Theoretical Analyses of Learning under Fairness

March 2018

Kazuto Fukuchi

Theoretical Analyses of Learning under Fairness

Graduate School of Systems and Information Engineering
University of Tsukuba

March 2018

Kazuto Fukuchi

Abstract

Machine learning algorithms have been widely incorporated into systems that make serious decisions, such as employment, loan, and insurance. Such serious decisions must be *fair*; that is, these decisions must not be biased on the individuals' sensitive attributes, including race, gender, religion, and ethnicity. Recently, social communities intensely demand fairness for decisions made by the machine learning algorithms as reported in (Barocas et al. 2016). In this thesis, we tackle problems of fairness in machine learning, especially in the viewpoint of the theoretical analysis. In particular, we deal with two difficulties in developing and analyzing machine learning algorithms with the fairness guarantee; that is, *disparate impact* and *populational fairness*. The disparate impact is a legal term meaning that unfair decisions can be made even if we make these decisions using the individual's non-sensitive attributes only. Even though the disparate impact is unintentional, it should be sufficiently suppressed when we make decisions (as discussed in Barocas et al. 2016). The problem of the populational fairness is that even if we learn a fair predictor for observed training samples, an unfair decision can be made for a test sample due to overfitting to the training samples. To make a fair decision for a test sample, we need to make the learned predictor fair in the underlying populational distribution from which the samples are generated.

Our contributions are developments of analyzing techniques and machine learning algorithms concerning the disparate impact and the populational fairness. In this thesis, we show three results. First, we develop a variety of maximum likelihood estimations that can remove the disparate impact even if the sensitive attributes are not involved in the given training samples. Second, we develop an empirical risk minimization framework with penalization on unfairness. By theoretically analyzing the framework, we show that a learning algorithm following this framework ensures the populational fairness. Third, we develop a minimax optimal estimator for the class of *additive functional*, which contains measures of the populational fairness.

Acknowledgements

First and foremost I wish to offer my immeasurable gratitude to my adviser, Dr. Jun Sakuma. He has supported my researches since I was an undergraduate student. He allowed me to go my own way with supportive counseling and insightful discussions. Thanks to that, I have conducted profoundly fulfilling studies.

I would particularly like to thank Dr. Toshihiro Kamishima for having fruitful discussions many times. His deep knowledge regarding fairness helps me to advance my studies.

I would like to express my gratitude to thesis committees for their insightful comments. I would like to again thank Toshihiro Kamishima to a participant as a thesis committee. Besides, I would like to show my greatest appreciation to Dr. Hiroyuki Kitagawa, Dr. Kazuhiro Fukui, and Dr. Hideitsu Hino for discussing my thesis as thesis committees, which is helpful for writing this thesis.

I gratefully acknowledge the work of past and present members of my laboratory. Eager discussion with them helps me to improve my studies.

At last but not least, I thank my family for their kindly support to provide a comfortable environment where I concentrate on my studies.

Contents

List of Figures

List of Tables

Notations

Notation	Description
$A, B, C, ..., Z$	A random variable is denoted by a upper case letter.
$a, b, c, ..., z$	An instantiation of a random variable is denoted by a corresponding lower case letter.
$\mathbb{R}$	Set of all reals.
$\mathbb{N}$	Set of all natural numbers where 0 is exclusive.
$(\cdot)_{t\in T}$	A family indexed by a set T.
$[m]$	$\{1, ..., m\}$ where m is a positive integer.
$a_m \lesssim b_m$	For sequences a_m and b_m, $a_m \lesssim b_m$ if there exists an universal constant $c > 0$ such that $a_m \leq cb_m$ for any $m \in \mathbb{N}$.
$a_m \gtrsim b_m$	For sequences a_m and b_m, $a_m \gtrsim b_m$ if there exists an universal constant $c > 0$ such that $a_m \geq cb_m$ for any $m \in \mathbb{N}$.
$a_m \asymp b_m$	The sequences satisfy $a_m \lesssim b_m$ and $a_m \gtrsim b_m$.
$\mathbb{P}\mathcal{E}$	Probability that an event $\mathcal{E}$ occurs.
$\mathrm{P}_\rho(\cdot)$	A probability density function with a probability measure ρ. For example, letting X be a random variable, $\mathrm{P}_\rho(X)$ denotes a function $x \to \frac{d\rho}{d\mu}(x)$ where μ is an appropriate base measure. The instance of X is denoted by a lowercase letter x, and we use $\mathrm{P}_\rho(x)$ to denote the density $\frac{d\rho}{d\mu}(x)$.

Notation	Description
$\mathrm{P}(\cdot)$	A probability density function with an appropriate probability measure. $\mathrm{P}(X)$ and $\mathrm{P}(x)$ are equivalent to $\mathrm{P}_\rho(X)$ and $\mathrm{P}_\rho(x)$ with an appropriate ρ, respectively. For random variables X and Y, $\mathrm{P}(X\|Y)$ denotes a function $(x,y) \to \frac{d\rho}{d\mu}(x,y) / \int_{x\in\mathcal{X}} \frac{d\rho}{d\mu}(x,y)dx$ where ρ is an appropriate measure of (X,Y).
$\mathbf{E}[X]$	Expectation of a random variable X.
$\mathbf{Var}[X]$	Variance of a random variable X.
$\mathbf{Bias}[X]$	Bias of a random variable X where the true parameter is 0, i.e., $\mathbf{Bias}[X] = \|\mathbf{E}[X]\|$.
TV	The total variation distance.
D_{KL}	The KL-divergence.
$\Delta(\mathcal{X})$	The set of all probability measure on $\mathcal{X}$ where $\mathcal{X}$ is some set.

Chapter 1

Introduction

Recently, machine learning algorithms have been widely incorporated into systems that make serious decisions, including systems that deal with employment, loans, and insurance. Social communities demand that such serious decisions must be fair and must not be biased on sensitive attributes of individuals, such as race, gender, religion, and ethnicity. For example, a 2014 White House Report (Podesta et al. 2014) mentioned "[t]he increasing use of algorithms to make eligibility decisions must be carefully monitored for potential discriminatory outcomes for disadvantaged groups, even absent discriminatory intent". The similar statement also appears in a 2016 White House Report (Munoz et al. 2016). Since we here want to avoid discriminatory or unfair treatment and make decisions neutral against the sensitive attribute, this problem is referred to as *anti-discrimination*, *fairness*, or *neutralization*.

With this requirement, we address supervised learning problems in which we want to learn a predictor so that it does not make unfair predictions. The application examples of this problem are listed as follows.

Example 1 (hiring decision)**.** A company collects personal information from employees and job applicants; this information includes age, gender, race or ethnicity, place of residence, and work experience. The company uses machine learning to predict the work performance of the applicants, using information collected from employees. The hiring decision is then based on this prediction. Here, if the learned predictor changes its output by the difference of the sensitive attributes, including gender, race, and ethnicity, the predictions made by the predictor can be biased against the sensitive attributes, and the hiring decisions might be deemed discriminatory. We here want to employ a learning algorithm such that the learned predictor is not biased against the sensitive attributes.

Example 2 (personalized advertisement and recommendation)**.** A company that provides web services records user behavior, including usage history and search logs, and uses machine learning to predict user attributes and preferences. The advertisements or recommendations displayed on web pages are thus personalized so that they match the predicted user attributes and preferences. When recommendations are accurately pinpointed to sensitive issues, such as political or religious affiliation, the result may be increasingly biased views. This is known as the problem of the filter bubble (Pariser 2011). For example, suppose supporters of the Democratic Party wish to read news articles related to politics. If the recommended articles are all related to their party and are absent of criticism, they may develop a biased view of the political situation. The objective of the learner is again to learn a predictor that is not biased against the sensitive attributes.

Machine learning algorithms are not designed to make unfair decisions intentionally, whereas it makes unfair decisions. For instance, Sweeney (2013) demonstrated that an online ads system for a search engine more frequently presented negative ads when querying names of African descent than when querying names of European descent. Although the system merely personalizes ads to maximize the click-through rate by utilizing the machine learning techniques, such discrimination can occur. Unintentional unfair decisions can occur when training samples are unfair potentially. As reported Barocas et al. (2016), biased decisions can be made by machine learning algorithms because of biased training samples.

Social communities demand to avoid such unintentional unfair decisions even if the training samples are biased against the sensitive attributes. As mentioned above, the line of the 2014 White House Report says that we must monitor for potential unfair decisions "even absent discriminatory intent". Consequently, we want to develop machine learning techniques to avoid unintentional unfair decisions made by the learning algorithms. If the training samples are biased against the sensitive attributes, to make an accurate prediction and to make a fair prediction are conflicting tasks. Therefore, the goal of the learning algorithm under fairness is to achieve the most efficient trade-off between prediction accuracy and fairness.

There are two main difficulties to learn a fair predictor; that is, *disparate impact* and *populational fairness*.

Disparate impact. The disparate impact is a legal term meaning that the unfair decisions can be made even if we make these decisions using the individual's non-sensitive attributes only. It comes from the dependent relationship between the sensitive attributes and the non-sensitive attributes. For example, suppose we make a hiring decision to an applicant from the information regarding his/her attributes, including name, address, gender, academic background, and job history. Since the gender is a sensitive attribute, we need to make the hiring decision from the attributes other than gender. However, name, academic background, and job history are often dependent on gender, and thus the decision made by these attributes also can depend on gender. Even though the disparate impact is unintentional, it should be sufficiently suppressed when we make decisions (as discussed in Barocas et al. 2016). Such dependent relationship between the non-sensitive attributes and the sensitive attributes might hiddenly exist even if the sensitive attributes are not explicitly observed. In the hiring decision example, if gender is not contained in the training samples, we can make a biased decision against gender because the name, academic background, and job history are dependent on gender.

Populational fairness. The problem of the populational fairness is that even if we learn a fair predictor for the observed training samples, an unfair decision can be made for a test sample due to overfitting to the training samples. To make a fair decision for a test sample, we need to make the learned predictor fair in the underlying population distribution from which the samples are generated.

Our contributions are developments of analyzing techniques and machine learning algorithms concerning the disparate impact and the populational fairness. In this thesis, we show three results. First, we develop a variety of maximum likelihood estimations that can remove the disparate impact even if the sensitive attributes do not present explicitly in the given training samples but implicitly exist (in Chapter 3). Second, we develop an empirical risk minimization framework with penalization on unfairness. By theoretically analyzing the framework, we show that a learning algorithm following this framework ensures the populational fairness (in Chapter 4). Third, we develop a minimax optimal estimator for the class of *additive functional*, which contains measures of the populational fairness (in Chapter 5).

Chapter 2

Background and Related work

In this chapter, we review some background of machine learning under fairness. Especially, we focus our attention on supervised learning.

