) \\
= (1 + \theta)\alpha R_X(h). \quad (\therefore T_p = R_X(h), \ |T_m| \leq R_X(h)).
\]

(60)
An upper bound of $\left| \frac{Z_3 + Z_4}{\alpha(\alpha - 1)} \right|$

Recall that

\[
Z_3 = q_{FP} + q_{FN} - (m_{FP} + m_{FN}),
\]

\[
Z_4 = (1 + \theta)^{\alpha}(m_{FP} - q_{FP}) + (1 - \theta)^{\alpha}(m_{FN} - q_{FN}).
\]

Using (41) and (40), we rewrite $Z_3$ and $Z_4$ as below:

\[
Z_3 = S_p - T_p,
\]

\[
Z_4 = \frac{(1 + \theta)^{\alpha} + (1 - \theta)^{\alpha}}{2}(T_p - S_p) + \frac{(1 + \theta)^{\alpha} - (1 - \theta)^{\alpha}}{2}(T_m - S_m)
\]

We find an upper bound of $\left| \frac{Z_3 + Z_4}{\alpha(\alpha - 1)} \right|$:

\[
\left| \frac{Z_3 + Z_4}{\alpha(\alpha - 1)} \right| = \left| \frac{(1 + \theta)^{\alpha} + (1 - \theta)^{\alpha} - 2}{2\alpha(\alpha - 1)}(T_p - S_p) + \frac{(1 + \theta)^{\alpha} - (1 - \theta)^{\alpha}}{2\alpha(\alpha - 1)}(T_m - S_m) \right|
\]

Note that $f(x) = \frac{x^{\alpha}}{\alpha(\alpha - 1)}$ is a convex function for $\alpha \neq 0, 1$. If we apply Lemma E.1 for $f(x) = \frac{x^{\alpha}}{\alpha(\alpha - 1)}$ with the values of $s = 1 - \theta$, $t = 1$, and $u = 1 + \theta$, it holds that

\[
\frac{1 - (1 - \theta)^{\alpha}}{\alpha(\alpha - 1)\theta} \leq \frac{(1 + \theta)^{\alpha} - (1 - \theta)^{\alpha}}{\alpha(\alpha - 1)\theta} \leq \frac{(1 + \theta)^{\alpha} - 1}{\alpha(\alpha - 1)\theta}.
\]  

(61)

By (61) and $\theta = \frac{1}{r} > 0$, it holds that

\[
\frac{(1 + \theta)^{\alpha} + (1 - \theta)^{\alpha} - 2}{\alpha(\alpha - 1)} \geq 0.
\]

Using the above property, we have

\[
\left| \frac{Z_3 + Z_4}{\alpha(\alpha - 1)} \right| \leq \left| \frac{(1 + \theta)^{\alpha} + (1 - \theta)^{\alpha} - 2}{2\alpha(\alpha - 1)} \right| \cdot |T_p - S_p| + \left| \frac{(1 + \theta)^{\alpha} - (1 - \theta)^{\alpha}}{2\alpha(\alpha - 1)} \right| \cdot |T_m - S_m|
\]

\[
\leq \frac{2 - (1 + \theta)^{\alpha} - (1 - \theta)^{\alpha}}{2\alpha(1 - \alpha)} \cdot |T_p - S_p| + \frac{(1 + \theta)^{\alpha} - (1 - \theta)^{\alpha}}{2\alpha(1 - \alpha)} \cdot |T_m - S_m|
\]

(62)

The second inequality holds since $0 < \alpha < 1$ (hence $\alpha - 1 < 0$).