2.1 Supervised Learning under Fairness

In this section, we setup a problem of supervised learning under fairness. Let X and Y be random variables representing *input* and *target* where these domains are denoted as $\mathcal{X}$ and $\mathcal{Y}$, respectively. In addition to X and Y, we introduce a random variable V representing *viewpoint* which denotes the sensitive attribute including gender, race, religion, and ethnicity, where the domain of V is denoted as $\mathcal{V}$. Suppose X, Y, and V are jointly distributed as a probability measure ρ. In the ordinary setting of supervised learning under fairness, the learner obtains n i.i.d. training samples $D_n = ((X_i, Y_i, V_i))_{i\in[n]} \in (\mathcal{X} \times \mathcal{Y})^n$ where (X_i, Y_i, V_i) are independent copies of (X, Y, V). Given the training samples D_n, the learner constructs a predictor f_n which is a (possibly stochastic) mapping from $\mathcal{X}$ and $\mathcal{V}$ to $\mathcal{Y}$. After constructing the predictor f_n, the learner can predict a target value by using values of input and viewpoint. Formally, a test sample (X_0, Y_0, V_0) is drawn independently from ρ, and X_0 and V_0 are revealed to the learner. Then, the learner makes an prediction on Y_0 from X_0 and V_0 as $\hat{Y}_0 = f_n(X_0, V_0)$. Here, the learner has two objectives:

Accuracy the prediction $\hat{Y}_0$ is accurate.
Fairness the prediction $\hat{Y}_0$ is not biased against V_0.

The learner's goal is to construct f_n that achieves these objectives simultaneously. If these objectives are conflicting and are in a trade-off relation, the goal then becomes constructing f_n that achieves the most efficient trade-off among them.

To achieve accurate prediction, the learner wants to construct f_n so that $\hat{Y}_0$ is closest to Y_0. The difficulty of this problem mainly comes from unknownness of ρ. The learner cannot directly access to ρ, and instead observes the training samples D_n drawn i.i.d. from ρ. On the other hand, the performance of the learned predictor f_n is evaluated at the test sample (X_0, Y_0, V_0) which is not contained in D_n, but is independently drawn from the unknown probability measure ρ. Hence, even if the learner constructs predictor f_n so that $f_n(X_i, V_i) = Y_i$ for $i \in [n]$, it can lead a large error for the test sample. The error for the test sample comes from a large complexity of the predictor f_n. For example, suppose the learner employs a quadratic function for the predictor f_n even when the relationship between (X, V) and Y is linear with no

noise. Then, if the sample size n is smaller than the dimensionality of (X, V), we can construct a quadratic function such that $f_n(X_i, V_i) = Y_i$ for $i \in [n]$. This phenomenon is called as *overfitting*; that is, by employing a too complex model that cannot be explained by the obtained samples, the learner incurs a large error at the test time.

To overcome this difficulty, many methodologies have been developed by machine learning researchers over recent decades. We introduce the most generic two approaches to supervised learning in the subsequent subsections.

2.1.1 Maximum Likelihood Estimation and Maximum a Posteriori

Suppose that the probability measure ρ is parametrized by $\theta \in \Theta$ as ρ_θ. Given the samples D_n drawn i.i.d. from ρ_θ, the goal of the learner is to find θ from D_n. To this end, the learner finds a parameter θ that maximizes the probability with which the samples D_n are obtained from ρ_θ. Formally, the estimated parameter θ_n is obtained as

$$\theta_n = \arg\max_{\theta \in \Theta} L(\theta) = \arg\max_{\theta \in \Theta} \left\{ \sum_{i=1}^{n} \ln \mathrm{P}_{\rho_\theta}(X_i, Y_i, V_i) \right\}. \tag{2.1}$$

Noting that the logarithmic function is a continuous and monotonically increasing function, Eq (2.1) maximizes the probability $\prod_{i=1}^{n} \mathrm{P}_{\rho_\theta}(X_i, Y_i, V_i)$. The function of θ that outputs the probability with which the samples D_n are obtained from ρ_θ is called as a *likelihood*, and hence this method is named as *maximum likelihood estimation (MLE)*. However, this method can lead overfitting if the parametrized model ρ_θ is too complex as explained above.

To avoid overfitting, we ordinary introduce a prior probability measure π over Θ, and then maximizes the posterior probability as follows

$$\theta_n = \arg\max_{\theta \in \Theta} \{ L(\theta) + \ln \mathrm{P}_\pi(\theta) \}.$$

This method is referred to as the *maximum a posteriori (MAP)*.

2.1.2 Empirical Risk Minimization

Let $\mathcal{F}$ be a set of predictors. Letting $\ell(f, (X, Y, V))$ be a *loss function* which assesses the error of f at a point (X, Y, V), we can define the optimal predictor

$f^* \in \mathcal{F}$ that maximizes the *generalization risk* as

$$f^* = \arg\max_{f \in \mathcal{F}} R(f) = \arg\max_{f \in \mathcal{F}} \big\{ \mathbf{E}_{(X,Y,V)\sim\rho}[\ell(f,(X,Y,V))] \big\}.$$

The goal of the learner is to find the optimal predictor f^* using the samples D_n. Since the learner cannot access the generalization risk directly due to unknown ρ, we instead find f_n that minimizes the *empirical risk* as

$$f_n = \arg\min_{f \in \mathcal{F}} R_n(f) = \arg\min_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(f,(X_i,Y_i,V_i)) \right\}.$$

This method is called as *empirical risk minimization (ERM)*.

If the domain of predictors $\mathcal{F}$ is too large, the problem of overfitting can occur. To suppress overfitting, we ordinary employ the *regularizer* $\Omega : \mathcal{F} \to \mathbb{R}$, which penalizes a large complexity of f_n, and find f_n as

$$f_n = \arg\min_{f \in \mathcal{F}} \{ R_n(f) + \Omega(f) \}.$$

This method is referred to as *regularized empirical risk minimization (RERM)*.

For the fairness objective, we first need to discuss a topic regarding when the prediction $\hat{Y}_0$ is biased against the viewpoint V_0. In the next subsection, we clarify this question by introducing the basic concept of the fairness definitions appeared in the literature.

2.1.3 Legal Theories of Liability for Fairness

With the setup described in Section 2.1, we here discuss about the situation that the prediction $\hat{Y}_0$ is biased against the viewpoint V_0. The basic concept behind this comes from the discussion of legal theories of liability (see Barocas et al. 2016, and references therein). Basically, there are two types of legal theories of liability with regarded to fairness; *disparate treatment* and *disparate impact*. Disparate treatment is a legal theory of liability for both explicit formal classification and intentional discrimination. More specifically, there is a disparate treatment if decisions made for an individual receives change with a difference of her/his sensitive attributes. This notion is equivalent to *direct discrimination* in (Pedreschi et al. 2008). In supervised learning, a predictor that uses the viewpoint V and changes its output with a change of V has a disparate treatment. Formally, a predictor f suffers from a disparate treatment if a prediction $\hat{Y} = f(X)$ satisfies for any $v \in \mathcal{V}$,

$$\mathrm{P}(\hat{Y}|X,v) \neq \mathrm{P}(\hat{Y}|X). \tag{2.2}$$

If a predictor f has a disparate treatment, f explicitly uses the viewpoint V as an argument and changes its output from the viewpoint V. Hence, we can say that f makes an unfair prediction against the viewpoint V intentionally. The intentional unfair prediction can be easily removed with a little effort. For example, we can easily remove the disparate treatment by removing the viewpoint from the training samples D_n and constructing f_n as a function of the input X.

Disparate impact occurs when the decisions is facially fair but have a bias against the viewpoint. There is a disparate impact if the decisions made for an individual outcomes a disproportionately adverse impact to members of a certain sensitive attribute value group. Formally, a predictor f suffers from a disparate impact if a prediction $\hat{Y} = f(X)$ satisfies for any $v \in \mathcal{V}$,

$$\mathrm{P}(\hat{Y}|v) \neq \mathrm{P}(\hat{Y}). \tag{2.3}$$

The disparate impact exists regardless of intentionality. We remark that machine learning algorithms can make an unfair prediction unintentionally. In the Sweeny's example, the ad provider does not intentionally display the unfair ads, but it merely distributes the ads so that the click-through rate is maximized. The cause of the disparate impact is a high dependent relationship between the input X and the viewpoint V. For example, a postal address of individuals is usually highly dependent on ethnicity due to historical reasons, and therefore a prediction made from the address can be biased against ethnicity. Such dependent relationship between the prediction and the viewpoint through the input is called as *indirect discrimination* (Pedreschi et al. 2008) or *red-lining effect* (Calders et al. 2010).

Biased unfair target values in the training samples account for the most of causes of the disparate impact. Typically, target values in training samples are assigned by a human; in the employment example, the employments of the applicants are finally decided by a person in the personnel section. In this case, the person perhaps determines the target values so that these are biased against the viewpoint values subconsciously. Also, as reported Barocas et al. (2016), biased decisions can be made by machine learning algorithms because of biased training samples. Consequently, learning with potentially unfair training samples results in the disparate impact. If the training samples are unfair, the perfect prediction, i.e., the prediction $f_n(X)$ equivalents to the target Y, can also be unfair. In this case, the accuracy objective and the fairness objective are conflicting, and the learner should aim to achieve the most efficient trade-off among them.

The most of the existing works that tackle the problems of supervised learning

under fairness aim to remove the disparate impact from the learned predictor f_n. The difficulty here is again in unknownness of ρ. Since ρ is unknown to the learner, she also does not know the actual dependency relationship between X and V. Hence, the learner should estimate the dependency relationship from given training samples D_n, and at the same time, she should carefully construct f_n so that it does not involve any factors in X depends on V by using the estimated dependency relationship. As well as the accuracy objective, the main part of difficulty in removing the disparate impact comes from overfitting. We will explain this problem in detail in the next section.

In addition to the difficulty coming from the unknownness of ρ, there is another difficulty of the hidden red-lining effect. Since the disparate impact comes from the indirect dependency relationship between $f_n(X)$ and V through X, the same phenomenon occurs even if the viewpoint hiddenly exists and its values are not contained in the training samples explicitly. Even though the disparate impact comes from the hidden viewpoint, it is often the case that we can check the occurence of the disparate impact easily. In the Sweeny's example, information about African-descent or Europian-descent is not contained in the search queries explicitly, whereas she determined whether a person name is African-descent or Europian-descent by using public information published by the government. Hence, by utilizing the public information, we can easily infer the value of the hidden viewpoint from the input. As Sweeny did, we can eaxily check the occurence of the disparate impact by using the inferred hidden viewpoint. Consequently, we need to remove the disparate impact causing from the hidden viewpoint. In Chapter 3, we will address this problem.

There is a line of literature that aims to avoid another legal theory of liability, *disparate mistreatment* (such as, Hardt et al. 2016; Woodworth et al. 2017; Zafar et al. 2017). This is specially designed for algorithms of supervised learning. Formally, a predictor f suffers from a disparate mistreatment if a prediction $\hat{Y} = f(X, V)$ satisfies for any $v \in \mathcal{V}$,

$$\mathrm{P}(\hat{Y}|Y, v) \neq \mathrm{P}(\hat{Y}|Y). \tag{2.4}$$

The concept of the disparate mistreatment comes from the reliability of target Y. As mentioned before, target values are usually assigned by human, and thus are unreliable. In this case, although the learner does not know whether or not the unrealiable target values are actually unfair, she should remove the disparate impact conservatively. On the other hand, there is the case that we can believe the given target values are fair. For example, suppose admission of a school is determined by whether or not the score of the entrance

examination is higher than 80%. In this case, admission is determined by applicant's academic ability systematically, and thus is reasonable even if enrolled students are biased against a sensitive attribute. Given such target Y, we permit that prediction $\hat{Y}$ is biased against V with the same reason of Y. In Eq (2.4), by adding Y as a condition, we permit that the predictor changes its output with a change of V along with a change of Y.

The cause of the disparate mistreatment is no longer the red-lining effect, but mainly consists of *underfitting*. Underfitting is a similar term of overfitting and means that, by employing a too simple model that cannot explains the obtained samples, the learner incurs a large error at both the training and test times. When the learned predictor f_n underfits the training samples, $f_n(X)$ is far from Y and hence can invalidate the condition in Eq (2.4). The line of works in (Hardt et al. 2016; Woodworth et al. 2017; Zafar et al. 2017) developed methods that remove the bias in the prediction against the viewpoint generated due to underfitting.

Quantification. Some of the existing works employed a quantified measure of Eq (2.2), Eq (2.3), or Eq (2.4). Typically, the disparity in Eq (2.2), Eq (2.3), or Eq (2.4) is measured by using some divergence, such as the total variation distance, KL-divergence, and χ^2 divergence. For example, letting the viewpoint be binary as $\mathcal{V} = \{v_+, v_-\}$, we can measure unfairness caused by the disparate impact by the total variation distance (Calders et al. 2010; Zemel et al. 2013):

$$\mathrm{TV}(\mathrm{P}(\hat{Y}|v_+), \mathrm{P}(\hat{Y}|v_-)).$$

We can use any divergence measure instead of the total variation distance. For another quantification, the mutual information between $\hat{Y}$ and V is employed as an unfairness measure (Kamishima et al. 2012b), where the mutual information is defined as

$$I(\hat{Y}; V) = D_{\mathrm{KL}}(\mathrm{P}(\hat{Y}, V), \mathrm{P}(\hat{Y})\mathrm{P}(V)) = \mathbf{E}\left[\ln\left(\frac{\mathrm{P}(\hat{Y}, V)}{\mathrm{P}(\hat{Y})\mathrm{P}(V)}\right)\right].$$

where D_{KL} denotes the KL-divergence.

2.1.4 Empirical and Populational Fairness

To achieve fairness objective, the learner constructs a predictor f_n so that Eq (2.2), Eq (2.3), or Eq (2.4) is satisfied. However, Eqs (2.2) to (2.4) cannot be evaluated directly due to unknown ρ. Instead, the learner empirically

evaluates Eqs (2.2) to (2.4) by using given training samples D_n. Then, the learner constructs a predictor f_n so that the empirically evaluated fairness is ensured. If a learning algorithm ensures such empirically evaluated fairness, we say the learning algorithm ensures *empirical fairness*. However, unfair predictions can be made at test time even if empirical fairness is guaranteed due to overfitting. Supervised learning aims to make predictions for the test samples, and thus a learning algorithm that does not ensure the test time fairness is far from our desire.

The actual desire of the fairness objective is to guarantee fairness at the test time. To this end, we aim to design the learning algorithm that ensures *populational fairness*, in which fairness is ensured for a sample generated from the underlying distribution ρ. Since ρ is unknown to the learner, what the learner can do is empirical evaluation of fairness. Thus, theoretical analysis of the learning algorithm is mandatory to ensure populational fairness. In Chapters 4 and 5, we will deal with problems of guaranteeing and evaluating the populational fairness.

2.1.5 Existing Fair Mechanisms

We clarified the definition of fairness in Section 2.1.3. We here discuss regarding when and how to achieve Eq (2.2), Eq (2.3), or Eq (2.4) in the existing studies. The existing methods of "when" can be divided into three types; that is, *preprocessing*, *interprocessing*, and *postprocessing*.

Preprocessing. A preprocessing method converts the training samples D_n to D'_n so that D'_n is to be fair samples. For example, Feldman et al. (2015) proposed methods of clarifying the existence of the disparate impact and of removing the disparate impact in a preprocessing manner. They prove that predictability of the viewpoint from the input accounts for a cause of the disparate impact, and introduce a support vector machine based method to clarify the disparate impact. This only ensures empirical fairness, and thus there is no guarantee of populational fairness.

Postprocessing. A postprocessing method modifies the learned predictor f_n to f'_n so that f'_n will make a fair prediction where f_n is constructed by some ordinary learning algorithm. As an instance of postprocessing method, Calders et al. (2010) presented the *Calders–Verwer 2 naïve Bayes method (CV2NB)*, which proactively removes the red-lining effect. Suppose the target and the viewpoint are both binary as $\mathcal{Y} = \{y_+, y_-\}$ and $\mathcal{V} = \{v_+, v_-\}$. Then, the Calders–Verwer (CV) score is defined by $\mathrm{CV}(D_n) = \mathrm{P}_{D_n}(Y = y_+|V =$

$v_+) - \mathrm{P}_{D_n}(Y = y_+|V = v_-)$ where P_{D_n} denotes the empirical probability density calculated from D_n. The CV2NB modifies the näive Bayes classifier in such a way that the CV score becomes zero concerning the given samples D_n. Based on this idea, various situations where discrimination can occur have been discussed in other studies (Kamiran et al. 2010; Zliobaite et al. 2011b). For example, it has been shown by Zliobaite et al. (2011a) that positive CV scores do not necessarily cause unfairness in some situations. These methods only ensure empirical fairness.

Hardt et al. (2016) developed postprocessing methods of removing the disparate mistreatment from the ordinary learned predictor. Their methods can be applied to a binary classification function mapping from an input to a binary target and a binary score-based classification function which maps an input to a real value where the target value is determined by thresholding the outputted real value. They analyzed the error caused by approximation of the probability distribution ρ, and in this sense, their method can ensure populational fairness.

Interprocessing. In interprocessing, a mechanism to make fair predictions is combined with the learning algorithm. As an interprocessing method, Kamishima et al. (2012b) presented a prejudice remover regularizer (PR) for removing the disparate impact in classification setting. This method is formulated as an optimization problem in which the objective function contains the loss term and the regularization term that penalizes mutual information between $\hat{Y}$ and V. The value of the mutual information represents the quantification of the disparate impact, and hence we can get a fair classifier by penalizing the objective function using the mutual information. This method does not have any guarantee of populational fairness.

Zemel et al. (2013) introduced an interprocessing method, the learning fair representations (LFR) model, for preserving the disparate impact in classification setting. LFR is designed to provide a map,from inputs to prototypes, which guarantees the classifiers that are learned with the prototypes will be fair. Hence, the prototypes are constructed so that predictions made from the prototypes are fair. This method also does not ensure populational fairness, either.

Zafar et al. (2017) proposed a method of removing the disparate mistreatment. Their method is based on the empirical risk minimization (ERM) where it is constrained so that the learned classifier does not have the disparate mistreatment. This method does not have any guarantee of populational fairness, either.

Woodworth et al. (2017) developed and analyzed a method of removing the disparate mistreatment. They showed that any postprocessing method as (Hardt et al. 2016) do fails to remove the disparate mistreatment. Hence, they combine interprocessing and postprocessing; in which, the learner first conducts the ERM with a fairness constraint, and then converts the learned predictor to remove the disparate mistreatment. They analyzed the two-phases method and provided a guarantee of populational fairness.

2.2 Related Work for Other Machine Learning Problems

We formalize the problem of supervised learning under fairness and discuss the difficulties in that thus far. In the meanwhile, many researchers addressed problems of fair machine learning other than supervised learning, such as unsupervised learning, bandit learning, and reinforcement learning. In this section, we introduce some related works that tackle such machine learning problems under fairness. Bolukbasi et al. (2016) demonstrated that an unfair relationship of words is learned by the word embedding algorithm using a deep learning technology. The word embedding algorithm enables us to extract relationship between words; for example, the algorithm can extract relationship men $-$ women $\approx$ king $-$ queen. They have reported that the word embedding system learns a relationship men $-$ women $\approx$ *computerprogrammer* $-$ homemaker; which is a biased view of the job against gender. They proposed a word embedding algorithm that can remove such a sensitive relationship.

Some researchers tackled the fairness problem appeared in the bandit problem or the reinforcement learning problem Jabbari et al. (2017) and Joseph et al. (2016). Their methods attempt to remove the disparate mistreatment caused by underfitting, and have guarantees of populational fairness.

Chapter 3

Fair Prediction with Model-based Sensitive Attribute

Disparate impact is a fairness definition for a situation when the learner can make unfair predictions unintentionally due to the potentially unfair training samples. As mentioned in Chapter 2, the main cause of the disparate impact is the red-lining effect; that is, use of biased input values results in biased prediction. The red-lining effect can occur even when the viewpoint hiddenly exists and its values are not contained in the training samples explicitly. In this chapter, we deal with a problem of removing the disparate impact coming from such a hidden viewpoint.

Several techniques that take account of fairness or discrimination have recently received attention (Bolukbasi et al. 2016; Calders et al. 2010; Dwork et al. 2012; Feldman et al. 2015; Kamiran et al. 2010; Kamishima et al. 2012b; Romei et al. 2013; Ruggieri et al. 2010; Zemel et al. 2013). One of the easiest ways to suppress unfair treatment is to remove the values of the viewpoint from the input values before the learning process of the prediction model. This procedure can remove the disparate treatment from the learned prediction model, whereas the disparate impact remains in the model due to the red-lining effect (Table 3.1, line 1).

Calders et al. (2010) presented the Calders–Verwer 2 naïve Bayes method (CV2NB), which proactively removes the red-lining effect. The CV2NB guarantees the elimination of unfairness regarding the CV score. The limitation of the CV2NB is that it cannot be used when the target or viewpoints are continuous. Moreover, this method requires the explicit viewpoint values in the training samples and thus is not applicable when the viewpoint is hidden; that is, when viewpoint values are not contained in training samples explicitly, but the disparate impact can occur through the input that highly depends on the viewpoint (Table 3.1, line 2). The methods based on the CV2NB, such as (Thanh et al. 2011; Zliobaite et al. 2011a), share its limitations.

Kamishima et al. (2012b) introduced the prejudice remover (PR) for a classification task. This method employs the mutual information between the target and the viewpoint, which is called as the *prejudice index*, as a quantification measure of unfairness. This method can work with a continuous target if it is approximated by a histogram, as demonstrated by Kamishima et al. (2012a, 2013). Continuous viewpoints, however, cannot be treated by the PR method. The PR method cannot deal with the hidden viewpoint, either (Table 3.1, line 3).