By (37), (58), (59), (60), (62), and applying Lemma E.2 with probability at least $1 - \delta$, we have

\[
|I_\alpha(h, P) - I_\alpha(b : n)| \leq \frac{\alpha\theta\varepsilon_2}{(1 - \theta R_S(h) - \theta\varepsilon_2)(1 - \theta R_S(h))^\alpha} \cdot \frac{1 - R_S(h) + \varepsilon_2 + (1 + \theta)^\alpha R_X(h)}{\alpha(1 - \alpha)} + \frac{1}{(1 - \theta R_S(h))^\alpha} \cdot \left[ \frac{2 - (1 + \theta)^\alpha - (1 - \theta)^\alpha}{2\alpha(1 - \alpha)} \cdot \varepsilon_2 + \frac{(1 + \theta)^\alpha - (1 - \theta)^\alpha}{2\alpha(1 - \alpha)} \cdot \varepsilon_2 \right]
\]

\[
\leq \frac{\varepsilon_2}{\alpha(1 - \alpha)(1 - \theta R_S(h))^\alpha} \cdot \left[ \frac{1}{1 - \theta R_S(h) - \theta\varepsilon_2} \cdot \left( 1 + (1 + \theta)^\alpha - 1 \right) R_S(h) + \varepsilon_2 \left( (1 + \theta)^\alpha + 1 \right) \right] + 1 - (1 - \theta)^\alpha
\]

\[
= \frac{\varepsilon_2}{\alpha(1 - \alpha)} \left( \frac{r}{r - R_S(h)} \right)^\alpha \left[ 1 - \left( 1 - \frac{1}{r} \right)^\alpha + \frac{\alpha C_\psi}{r - R_S(h) - \varepsilon_2} \right] \]

51
where
\[ C_\psi = 1 + ((1 + \frac{1}{r})^\alpha - 1)RS(h) + \varepsilon_2 ((1 + \frac{1}{r})^\alpha + 1). \]
The second inequality holds since \( R_X(h) \leq R_S(h) + \varepsilon_2 \) by Theorem 2.2 and the first equality is obtained by \( \theta = \frac{1}{r} \).

By Mean Value Theorem and \( f(x) = x^{\alpha-1} \) is decreasing for \( 0 < \alpha < 1 \) and \( x > 0 \), we have \( z_0 \in (1 - \frac{1}{r}, 1) \) such that
\[ 1 - \left(1 - \frac{1}{r}\right)^\alpha = \frac{\alpha Z_\theta}{r} \leq \alpha. \]
Using the above inequality, we have found an upper bound of \( |I_\alpha(h, P) - I_\alpha(b : n)| \);
\[ |I_\alpha(h, P) - I_\alpha(b : n)| \leq \frac{\varepsilon_2}{1 - \alpha} \left(\frac{r}{r - R_S(h)}\right)^\alpha \left(1 + \frac{C_\psi}{r - R_S(h) - \varepsilon_2}\right) \]
\[ = \frac{1}{1 - \alpha} \left(\frac{r}{r - R_S(h)}\right)^\alpha \left(1 + \frac{C_\psi}{r - R_S(h) - \varepsilon_2}\right) \sqrt{8d_\mathcal{H} \ln \left(\frac{2e^\delta r + 8 \ln \frac{\delta}{2}}{n}\right)}. \]