Zemel et al. (2013) proposed *learning fair representation (LFR)*, aiming to obtain an invariant intermediate representation against the viewpoint. The fairness of LFR is based on statistical parity with quantification by the total variation distance. This method can be applied to the binary or multiple

Table 3.1: Summary of learning algorithms with fairness guarantee.

method	fairness guarantee	target domain	viewpoint domain	viewpoint model
elimination of viewpoint	no guarantee	any	any	×
CV2NB (Calders et al. 2010)	CV Score	multiple	multiple	×
PR (Kamishima et al. 2012b)	mutual information	any	multiple	×
LFR (Zemel et al. 2013)	statistical parity	multiple	multiple	×
η-neutral logistic regression (proposal)	η-neutrality	multiple	multiple	√
η-neutral linear regression (proposal)	η-neutrality	continuous	continuous	√

target and viewpoint; however, continuous viewpoints are not considered. The hidden viewpoint is also unavailable (Table 3.1, line 4).

Our contributions.

Our contributions in this chapter are three-folds; modeling viewpoint, maximum likelihood estimation with η-neutrality, and comparison of fairness measures.

Modeling viewpoint Existing methods assume the viewpoint is observed and is explicitly provided in the input. Therefore, these methods are not applicable the hidden viewpoint. We here aim to remove the disparate impact of the hidden viewpoint. If the learner has no information about the hidden viewpoint, she cannot make any effort for removing the disparate impact. Instead, we here assume that the learner obtains a probabilistic model of the viewpoint, which is a model that predicts the viewpoint from the input. In many cases, we can easily construct such a probabilistic model; for example, we can construct prediction model from name to race by using the statistical survey published by the government. By using this strategy, Sweeney determined that a person name is either of Europian descent or African descent (Sweeney 2013).

We provide a method to remove the disparate impact from the target prediction model against a given probabilistic model of the viewpoint when the viewpoint is hidden. To remove the disparate impact of the hidden view-

point, we define η-neutrality (in Section 3.1.1), which measures dependency between the target model and the viewpoint model. With η-neutrality, we can evaluate the unfairness of a target prediction model against any hidden viewpoint, as long as its probabilistic model is given to the learner (Table 3.1, the rightmost column). Furthermore, since η-neutrality is measured for probabilistic models, populational fairness is expected to be effectively guaranteed. This is demonstrated by experiments in Section 3.5.

Maximum likelihood estimation with η-neutrality Following the definition of η-neutrality, we introduce a systematic way to remove the disparate impact from the target prediction model when viewpoints are hidden. This framework is based on the maximum likelihood estimation. Our methods can treat a target and viewpoint that are either discrete (Table 3.1, line 5) or continuous (Table 3.1, line 6), as demonstrated by η-neutrality with logistic regression in Section 3.3.1 and linear regression in Section 3.3.2. The effectiveness of our methods is examined by experiments with both artificial and real datasets in Section 3.5.

Comparison of fairness measures We clarify the relationship between the existing fairness measures, η-neutrality, the CV Score (Calders et al. 2010), statistical parity (Zemel et al. 2013) and the prejudice index (Kamishima et al. 2012b). For a comprehensive discussion, we introduce a *neutrality factor* (in Section 3.4), which represents fairness of a pair of a target value and viewpoint value. We show that existing fairness measures are universally represented by the aggregation of the neutrality factors.

In Figure 3.1, we illustrate the relationship between the fairness measures. The dashed arrow between statistical parity and the CV score shows that statistical parity with a binary viewpoint is equivalent to CV score. Furthermore, η-neutrality is interpreted as an upper-bound of the other fairness measures represented by the solid arrows. We will prove these relations in Section 3.4.

3.1 Problem Setting

The most of the problem setting in this chapter is borrowed from the setup of supervised learning under fairness and Section 2.1.1 described in Section 2.1. Given the training samples D_n, the goal of the learner is to find the model parameter that achieves the most efficient trade-off between the accuracy objective and the fairness objective. For a prediction model of the target, we

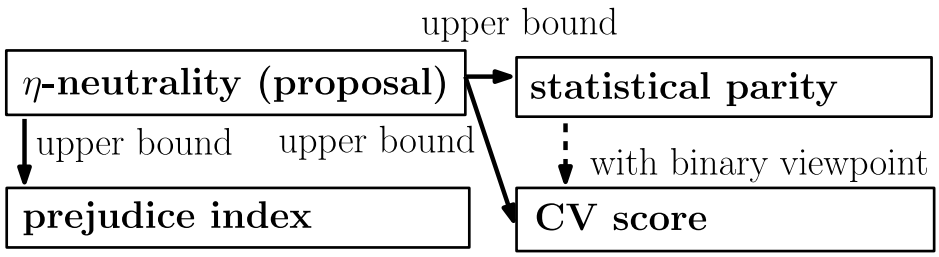


Figure 3.1: Relationship between the fairness measures. Statistical parity with binary viewpoint is equivalent to CV score. η-neutrality upper bounds all of other fairness measures.

model the conditional probability $\mathrm{P}(Y|X)$ by using some parametric model $f(Y|X;\theta)$ where $\theta \in \Theta$ is a parameter of the target prediction model.

As mentioned before, we attempt to remove the disparate impact of the hidden viewpoint by utilizing a given prediction model of the hidden viewpoint. As we did for the target, we model the conditional probability $\mathrm{P}(V|X)$ by a prediction model $g(V|X;\phi)$ where $\phi \in \Phi$ is a parameter of the viewpoint prediction model. Since the viewpoint prediction model is revealed to the learner in advance of the learning phase, we can assume the viewpoint prediction model is fixed, and so the model parameter ϕ is omitted and g is described by $g(V|X)$.

Noting that the values of the target and the viewpoint are predicted independently, we can assume the joint probability is

$$\mathrm{P}_\rho(X,Y,V) = \mathrm{P}(X)\mathrm{P}(Y|X)\mathrm{P}(V|X).$$

The target and the viewpoint prediction models describe the conditional probability density functions $f(Y|X;\theta) = \mathrm{P}(Y|X)$ and $g(V|X) = \mathrm{P}(V|X)$, respectively. Thus, given the target prediction model $f(Y|X;\theta)$ and the viewpoint prediction model $g(V|X)$, the probabilistic model of $\mathrm{P}_\rho(X,Y,V)$ becomes

$$M(X,Y,V;\theta) = f(Y|X;\theta)g(V|X)\mathrm{P}(X). \tag{3.1}$$

We first propose a novel definition of neutrality, η-neutrality. We then describe the goal of the learning algorithm. In the following discussion, we assume the input X is continuous. We can treat a discrete X by replacing the integral with a sum. For Y, the discussion below is valid for both discrete and continuous variables. As is the case with Y, the discussion below for V is also valid for both discrete and continuous variables.

3.1.1 η-Neutrality

With the setup above, we consider the dependency of the target Y and the viewpoint V. When V and Y are statistically independent, for any $v \in \mathcal{V}$ and any $y \in \mathcal{Y}$, $\mathrm{P}(v, y)/\mathrm{P}(v)\mathrm{P}(y) = 1$. When $\mathrm{P}(v, y)/\mathrm{P}(v)\mathrm{P}(y) > 1$, y and v are more dependent than independent. Hence, our definition of fairness is defined as the ratio of the marginal probabilities, as follows.

Definition 1 (η-neutrality)**.** Let X, Y, and V be the random variables representing the input, the target, and the viewpoint, respectively. Given $\eta \geq 0$, the probability distribution $\mathrm{P}(X, Y, V)$ is η-neutral if for all $v \in \mathcal{V}$ and all $y \in \mathcal{Y}$,

$$\frac{\mathrm{P}(v, y)}{\mathrm{P}(v)\mathrm{P}(y)} \leq 1 + \eta. \tag{3.2}$$

Our neutrality definition simply bounds the ratio. As a variation of this definition, the ratio can be bounded above and below as

$$\forall v \in \mathcal{V}, y \in \mathcal{Y}, \quad 1 - \eta \leq \frac{\mathrm{P}(v, y)}{\mathrm{P}(v)\mathrm{P}(y)} \leq 1 + \eta. \tag{3.3}$$

If both the target and the viewpoint are binary, the ratio of our definition is bounded below as Eq (3.3). If either or both of the target and the viewpoint take M multiple values, the ratio of our definition is bounded below by $1 - M\eta$ which is different from Eq (3.3). We employed Definition 1 for optimization efficiency. The number of constraints derived from Definition 1 can be reduced to half compared to Eq (3.3).

Next, given the probabilistic models of $\mathrm{P}(Y|X)$ and $\mathrm{P}(V|X)$, we derive conditions that the model of the joint probability distribution satisfies η-neutrality. The following theorem shows the condition that the model of Eq (3.1) is η-neutral.

Theorem 1. *Suppose the joint probability distribution of input X, target Y, and viewpoint V follows the model $M(X, Y, V; \theta) = f(Y|X; \theta)g(V|X)\mathrm{P}(X)$. Then M is η-neutral if for all $v \in \mathcal{V}$ and all $y \in \mathcal{Y}$,*

$$\int_x \mathrm{P}(x) f(y|x; \theta)[g(v|x) - (1 + \eta)\bar{g}(v)]dx \leq 0, \tag{3.4}$$

where $\bar{g}(v) = \int_x \mathrm{P}(x)g(v|x)dx$.

Proof. By the marginalization of $\mathrm{P}(x, y, v)$ with respect to x, (x, y), and (x, v), we have

$$\begin{aligned}
\mathrm{P}(y, v) &= \int_x \mathrm{P}(x, y, v)dx = \int_x \mathrm{P}(x)f(y|x;\theta)g(v|x)dx, \\
\mathrm{P}(y) &= \int_x \int_v \mathrm{P}(x, y, v)dvdx = \int_x \mathrm{P}(x)f(y|x;\theta)dx, \\
\mathrm{P}(v) &= \int_x \int_y \mathrm{P}(x, y, v)dydx = \int_x \mathrm{P}(x)g(v|x)dx \\
&= \bar{g}(v).
\end{aligned}$$

By substituting the above equations into Eq (3.2), we have

$$\begin{aligned}
&\forall v, y, \int_x \mathrm{P}(x)f(y|x;\theta)g(v|x)dx - (1+\eta)\bar{g}(v)\int_x \mathrm{P}(x)f(y|x;\theta)dx \leq 0, \\
&\forall v, y, \int_x \mathrm{P}(x)f(y|x;\theta)[g(v|x) - (1+\eta)\bar{g}(v)]dx \leq 0.
\end{aligned}$$

□

With the definition of η-neutrality, the learner's goal is to find the target model parameter θ_n that is closest to the true parameter θ under the constraint of η-neutrality. In the next section, we describe the learning algorithm we propose.