### E.4 When \( \alpha \geq 1 \)

The analysis is almost same as in the case of \( 0 < \alpha < 1 \) but needs some modification because the value of \( \alpha \) is bigger than one. The upper bounds of \( \left|\frac{1}{K^\alpha} - \frac{1}{J^\alpha}\right| \) and \( \frac{Z_3 + Z_4}{\alpha(\alpha - 1)} \) have different values when \( \alpha > 1 \). For the case \( \alpha \geq 1 \), we can show that (60) and (62) are replaced by the following inequalities, respectively,
\[ \left|\frac{1}{K^\alpha} - \frac{1}{J^\alpha}\right| \leq \frac{\alpha \theta |T_m - S_m|}{(1 - \theta R_S(h)) \cdot (1 - \theta R_S(h) - \theta |T_m - S_m|)^\alpha} \]
and
\[ \left|\frac{Z_3 + Z_4}{\alpha(\alpha - 1)}\right| \leq \frac{(1 + \theta)^\alpha + (1 - \theta)^\alpha - 2}{2 \alpha(\alpha - 1)} \cdot |T_p - S_p| + \frac{(1 + \theta)^\alpha - (1 - \theta)^\alpha}{2 \alpha(\alpha - 1)} \cdot |T_m - S_m|. \]
By (57), (63), (59), (60), and (64), we have
\[ |I_\alpha(h, P) - I_\alpha(b : n)| \]
\[ \leq \frac{\alpha \theta \varepsilon_2}{(1 - \theta R_S(h))(1 - \theta R_S(h) - \varepsilon_2)^\alpha} \cdot \frac{1 - R_S(h) + \varepsilon_2 + (1 + \theta)^\alpha R_X(h)}{\alpha(\alpha - 1)} \]
\[ + \frac{1}{(1 - \theta R_S(h))^\alpha} \cdot \frac{(1 + \theta)^\alpha + (1 - \theta)^\alpha - 2}{2 \alpha(\alpha - 1)} \varepsilon_2 \]
\[ \leq \frac{\alpha \theta \varepsilon_2}{(1 - \theta R_S(h))(1 - \theta R_S(h) - \varepsilon_2)^\alpha} \cdot \frac{1 - R_S(h) + \varepsilon_2 + (1 + \theta)^\alpha (R_S(h) + \varepsilon_2)}{\alpha(\alpha - 1)} \]
\[ + \frac{1}{(1 - \theta R_S(h))^\alpha} \cdot \frac{(1 + \theta)^\alpha + (1 - \theta)^\alpha - 2}{2 \alpha(\alpha - 1)} \varepsilon_2 \]
\[ \leq \varepsilon_2 \cdot \frac{(1 + \theta)^\alpha - 1 + \frac{\alpha \theta}{1 - \theta R_S(h)} \left[\frac{1 - \theta R_S(h)}{1 - \theta R_S(h) + \varepsilon_2}\right]^{\alpha} \left[1 + ((1 + \theta)^\alpha - 1) R_S(h) + \varepsilon_2 ((1 + \theta)^\alpha + 1)\right]}{\alpha(\alpha - 1)(1 - \theta R_S(h))^\alpha}. \]
We rewrite $|I_{\alpha}(h, P) - I_{\alpha}(b : n)|$ in terms of $r = \frac{1}{\theta}$ instead of $\theta$. Recalling that $C_\psi = 1 + ((1 + \frac{1}{\theta})^\alpha - 1) R_S(h) + \varepsilon_2((1 + \frac{1}{\theta})^\alpha + 1)$, we have

$$|I_{\alpha}(h, P) - I_{\alpha}(b : n)| \leq \frac{\varepsilon_2}{\alpha(\alpha - 1)(1 - R_S(h))} \left[ \left(1 + \frac{1}{r}\right)^\alpha - 1 + \frac{\alpha(1 + \frac{\varepsilon_2}{r - R_S(h)} - \varepsilon_2) C_\psi}{r - R_S(h)} \right].$$

By Mean Value Theorem and the fact that $f(x) = x^{\alpha - 1}$ is increasing with $x > 0$ for $\alpha > 1$, we have

$$\left(1 + \frac{1}{r}\right)^\alpha - 1 \leq \frac{\alpha}{r} \left(1 + \frac{1}{r}\right)^{\alpha - 1}.$$

Hence

$$|I_{\alpha}(h, P) - I_{\alpha}(b : n)| \leq \frac{\varepsilon_2}{\alpha(\alpha - 1)(1 - R_S(h))} \left[ \frac{\alpha}{r} \left(1 + \frac{1}{r}\right)^\alpha + \frac{\alpha(1 + \frac{\varepsilon_2}{r - R_S(h) - \varepsilon_2}) C_\psi}{r - R_S(h)} \right].$$

Recalling the definition of $\varepsilon_2$, we have proved the case of $\alpha > 1$ of Theorem 4.6.

**F Proof of Theorem 5.1**

We provided the proof for the completeness of the paper and for the objective that it indeed holds even for some modifications we made to fit for our special case, even though the proof can be basically found in [11, 12].

We need the following Theorem and its corollary in [11, 12].

**Theorem F.1.** For any $j \in \{0, 1\}$,

$$\max_{\lambda \in \Lambda} \frac{1}{T} \sum_{t=1}^{T} l(\hat{h}^{(t)}, \lambda) \leq \kappa + \frac{\ln 2}{\kappa T} + \frac{1}{T} \sum_{t=1}^{T} l(\hat{h}^{(t)}, \lambda) \sum_{j=0}^{1} \frac{w_j^{(t)} \lambda_j}{W^{(t)}}.$$ 

**Corollary F.2.** If we choose $\kappa = \sqrt{\frac{\ln 2}{T}}$, then

$$\max_{\lambda \in \Lambda} \frac{1}{T} \sum_{t=1}^{T} l(\hat{h}^{(t)}, \lambda) \leq 2\sqrt{\frac{\ln 2}{T}} + \frac{1}{T} \sum_{t=1}^{T} l(\hat{h}^{(t)}, \hat{\lambda}^{(t)}).$$

Since $\Delta H$ is a convex hull of $H$, $\min_{D \in \Delta H}$ occurs at a pure hypothesis $h \in H$. Hence we consider $\min_{h \in H}$ instead of $\min_{D \in \Delta H}$. Note that $\arg \min_{h \in H} L(h, \lambda) = \arg \min_{h \in H} l(h, \lambda)$ and
arg max_{\lambda \in \Lambda} L(D, \lambda) = \arg \max_{\lambda \in \Lambda} l(D, \lambda) \text{ since adding and multiplying a positive constant has no effect on optimization.}

Then the following holds:

\[
\begin{align*}
\max_{\lambda} \min_{h \in \mathcal{H}} l(h, \lambda) & \geq \min_{h \in \mathcal{H}} l(h, \tilde{\lambda}) \\
& = \min_{h \in \mathcal{H}} l(h, \frac{1}{T} \sum_{t=1}^{T} \hat{\lambda}^{(t)}) \\
& = \min_{h \in \mathcal{H}} \frac{R_S(h) + \left[ \frac{1}{T} \sum_{t=1}^{T} \hat{\lambda}^{(t)} \right] \cdot (I_\alpha(b_h; n) - \gamma) + B}{A_\alpha} \\
& \geq \frac{1}{T} \sum_{t=1}^{T} \min_{h \in \mathcal{H}} \frac{R_S(h) + \hat{\lambda}^{(t)} \cdot (I_\alpha(b_h; n) - \gamma) + B}{A_\alpha} \\
& = \frac{1}{T} \sum_{t=1}^{T} \min_{h \in \mathcal{H}} l(h, \hat{\lambda}^{(t)}) \\
& = \frac{1}{T} \sum_{t=1}^{T} l(\hat{\lambda}^{(t)}, \hat{\lambda}^{(t)}) \\
& \geq \max_{\lambda} \frac{1}{T} \sum_{t=1}^{T} l(\hat{\lambda}^{(t)}, \lambda) - 2\sqrt{\frac{\ln 2}{T}} \quad \text{(by Corollary F.2)} \\
& = \max_{\lambda} l\left( \frac{1}{T} \sum_{t=1}^{T} \hat{\lambda}^{(t)}, \lambda \right) - 2\sqrt{\frac{\ln 2}{T}} \\
& \geq \min_{h} \max_{\lambda} l(h, \lambda) - 2\sqrt{\frac{\ln 2}{T}}. 
\end{align*}
\]