3.2 Maximum Likelihood Estimation with η-Neutrality

With the setup described in Section 3.1, we here introduce a systematic framework that estimates the target model parameter under the constraint of η-neutrality. Our framework is based on the maximum likelihood estimation (MLE), whereas it can be easily extended to maximum a posteriori (MAP). In MLE, the parameter of the target prediction model is estimated by minimization of the negative log-likelihood:

$$\theta_n = \mathrm{argmin}_{\theta\in\Theta} L(\theta),$$

where

$$L(\theta) = -\sum_{i=1}^{n} \ln f(Y_i|X_i;\theta). \tag{3.5}$$

3.2.1 Approximation of η-Neutrality

Since the probability distribution P is unknown to the learner, we cannot evaluate Eq (3.4) directly due to the term of $\mathrm{P}(X)$. When $\mathrm{P}(x)$ cannot be obtained, η-neutrality can be empirically evaluated with the frequency distribution of $\mathrm{P}(X)$ using the training samples D_n. The neutrality condition with respect to this frequency distribution is derived similarly, as follows. Given a set of samples D_n, we approximate η-neutrality with the frequency distribution

$$\hat{\mathrm{P}}(X = x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{X_i = x},$$

where $\mathbf{1}_{\cdot}$ denotes the indicator function. From this, we have

$$\hat{\mathrm{P}}(X, Y, V) = \hat{\mathrm{P}}(X)\mathrm{P}(Y|X)\mathrm{P}(V|X),$$

and an approximation of η-neutrality is defined by this $\hat{\mathrm{P}}(X, Y, V)$.

Definition 2 (Empirical η-neutrality)**.** Let X, Y, and V be the random variables representing the input, the target, and the viewpoint, respectively. Let $\hat{\mathrm{P}}(X)$ be the frequency distribution of X obtained from the samples D_n. Given $\eta \geq 0$, if $\hat{\mathrm{P}}(X, Y, V)$ is η-neutral, $\mathrm{P}(X, Y, V)$ is said to be empirically η-neutral with respect to the samples D_n.

The following theorem shows the condition that the model in Eq (3.1) is empirically η-neutral with respect to the given samples.

Theorem 2. *Suppose the joint probability distribution of the input X, target Y, and viewpoint V follows the model $M(X, Y, V; \theta) = \mathrm{P}(X)f(Y|X;\theta)g(V|X)$. Then, given the samples D_n, M is empirically η-neutral if for all $y \in \mathcal{Y}$ and all $v \in \mathcal{V}$,*

$$\sum_{i=1}^{n} f(y|x_i;\theta)[g(v|x_i) - (1+\eta)\tilde{g}(v)] \leq 0,$$

where $\tilde{g}(v) = \frac{1}{n}\sum_{i=1}^{n} g(v|x_i)$.

Proof. Theorem 2 states that $\mathrm{P}(X, Y, V)$ is η-neutral if Eq (3.4) holds. By substituting $\hat{\mathrm{P}}(X)$ into Eq (3.4), the neutrality condition is rewritten as

$$\forall y, v, \frac{1}{n}\sum_{i=1}^{n} f(y|x_i)[g(v|x_i) - (1+\eta)\tilde{g}(v)] \leq 0.$$

□

For convenience in the following discussion, the neutrality condition is notated as

$$N_\eta(y, v) = \sum_{i=1}^{n} f(y|x_i)[g(v|x_i) - (1+\eta)\tilde{g}(v)] \leq 0. \tag{3.6}$$

3.2.2 Maximum Likelihood Estimation with η-neutrality

Given training samples and a viewpoint prediction model, we performed the maximum likelihood estimation (MLE) with the guarantee of empirical η-neutrality. We wanted a target prediction model that would achieve the maximum log-likelihood for the given samples. At the same time, we wanted a target prediction function that would make $\mathrm{P}(X, Y, V)$ empirically η-neutral with respect to the given samples and viewpoint prediction model. This problem is the following constrained optimization problem:

$$\min_{\theta \in \Theta} L(\theta) \text{ sub to } N_\eta(y, v; \theta) \leq 0, \forall y \in \mathcal{Y}, v \in \mathcal{V}.$$

Existing fairness indices measure fairness with certain statistics, such as differences in the conditional probabilities (Calders et al. 2010) or mutual information (Kamishima et al. 2012b). If such measures are used to guarantee fairness, the fairness of the model is statistically guaranteed for the set of the training samples. In principle, it is desirable to guarantee fairness for individuals contained in the training samples. However, such prediction functions tend to overfit to the given samples and do not provide fairness of unseen samples.

Assuming the model of the viewpoint correctly represents the true distribution, a model that satisfies our η-neutrality condition guarantees statistical independence between every combination of target value y and viewpoint value v. Note that η-neutrality can be realized even when the viewpoint values are not contained in the given samples. This is because the evaluation of fairness is not dependent on the value of the viewpoint but the model of the viewpoint.

3.2.3 Prediction Model for Viewpoints

In principle, we assume $g(V|X)$ accurately represents the true probabilistic distribution $\mathrm{P}(V|X)$, but in reality, this does not always hold. In this subsection, we consider three types of possible viewpoint models.

The first case assumes an extreme example; model $g(V|X)$ is the probabilistic model that outputs random or constant values independent of input X. If we have no knowledge about the viewpoint, we have no choice other than this. Since $g(V|X)$ takes a constant value independent of X, η-neutrality is guaranteed for any $f(Y|X;\theta)$ in this model; however, such prediction is meaningless.

The second case assumes that model $g(V|X)$ is taken as the empirical distribution of the training samples. Existing methods, including CV2NB, statistical parity, and PR, achieve fairness for this empirical distribution. This model realizes empirical fairness with respect to the given training samples, but populational fairness is not guaranteed.

The third case considers the situation that is our focus; model $g(V|X)$ is given as a parametrized probabilistic model. In this case, if $g(V|X)$ accurately represent the true distribution without overfitting, the output of the target prediction model is expected to be fair against not only to the training samples, but also to the unseen samples; this is demonstrated in the following sections by experiments.

The definition of η-neutrality contains all of the above cases, but we specifically consider only the third case, the parametric model.

3.3 Applications of Maximum Likelihood Estimation with η-Neutrality

In this section, we demonstrate two applications of maximum likelihood estimation with a guarantee of empirical η-neutrality: η-neutral logistic regression and η-neutral linear regression.

3.3.1 η-Neutral Logistic Regression

We incorporate our neutrality definition into logistic regression. In logistic regression, the domain of the input variable is $\mathcal{X} = \mathbb{R}^d$, and the domain of the target is binary, $\mathcal{Y} = \{0, 1\}$. Letting $\boldsymbol{\theta} \in \mathbb{R}^d$ be the model parameter, the target prediction model for logistic regression is

$$f(y|\boldsymbol{x};\boldsymbol{\theta}) = \sigma(\boldsymbol{\theta}^T\boldsymbol{x})^y(1-\sigma(\boldsymbol{\theta}^T\boldsymbol{x}))^{1-y}, \tag{3.7}$$

where $\sigma(a)$ is the logistic sigmoid function.

Letting Eq (3.7) be the target prediction model, the log-likelihood is given by Eq (3.5), and then the problem of η-neutral logistic regression is

$$\min L(\boldsymbol{\theta}) \text{ sub to } N_\eta(y, v; \boldsymbol{\theta}) \leq 0, \ \forall v, y.$$

Note that the viewpoint prediction model $g(v|\boldsymbol{x})$ can be any probabilistic model.

We consider the optimization of η-neutral logistic regression. The gradient and Hessian matrix of $L(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ are, respectively,

$$\begin{aligned}\nabla L(\boldsymbol{\theta}) &= \sum_{i=1}^{N} \big(\sigma(\boldsymbol{\theta}^T \boldsymbol{x}_i) - y_i\big)\boldsymbol{x}_i, \\ \nabla^2 L(\boldsymbol{\theta}) &= \sum_{i=1}^{N} \sigma(\boldsymbol{\theta}^T \boldsymbol{x}_i)(1 - \sigma(\boldsymbol{\theta}^T \boldsymbol{x}_i))\boldsymbol{x}_i \boldsymbol{x}_i^T.\end{aligned}$$

Due to the nature of the logistic sigmoid function, the Hessian matrix is positive semidefinite. Hence, the log-likelihood function is convex.

Next, we examine the convexity of the constraints associated with the η-neutrality condition. Since $N_\eta(y, v; \boldsymbol{\theta})$ is a linear combination of f, the convexity of f is investigated. The gradient of f with respect to the parameter $\boldsymbol{\theta}$ is

$$\begin{aligned}\nabla f(y, \boldsymbol{x}; \boldsymbol{\theta}) &= \nabla \exp\left(\ln f(y|\boldsymbol{x}; \boldsymbol{\theta})\right) \\ &= \big(y - \sigma(\boldsymbol{\theta}^T \boldsymbol{x})\big)\, f(y|\boldsymbol{x}; \boldsymbol{\theta})\boldsymbol{x}.\end{aligned}$$

The Hessian is similarly obtained as

$$\nabla^2 f(y|\boldsymbol{x}; \boldsymbol{\theta}) = \alpha(\boldsymbol{x}, y, \boldsymbol{\theta}) f(y|\boldsymbol{x}; \boldsymbol{\theta})\boldsymbol{x}\boldsymbol{x}^T,$$

where $\alpha(\boldsymbol{x}, y, \boldsymbol{\theta}) = 2\sigma(\boldsymbol{\theta}^T \boldsymbol{x})^2 + y^2 - (2y + 1)\sigma(\boldsymbol{\theta}^T \boldsymbol{x})$. Since $\alpha(\boldsymbol{x}, y, \boldsymbol{\theta}) \in \mathbb{R}$ can be negative, the Hessian is not positive definite, and f is nonconvex with respect to $\boldsymbol{\theta}$. Thus, unfortunately, the neutrality condition in logistic regression is nonconvex, regardless of the choice of $g(v|\boldsymbol{x})$.

In our experiments with η-neutral logistic regression, we used the nonlinear optimization package, Ipopt, that provides the implementation of the primal-dual interior point method (Wächter et al. 2006). As the initial point of the primal-dual interior point method to solve the optimization problem of η-neutral logistic regression, we use the optimal point of non-fair logistic regression. Although the constraint is nonconvex, we show by experiments that η-neutrality can be achieved without sacrificing too much of the accuracy of the prediction in Section 3.5. This nonconvexity arises in part from the nonconvexity of the probability distribution. Further research on convexifying the neutrality constraint is left as an area of future work.

3.3.2 η-Neutral Linear Regression

We now consider η-neutral linear regression and demonstrate that maximum likelihood estimation with η-neutrality can work with a continuous viewpoint. In linear regression, the domain of the target is $\mathcal{Y} = \mathbb{R}$, and the input domain is $\mathcal{X} = \mathbb{R}^d$. The target prediction function is given by

$$f(y|\boldsymbol{x}; \boldsymbol{w}, \beta) = \frac{\beta}{\sqrt{2\pi}} \exp\left[-\frac{\beta(\boldsymbol{w}^T\boldsymbol{x} - y)^2}{2}\right],$$

where $\boldsymbol{w}$ denotes the regression coefficient for the target and β denotes the parameter representing the inversed variance of the prediction error of the target. The linear regression problem is solved by minimization of the negative log-likelihood, as given by Eq (3.5).

The domain of the viewpoint is $\mathcal{V} = \mathbb{R}$. As well as the target prediction model, we assume the viewpoint prediction model is

$$g(v|\boldsymbol{x}; \boldsymbol{w}_v, \beta_v) = \frac{\beta_v}{\sqrt{2\pi}} \exp\left[-\frac{\beta(\boldsymbol{w}_v^T\boldsymbol{x} - v)^2}{2}\right],$$

where $\boldsymbol{w}_v$ denotes the regression coefficient for the viewpoint and β_v denotes the parameter representing the inversed variance of the prediction error of the viewpoint.