We elaborate how the second inequality, (67), holds below. Let

\[
\hat{h}(\hat{\lambda}^{(t)}) = \min_{h \in \mathcal{H}} R_S(h) + \hat{\lambda}^{(t)} (I_\alpha(b_h; n) - \gamma),
\]

\[
\tilde{h} = \min_{h \in \mathcal{H}} R_S(h) + \left[ \frac{1}{T} \sum_{t=1}^{T} \hat{\lambda}^{(t)} \right] \cdot (I_\alpha(b_h; n) - \gamma).
\]

Note that by the definition of \( \tilde{h} \), it holds that

\[
\min_{h \in \mathcal{H}} R_S(h) + \left[ \frac{1}{T} \sum_{t=1}^{T} \hat{\lambda}^{(t)} \right] \cdot (I_\alpha(b_h; n) - \gamma) = R_S(\tilde{h}) + \left[ \frac{1}{T} \sum_{t=1}^{T} \hat{\lambda}^{(t)} \right] \cdot (I_\alpha(b_h; n) - \gamma)
\]

\[
= \frac{1}{T} \sum_{t=1}^{T} \left[ R_S(\tilde{h}) + \hat{\lambda}^{(t)} (I_\alpha(b_h; n) - \gamma) \right].
\]

54
By the definition of $\hat{h}(\hat{\lambda}(t))$,

\[ R_S(\hat{h}) + \hat{\lambda}(t)(I_\alpha(b_h; n) - \gamma) \geq R_S(\hat{h}(\hat{\lambda}(t))) + \hat{\lambda}(t)(I_\alpha(b_{\hat{h}(\hat{\lambda}(t)); n}) - \gamma) \]

\[ \Leftrightarrow \frac{1}{T} \sum_{t=1}^{T} R_S(\hat{h}) + \hat{\lambda}(t)(I_\alpha(b_h; n) - \gamma) \geq \frac{1}{T} \sum_{t=1}^{T} R_S(\hat{h}(\hat{\lambda}(t))) + \hat{\lambda}(t)(I_\alpha(b_{\hat{h}(\hat{\lambda}(t)); n}) - \gamma) \]

\[ \Leftrightarrow R_S(\hat{h}) + \frac{1}{T} \sum_{t=1}^{T} \hat{\lambda}(t)(I_\alpha(b_h; n) - \gamma) \geq \frac{1}{T} \sum_{t=1}^{T} R_S(\hat{h}(\hat{\lambda}(t))) + \hat{\lambda}(t)(I_\alpha(b_{\hat{h}(\hat{\lambda}(t)); n}) - \gamma) \]

\[ \Leftrightarrow \min_{h \in H} R_S(h) + \frac{1}{T} \sum_{t=1}^{T} \hat{\lambda}(t) \cdot (I_\alpha(b_h; n) - \gamma) \geq \frac{1}{T} \sum_{t=1}^{T} \min_{h \in H} R_S(h) + \hat{\lambda}(t) \cdot (I_\alpha(b_h; n) - \gamma) \]

(by the definition of $\hat{h}$ and $\hat{h}(\hat{\lambda}(t))$)

which shows that the second inequality, (67), holds.

By (65) and (68), it hold that

\[ \max_{\lambda} L(\bar{h}, \lambda) \leq L^*_\alpha + 2A_\alpha \sqrt{\frac{\ln 2}{T}}. \]

By (60) and (69), it holds that

\[ \min_h l(\hat{h}, \hat{\lambda}) \geq L^*_\alpha - 2A_\alpha \sqrt{\frac{\ln 2}{T}}. \]

If we set $\nu = 2A_\alpha \sqrt{\frac{\ln 2}{T}}$, then we get the value of $T$, which is $T = \frac{4A^2_\alpha \ln 2}{\nu^2}$. Hence after $T$ iterations, the outcome of Algorithm 1 satisfies (12) and (12).

**G Proof of Theorem 5.2**

From the assumption $\inf_{D \in \mathcal{H}} I_\alpha(b_h; n) \leq \gamma$, we can find an optimal $D^*_\alpha \in \Delta \mathcal{H}$ such that

\[ D^*_\alpha \in \arg\min_{D \in \Delta \mathcal{H}} R_S(D) \] subject to $I_\alpha(D; n) \leq \gamma$.

Since $I_\alpha(D^*_\alpha; n) - \gamma \leq 0$, we have for any $\lambda \in [0, \lambda_{\max}]$,

\[ L(D^*_\alpha, \lambda) = R_S(D^*_\alpha) + \lambda(I_\alpha(D^*_\alpha; n) - \gamma) \leq R_S(D^*_\alpha). \tag{70} \]

Case i) $I_\alpha(D; n) - \gamma \leq 0$:

In this case, it is enough to check $R_S(D) \leq L^* + 2\nu$. From the assumption $I_\alpha(D; n) - \gamma \leq 0$, we have that

\[ L(D, \lambda) = R_S(D) + \lambda(I_\alpha(D; n) - \gamma) \leq R_S(D) \]

55
and know that $\max_{\lambda \in \Lambda} L(\bar{D}, \lambda) = R_S(\bar{D})$ when $\lambda = 0$. Therefore

$$R_S(\bar{D}) = \max_{\lambda} L(\bar{D}, \lambda)$$

$$\leq L^* + \nu \quad \text{(by Theorem 5.1)}$$

$$\leq \min_{\bar{D}} L(D, \bar{\lambda}) + 2\nu \quad \text{(by Theorem 5.1)}$$

$$\leq L(D^*_\Lambda, \bar{\lambda}) + 2\nu$$

$$\leq \max_{\lambda \in \Lambda} L(D^*_\Lambda, \lambda) + 2\nu$$

$$= L(D^*_\Lambda, \lambda^*_\Lambda) + 2\nu = L^* + 2\nu.$$

Case ii) $I_\alpha(\bar{D}; n) - \gamma > 0$:

Since $(\bar{D}, \bar{\lambda})$ satisfies (11) and (12) by Theorem 5.1, we have

$$\max_{\lambda \in \Lambda} L(\bar{D}, \lambda) - \nu \leq L^* \leq \min_{\bar{D} \in \Delta H} L(D, \bar{\lambda}) + \nu. \quad (71)$$

By (71), we have

$$\max_{\lambda \in \Lambda} L(\bar{D}, \lambda) \leq \min_{\bar{D} \in \Delta H} L(D, \bar{\lambda}) + 2\nu \quad (72)$$

By the assumption $I_\alpha(\bar{D}; n) > \gamma$, we have

$$\max_{\lambda \in \Lambda} L(\bar{D}, \lambda) = R_S(\bar{D}) + \lambda_{\max}(I_\alpha(\bar{D}; n) - \gamma) \quad (73)$$

$$\geq R_S(\bar{D}). \quad (74)$$

By (72) and (74), we have

$$R_S(\bar{D}) \leq \max_{\lambda \in \Lambda} L(\bar{D}; n) \leq L^* + 2\nu.$$

Moreover,

$$\lambda_{\max}(I_\alpha(\bar{D}; n) - \gamma) \leq \max_{\lambda \in \Lambda} L(\bar{D}, \lambda) \quad \text{(by (73))}$$

$$\leq L^* + 2\nu \quad \text{(by (72))}$$

$$\leq R_S(D^*_\Lambda) + 2\nu \quad \text{(by (70) and $L^* = L(D^*_\Lambda, \lambda^*_\Lambda)$)}$$

$$\leq 1 + 2\nu. \quad \text{(since $R_S(D^*_\Lambda) \leq 1$)}$$

Hence, $I_\alpha(\bar{D}; n) < \gamma + \frac{2\nu}{\lambda_{\max}}$.  

56