Predictions of the target Y and the viewpoint V are obtained, respectively, by

$$\hat{y} = \arg\max_y f(y|\boldsymbol{x}; \boldsymbol{w}, \beta), \quad \hat{v} = \arg\max_v g(v|\boldsymbol{x}; \boldsymbol{w}_v, \beta_v).$$

Then, η-neutral linear regression is formulated as an optimization problem with the same constraints as in Eq (3.6):

$$\begin{aligned} &\min \frac{1}{2}\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}^T\boldsymbol{X}\boldsymbol{w} \\ &\text{sub to } \max_{i\in[n]}\left\{N_\eta(\boldsymbol{w}^T\boldsymbol{x}, \boldsymbol{w}_v^T\boldsymbol{x}; \boldsymbol{w}, \beta)\right\} \leq 0, \end{aligned}$$

where $\boldsymbol{X} = (\boldsymbol{x}_1^T, \boldsymbol{x}_2^T, ..., \boldsymbol{x}_n^T)^T$ is the design matrix and $\boldsymbol{y} = (y_1, y_2, ..., y_n)^T$ is the vector of the target values.

As in the case with η-neutral logistic regression, we investigate the convexity of the neutrality constraint given models f and g by investigating the convexity

of f. The gradient and Hessian matrix of f are, respectively,

$$\begin{aligned}
&\nabla_{\boldsymbol{w}} f(\boldsymbol{w}^T\boldsymbol{x}'|\boldsymbol{x};\boldsymbol{w},\beta)\\
=&\nabla_{\boldsymbol{w}} \exp(\ln f(\boldsymbol{w}^T\boldsymbol{x}'|\boldsymbol{x};\boldsymbol{w},\beta))\\
=&-\beta(\boldsymbol{w}^T\boldsymbol{x}-\boldsymbol{w}^T\boldsymbol{x}')f(\boldsymbol{w}^T\boldsymbol{x}'|\boldsymbol{x};\boldsymbol{w},\beta)(\boldsymbol{x}-\boldsymbol{x}'),\\
&\nabla^2_{\boldsymbol{w}} f(\boldsymbol{w}^T\boldsymbol{x}'|\boldsymbol{x};\boldsymbol{w},\beta)\\
=&\alpha(\boldsymbol{x},\boldsymbol{x}',\boldsymbol{w},\beta)\beta f(\boldsymbol{w}^T\boldsymbol{x}'|\boldsymbol{x};\boldsymbol{w},\beta)(\boldsymbol{x}-\boldsymbol{x}')(\boldsymbol{x}-\boldsymbol{x}')^T,
\end{aligned}$$

where $\alpha(\boldsymbol{x},\boldsymbol{x}',\boldsymbol{w},\beta) = \beta(\boldsymbol{w}^T\boldsymbol{x}-\boldsymbol{w}^T\boldsymbol{x}')^2 - 1$. Since, depending on $\boldsymbol{w}$, $f(\boldsymbol{w}^T\boldsymbol{x}'|\boldsymbol{x};\boldsymbol{w},\beta) \geq 0$ and $\alpha(\boldsymbol{x},\boldsymbol{x}',\boldsymbol{w},\beta) \in \mathbb{R}$ can take negative values, the Hessian is not positive definite. Hence, unfortunately, f is not convex with respect to $\boldsymbol{w}$. For this non-convex constraint optimization, we again use the primal-dual interior point method of Ipopt in our experiments (Wächter et al. 2006).

3.4 Comparison of Fairness Measures

One of the largest difference between η-neutrality and the CV score (Calders et al. 2010) or statistical parity (Dwork et al. 2012) or the prejudice index (Kamishima et al. 2012b) is its situation; while fairness of η-neutrality is defined for the viewpoint prediction model, that of others is defined for the viewpoint value. To discuss the difference of these fairness measures, we assume the samples D_n contain the viewpoint values in the subsequent subsections.

The CV score and the prejudice index are defined as quantities that measure fairness, whereas η-neutrality and statistical parity are defined as conditions required for the prediction model to be fair. More precisely, for example, the prediction model is said to be η-neutral only if the ratio of the particular probabilities is upper bounded by η for all y and all v. We employ the upper bound of η-neutrality and statistical parity as the fairness measure for η-neutrality and statistical parity, respectively.

For a comprehensive discussion of the comparison of the fairness measures, we define the *neutrality factor* to universally represent all the fairness measures. The neutrality factor quantifies unfairness of a specific pair of the target value y and viewpoint value v:

Definition 3 (Neutrality factor)**.** Let X, Y, and V be random variables representing the input, the target, and the viewpoint, respectively. Then, the

neutrality factor with respect to a target $y \in \mathcal{Y}$ and a viewpoint $v \in \mathcal{V}$ is defined by

$$\nu(y, v) = \frac{\hat{\mathrm{P}}(y, v)}{\hat{\mathrm{P}}(y)\hat{\mathrm{P}}(v)}.$$

From the definition of η-neutrality, we can say that η-neutrality evaluates the maximum value of the neutrality factors with respect to y and v. In subsequent subsections, we represent the CV score, statistical parity and the prejudice index by using the neutrality factor. Furthermore, we clarify the relationship between these fairness measures.

3.4.1 Comparison of η-neutrality, CV score and Statistical Parity

In this subsection, we first show that the CV score is a variant measure of statistical parity. Then, we derive the relation between η-neutrality and statistical parity. In what follows, we assume Y is the discrete target and V is the binary viewpoint.

Let $y \in \{y_+, y_-\}$ and $v \in \{v_+, v_-\}$ be the binary target and the binary viewpoint, respectively. The CV score (Calders et al. 2010) with respect to the given samples D_n is defined by the difference of the conditional probabilities:

$$\mathrm{CV}(D_n) = \hat{\mathrm{P}}(y_+|v_+) - \hat{\mathrm{P}}(y_+|v_-), \tag{3.8}$$

where $\hat{\mathrm{P}}(Y|V)$ is empirically evaluated with the given samples D_n. We can assume $\hat{\mathrm{P}}(y_+|v_+) \geq \hat{\mathrm{P}}(y_+|v_-)$ without loss of generality. If the CV score equals to zero, the classification is empirically fair with respect to the given samples D_n.

Statistical parity (Dwork et al. 2012) employs the total variation distance to quantify fairness. The total variation distance of the two probabilistic distributions of target y, $P(y)$ and $Q(y)$, is defined as

$$\mathrm{TV}(P, Q) = \frac{1}{2}\sum_{y \in \mathcal{Y}} |P(y) - Q(y)|. \tag{3.9}$$

Given $\epsilon \geq 0$ as a neutrality parameter, we say ϵ-statistical parity holds with respect to the given samples D_n if

$$\mathrm{TV}(\hat{\mathrm{P}}(Y|v_+), \hat{\mathrm{P}}(Y|v_-)) \leq \epsilon.$$

First, we show that the CV score is a variant measure of statistical parity.

Lemma 1. *Let Y and V be the binary target and the binary viewpoint, respectively. For any $\epsilon \geq 0$ and any samples D_n, $CV(D_n) \leq \epsilon$ if and only if ϵ-statistical parity with respect to D_n holds.*

Proof. By the definition of statistical parity, if ϵ-statistical parity holds

$$\mathrm{TV}(\hat{\mathrm{P}}(Y|v_+), \hat{\mathrm{P}}(Y|v_-)) \leq \epsilon.$$

By the definition of the probability, $\hat{\mathrm{P}}(y_+|v) + \hat{\mathrm{P}}(y_-|v) = 1 \; \forall v \in \mathcal{V}$ and we have

$$\begin{aligned} &\hat{\mathrm{P}}(y_+|v_+) - \hat{\mathrm{P}}(y_+|v_-) \\ =&(1 - \hat{\mathrm{P}}(y_-|v_+)) - (1 - \hat{\mathrm{P}}(y_-|v_-)) \\ =&\hat{\mathrm{P}}(y_-|v_-) - \hat{\mathrm{P}}(y_-|v_+). \end{aligned} \tag{3.10}$$

By substituting Eq (3.10) into Eq (3.8), we have

$$\begin{aligned} &CV(D_n) \\ =&\hat{\mathrm{P}}(y_+|v_+) - \hat{\mathrm{P}}(y_+|v_-) \\ =&\frac{1}{2}(\hat{\mathrm{P}}(y_+|v_+) - \hat{\mathrm{P}}(y_+|v_-) + \hat{\mathrm{P}}(y_-|v_-) - \hat{\mathrm{P}}(y_-|v_+)) \end{aligned}$$

We can assume $\hat{\mathrm{P}}(y_+|v_+) - \hat{\mathrm{P}}(y_+|v_-) \geq 0$ and $\hat{\mathrm{P}}(y_-|v_-) - \hat{\mathrm{P}}(y_-|v_+) \geq 0$ without loss of generality. Hence, we have

$$\begin{aligned} CV(D_n) =&\frac{1}{2}\sum_{y \in \mathcal{Y}} |\hat{\mathrm{P}}(y|v_+) - \hat{\mathrm{P}}(y|v_-)| \\ =&\mathrm{TV}(\hat{\mathrm{P}}(Y|v_+), \hat{\mathrm{P}}(Y|v_-)). \end{aligned}$$

□

As proved by Lemma 1, the statistical parity with the binary target can be interpreted as the CV score.

Next, we provide the relation between η-neutrality and statistical parity. The following theorem shows that η-statistical parity with respect to the given samples D_n holds if η-neutrality holds.

Theorem 3. *Let X and Y be the input variable and the discrete target random variable, respectively. Let V denote the binary viewpoint random variable. If the probability $\mathrm{P}(X, Y, V)$ is empirically η-neutral, then Y is η-statistical parity with respect to V.*

In order to prove Theorem 3, we use the following lemma that shows another representation of the total variation distance in statistical parity by using the neutrality factors.

Lemma 2. *Let* $\mathrm{TV}(P, Q)$ *be the total variation distance between* P *and* Q *defined Eq* (3.9)*. Then,*

$$\mathrm{TV}(\hat{\mathrm{P}}(Y|v_+), \hat{\mathrm{P}}(Y|v_-)) = \mathbf{E}\left[\max_{v\in\{v_+,v_-\}} \nu(Y, v)\right] - 1.$$

The proof of Lemma 2 is shown in the Section 3.A. As proved by Lemma 2, statistical parity is the expectation of the maximum value with respect to v of the neutrality factors. By using Lemma 2, we prove Theorem 3.

Proof of Theorem 3. If η-neutrality holds, $\nu(y, v) \leq 1 + \eta \ \forall y \in \mathcal{Y}, v \in \mathcal{V}$. Then, we have

$$\mathrm{TV}(\hat{\mathrm{P}}(Y|v_+), \hat{\mathrm{P}}(Y|v_-)) \leq \mathbf{E}_Y[1 + \eta] - 1 \leq \eta.$$

□

As proved by the Theorem 3, statistical parity holds if η-neutrality holds. We can immediately show that the CV score is bounded by a certain function of η if η-neutrality holds by using Theorem 3 and Lemma 1.

3.4.2 Comparison of η-neutrality and Prejudice Index

In this subsection, we compare our η-neutrality with the prejudice index (Kamishima et al. 2012b). The prejudice index is defined as the mutual information of the target Y and the viewpoint V:

$$\mathrm{P}I = I(Y; V) = \mathbf{E}[\ln \nu(Y, V)],$$

where $I(Y; V)$ is the mutual information of the target Y and the viewpoint V.

While the prejudice index is the expectation of the logarithm of the neutrality factors $\nu(y, v)$, the parameter η of η-neutrality denotes the upper bound of the neutrality factor $\nu(y, v)$. This indicates that prejudice index can be upper bounded with the parameter η if η-neutrality holds. Following proposition provides this indication.

Table 3.2: Summary of fairness measures

Aggregation	$\nu(y, v)$	$\ln \nu(y, v)$
Maximum w.r.t y and v	η-neutrality	equivalent to η-neutrality
Maximum w.r.t v and expectation w.r.t. y	CV-score, statistical parity	-
Expectation w.r.t y and v	-	prejudice index

Proposition 1. *Let X and Y be random variables representing the input and the target, respectively. Let V denote a random variable representing the viewpoint. If the probability $\mathrm{P}(X, Y, V)$ is empirically η-neutral with respect to given D_n with $\eta \geq 0$, then*

$$I(V; Y) \leq \ln(1 + \eta).$$

Proof. From empirical η-neutrality of the probability $\mathrm{P}(X, Y, V)$, we have

$$\forall v \in \mathcal{V}, y \in \mathcal{Y}, \quad \frac{\hat{\mathrm{P}}(v, y)}{\hat{\mathrm{P}}(v)\hat{\mathrm{P}}(y)} \leq 1 + \eta.$$

Since natural logarithm is a monotonically increasing function, we have

$$\forall v \in \mathcal{V}, y \in \mathcal{Y}, \quad \ln \frac{\hat{\mathrm{P}}(v, y)}{\hat{\mathrm{P}}(v)\hat{\mathrm{P}}(y)} \leq \ln(1 + \eta). \tag{3.11}$$

Expectation of Eq (3.11) with respect to Y and V derives as follows:

$$\mathbf{E}\left[\ln \frac{\hat{\mathrm{P}}(V, Y)}{\hat{\mathrm{P}}(V)\hat{\mathrm{P}}(Y)}\right] = I(v; y) \leq \ln(1 + \eta).$$

□

As proved by the Proposition 1, the prejudice index is upper bounded by $\ln(1 + \eta)$ if η-neutrality holds.

3.4.3 Summary of Comparisons

Table 3.2 shows the summary of the fairness measures. By definition of η-neutrality, η-neutrality is the maximum value of the neutrality factors. Due to monotonicity of the logarithm function, η-neutrality is equivalent to the

maximum value of the logarithm of the neutrality factors (Table 3.2, line 1). Statistical parity can be represented as the expectation of the maximum value with respect to v of the neutrality factors as indicated by Lemma 2. Similarly, as indicated by Lemma 1, the CV score is equivalent to statistical parity with binary targets (Table 3.2, left of line 2). The prejudice index is defined as the expectation of logarithm of the neutrality factor (Table 3.2, left of line 2). As indicated by Theorem 3, statistical parity and the CV score can be upper bounded by η-neutrality. Moreover, as indicated by Proposition 1, the prejudice index can be upper bounded by η-neutrality. As shown in Table 3.2, all of the fairness measures can be represented with the neutrality factors. The difference of these fairness measures is only in the way of aggregation.

The prejudice index is defined by the mutual information which represents the statistical dependency between the target and the viewpoint. Thus, the fairness measures are closely connected to the measures of statistical dependency (Suzuki et al. 2009; Torkkola 2003).

3.5 Experiments

3.5.1 Classification

Settings. To examine and compare the classification performance and the fairness performance of η-neutral logistic regression with other methods, we performed experiments on five real data sets specified in Tables 3.3 and 3.4. In Table 3.3, #Inst. and #Attr. denote the number of samples and the number of the attributes, respectively; "Viewpoint" and "Target" denote the attribute used as the target and the viewpoint, respectively. Table 3.4 also shows the number of samples with the target ($\#y_+$) and the viewpoint ($\#v_+$). Also, the table shows the prediction accuracy of the non-fair logistic regression with respect to the target ($\mathrm{Acc}(y)$) and the viewpoint ($\mathrm{Acc}(v)$).

We compared the following methods: logistic regression (LR, no fairness guarantee), logistic regression that learns without using the values of viewpoint (LRns), the Naïve Bayes classifier (NB, no fairness guarantee), the Naïve Bayes classifier that learns without the values of viewpoint (NBns), CV2NB (Calders et al. 2010), logistic regression that uses the PR (Kamishima et al. 2012a), and η-neutral logistic regression (ηLR, proposal). In the PR method, the regularizer parameter λ, which balances the loss minimization and fairness, was varied as $\lambda \in \{0, 5, 10, 15, 20, 30\}$. The parameter η, which determines the degree of fairness, was varied as $\eta \in \{0.00, 0.01, ..., 0.40\}$. All dataset

attributes were discretized by the same procedure described in (Calders et al. 2010) and coded by 1-of-K representation for LR, LRns, PR and ηLR.

As fairness indices of prediction models, normalized prejudice index (NPI) and $\hat{\eta}$ are introduced. NPI is defined as the normalized mutual information of Y and V, normalized by the entropy of Y and V (Kamishima et al. 2012b):

$$\mathrm{N}PI = \frac{I(Y;V)}{\sqrt{H(Y)H(V)}},$$

where $I(Y;V)$ is the mutual information of target Y and viewpoint V, $I(Y;V)/H(Y)$ is the ratio of information of V used for predicting Y; $I(Y;V)/H(V)$ is the ratio of information that is exposed if a value of Y is known. Thus NPI can be interpreted as the geometrical mean of these two ratios. The range of NPI is $[0, 1]$.

The fairness measure $\hat{\eta}$ is defined as

$$\hat{\eta} = \max_{y\in\mathcal{Y},v\in\mathcal{V}} \frac{\hat{\mathrm{P}}(v,y)}{\hat{\mathrm{P}}(v)\hat{\mathrm{P}}(y)} - 1,$$

where $\hat{\eta}$ can be interpreted as the degree of the dependency of y and v with which the largest dependency occurs. If Y and V are mutually independent, $\hat{\eta} = 0$. If the fairness measure with respect to a target prediction model is $\hat{\eta}$, it means the model Eq (3.1) is empirically $\hat{\eta}$-neutral with respect to the given samples.

We compared the three measures: accuracy, normalized prejudice index (NPI), and $\hat{\eta}$ of η-neutrality. These indices were evaluated with five-fold cross validation and the average values of ten different folds are shown in the plots.

The values used for the learning of $f(y|x)$, the guarantee of fairness, and the measurement of fairness are summarized in Table 3.5. For the guarantee of fairness, we consider the following two cases.

Case 1 assumes that the values of the viewpoint are provided in the samples. In this case, our method uses the viewpoint model learned from the samples, whereas other methods uses the actual viewpoint values.

Case 2 assumes that the values of the viewpoint are not provided. Instead, the viewpoint model, $g(v|\boldsymbol{x})$, is provided. In this case, our method learns the model of the target using the given viewpoint model g. Other methods need the values of the viewpoint, so these are estimated as $\hat{v} = \arg\max_v g(v|\boldsymbol{x})$. Other methods then learn the model of the target with $(x, \hat{v})$.

As a measurement of fairness, all methods used the true viewpoint value v in both cases.

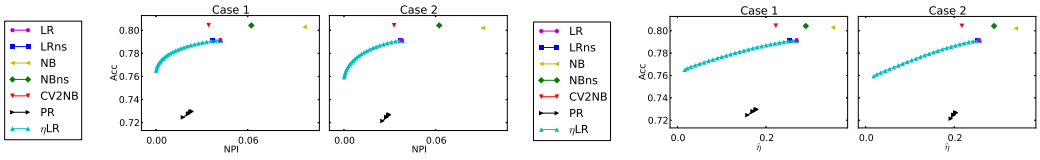
Results. Figure 3.2 shows the experimental results. In the graphs, the best result is at the left top. Comparing the results of NB and NBns in Adult, Dutch Census, and Bank Marketing, we can see that the improvement of fairness by an elimination of the viewpoint is limited in both cases. The same applies to LR and LRns.

In both cases, CV2NB achieves better fairness than NBns regarding both NPI and $\hat{\eta}$ in Adult and Dutch Census. Besides, the decrease in the accuracy of the prediction is less than 1% in the Adult dataset and 5% in the Dutch Census. On the other hand, CV2NB fails to achieve a fair target model in Bank Marketing and German Credit Data, because the fair level of CV2NB is worse than NBns. As shown in Table 3.4, the number of the positive viewpoint of these datasets is fewer than the negative viewpoint values, in comparison with the other datasets. The degradation of performance of CV2NB in Bank Marketing and German Credit Data can be caused by such imbalanced viewpoint labels.

In both cases, PR successfully balances the NPI or the $\hat{\eta}$ and the accuracy for Adult and Dutch Census datasets, but dominated by ηLR. To ensure fairness of the target prediction model, PR adds non-convex NPI term to the objective function. Due to the non-convexity of the objective function, both the accuracy and the fair level of the prediction model can be worsened.

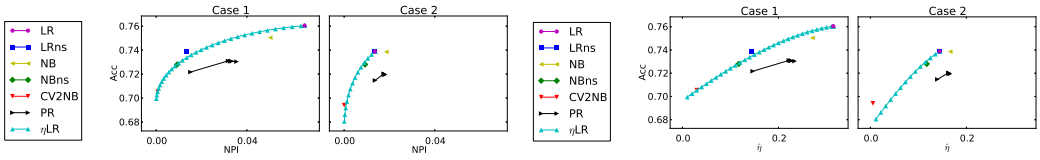
In both cases, our η-neutral logistic regression successfully balances fairness and accuracy of the predication by changing η in Adult, Dutch Census and Bank Marketing. Particularly, in Bank Marketing, even though the fair level of the other methods is almost the same as its baseline (LR or NB), our ηLR can achieve the prediction model with a high fair level. Furthermore, the decrease in the accuracy of the prediction was at most 5% in these datasets, even with small η. Thus, ηLR empirically works well even if its constraints are non-convex. In contrast, we can see from the results of PR that the non-convex objective function can worsen both of the fair level and the accuracy. Hence, although fairness measure is usually non-convex, we can achieve small unfair level by employing the fairness measure as constraints.

In German Credit Data, the fair level of ηLR is lower than LRns in both cases. It is noteworthy that the fair level of ηLR is even lower than LR in Case 2. This was again due to imbalanced viewpoint labels of the dataset. The given model of the viewpoint is trained so that it ignores minor viewpoint label. Hence, due to the overfitting of the model learned by ηLR with such model



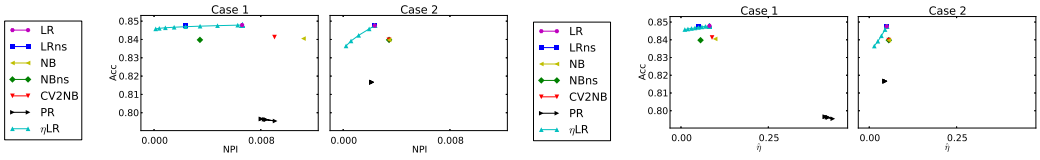
(a) Adult, NPI

(b) Adult, $\hat{\eta}$



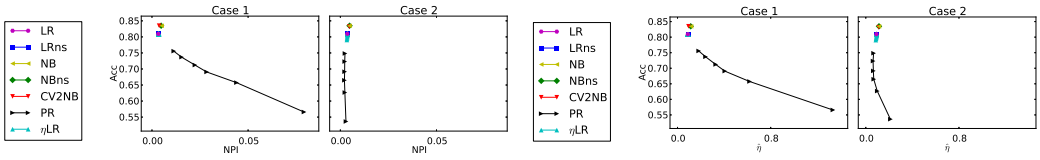
(c) Dutch Census, NPI

(d) Dutch Census, $\hat{\eta}$



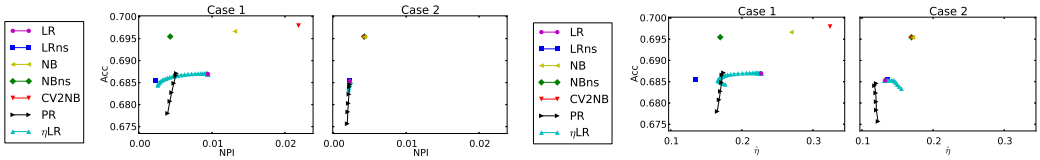
(e) Bank Marketing, NPI

(f) Bank Marketing, $\hat{\eta}$



(g) Credit Approval, NPI

(h) Credit Approval, $\hat{\eta}$



(i) German Credit Data, NPI

(j) German Credit Data, $\hat{\eta}$

Figure 3.2: Accuracy vs. fairness measure. Each subplot displays the result of Case 1 (left) and the result of Case 2 (right) corresponding to the datasets and the fairness measure ($\hat{\eta}$ or NPI).

of the viewpoint to the major viewpoint label, the fair level of ηLR may be lower than LR.

In Credit Approval, all technique did not work well in both cases. From Table 3.4, the number of the samples of the dataset is up to seven hundred. This result can indicate that estimation of the fairness measures for test dataset needs a sufficiently large number of the samples.

3.5.2 Regression

Settings. To investigate the behaviors of the fair learning algorithm in linear regression, we performed experiments of η-neutral linear regression on three real datasets specified in Table 3.6. As with the specification of the dataset for the classification, the table shows #Inst., #Attr., "Viewpoint" and "Target". Also, the table also provides "Corr", the correlation coefficient between the target and the viewpoint. We chose the viewpoints for each dataset as the attribute of which the correlation coefficient with respect to the target maximizes. All the attributes, the target, and the viewpoint were scaled into the range $[-1, 1]$. Letting the regression parameters of the target f and viewpoint g be $\boldsymbol{w}$ and $\boldsymbol{w}_v$, respectively, the predicted values were $\hat{y} = \boldsymbol{w}^T\boldsymbol{x}$ and $\hat{v} = \boldsymbol{w}_v^T\boldsymbol{x}$. The parameter η was varied as $\eta \in \{2^{-12}, 2^{-11}, ..., 2^2\}$. The accuracy of the prediction was measured by root-mean-square error (RMSE); $\hat{\eta}$ and the correlation coefficient between the target and the viewpoint were used as the measure of fairness.

Results. Figure 3.3 shows RMSE and the absolute value of the correlation coefficient between the predicted target value $\hat{y}$ and the viewpoint value v corresponding to the parameter η. For all datasets, the plots explicitly show that the correlation coefficient becomes lower and RMSE becomes higher as η decreases. These results show that our η-neutral linear regression with low η can obtain the fair regression model in the sense of the correlation coefficient. Furthermore, these results indicate that our η-neutral linear regression can use η to successfully control the fair level of the regression model.

Figure 3.4 shows the scatter plots of $(\hat{y}, y)$ (the top row) and $(\hat{y}, \hat{v})$ (the bottom row) with $\eta \in \{2^{-9}, 2^{-6}, 2^{-3}\}$ on the Housing dataset. From left to right, the parameter η was varied as $\eta \in \{2^{-9}, 2^{-6}, 2^{-3}\}$. The right-most figures show the results of the most unfair linear regression. The $(\hat{y}, \hat{v})$ plot represents the prediction accuracy of the regression model. When the model achieves a better RMSE, the points in the $(\hat{y}, y)$ plot concentrate more along the diagonal line. At the same time, the $(\hat{y}, \hat{v})$ plot represents fairness. If the fairness is low, the

correlation between $\hat{y}$ and $\hat{v}$ appears in the $(\hat{y}, \hat{v})$ plot.

In Figure 3.4 (h), a strong negative correlation between $\hat{y}$ and $\hat{v}$ can be found, and thus this regression model has a low fairness. In Figure 3.4, the level of fairness increases from right to left. The plots show that the dependency of $\hat{y}$ on $\hat{v}$ becomes weaker as η decreases. In Figure 3.4 (e), we can see that outputs of the target regression model is not changed by a change of the input; such regression is useless even if the model is fair. Thus, selection of η is important to obtain a fair regression model with high accuracy.

3.6 Conclusion

In this chapter, we propose a framework using the maximum likelihood estimation (MLE) for learning probabilistic models, with which the disparate impact coming from the hidden viewpoint can be removed. There are two key points in which our proposal is different from existing methods.

First, our method guarantees fairness of the target prediction model with respect to a given viewpoint prediction model. Due to this model-based fairness, our method allows to remove the disparate impact from target prediction models with respect to viewpoints arbitrarily defined by users, as long as the viewpoint prediction model is provided in the form of a probabilistic distribution.

Second, our fairness measure, η-neutrality, is based on the principle that the model should guarantee fairness with respect to every combination of target and viewpoint value that appears in the dataset.

To clarify the relationship between the fairness measures, we define the neutrality factor. Then, we showed that all of the fairness measures are represented by the aggregation of the neutrality factors. We also show that η-neutrality can upper bound all of the other fairness measures.

Experimental results show that our method with model-based fairness successes to achieve a fair model even when only a model of the viewpoint is provided. Besides, it balances accuracy and fairness of the target prediction. As discussed in Section 3.3.1 and Section 3.3.2, likelihood maximization with the η-neutrality constraint is nonconvex optimization; this is due the nonconvexity of the constraint function. In the next chapter, we intend to find a way to convexify the constraints induced by the fairness condition.

The privacy problem is strongly related to the fairness problem. The differ-

ence between the privacy problem and the fairness problem causes from the treatments of the sensitive information. The sensitive information in the privacy problem is individuals' information that they want not to be published. On the other hand, the sensitive information in the fairness problem, which is equivalent to the viewpoint, is individuals' information that they want not to be made decisions depending on this information. Following the treatment of the sensitive information, we can define the adversaries in the privacy problem as entities that can predict the sensitive information. The adversaries in the fairness problem can be defined in the same manner as entities that make decisions depending on the sensitive information.

Appendix 3.A Proof of Lemma 2

Proof. Let $\mathcal{Y}^+ = \{y \in \mathcal{Y} | \hat{\mathrm{P}}(y|v_+) \geq \hat{\mathrm{P}}(y|v_-)\}$ and let $\mathcal{Y}^- = \{y \in \mathcal{Y} | \hat{\mathrm{P}}(y|v_+) \leq \hat{\mathrm{P}}(y|v_-)\}$. Then, we have

$$
\begin{aligned}
&\mathrm{TV}(\hat{\mathrm{P}}(Y|v_+), \hat{\mathrm{P}}(Y|v_-)) \\
=&\frac{1}{2}\sum_{y\in\mathcal{Y}} |\hat{\mathrm{P}}(y|v_+) - \hat{\mathrm{P}}(y|v_-)| \\
=&\frac{1}{2}[\sum_{y\in\mathcal{Y}^+} (\hat{\mathrm{P}}(y|v_+) - \hat{\mathrm{P}}(y|v_-)) \\
&\quad + \sum_{y\in\mathcal{Y}^-} (\hat{\mathrm{P}}(y|v_-) - \hat{\mathrm{P}}(y|v_+))] \\
=&\frac{1}{2}[\sum_{y\in\mathcal{Y}^+} (\hat{\mathrm{P}}(y|v_+) - \hat{\mathrm{P}}(y|v_-)) \\
&\quad + \sum_{y\in\mathcal{Y}^+} (\hat{\mathrm{P}}(y|v_+) - \hat{\mathrm{P}}(y|v_-))] \\
&(\because \sum_{y\in\mathcal{Y}^-} \hat{\mathrm{P}}(y|v) = 1 - \sum_{y\in\mathcal{Y}^+} \hat{\mathrm{P}}(y|v)\ \forall v \in \mathcal{V}) \\
=&\sum_{y\in\mathcal{Y}^+} (\hat{\mathrm{P}}(y|v_+) - \hat{\mathrm{P}}(y|v_-)).
\end{aligned} \tag{3.12}
$$

Similarly, we obtain

$$
D_{tv}(\hat{\mathrm{P}}(Y|v_+), \hat{\mathrm{P}}(Y|v_-)) = \sum_{y\in\mathcal{Y}^-} (\hat{\mathrm{P}}(y|v_-) - \hat{\mathrm{P}}(y|v_+)). \tag{3.13}
$$

Combining Eq (3.12) and Eq (3.13) and with the fact that $\hat{\mathrm{P}}(v_+)+\hat{\mathrm{P}}(v_-)=1$, we have

$$
\begin{aligned}
& D_{tv}(\hat{\mathrm{P}}(Y|v_+),\hat{\mathrm{P}}(Y|v_-)) \\
=& \hat{\mathrm{P}}(v_-)\sum_{y\in\mathcal{Y}^+}(\hat{\mathrm{P}}(y|v_+)-\hat{\mathrm{P}}(y|v_-)) \\
& \quad +\hat{\mathrm{P}}(v_+)\sum_{y\in\mathcal{Y}^-}(\hat{\mathrm{P}}(y|v_-)-\hat{\mathrm{P}}(y|v_+)) \\
=& \sum_{y\in\mathcal{Y}^+}(\hat{\mathrm{P}}(v_-)\hat{\mathrm{P}}(y|v_+)-\hat{\mathrm{P}}(y,v_-)) \\
& \quad +\sum_{y\in\mathcal{Y}^-}(\hat{\mathrm{P}}(v_+)\hat{\mathrm{P}}(y|v_-)-\hat{\mathrm{P}}(y,v_+)) \\
=& \sum_{y\in\mathcal{Y}^+}((1-\hat{\mathrm{P}}(v_+))\hat{\mathrm{P}}(y|v_+)-\hat{\mathrm{P}}(y,v_-)) \\
& \quad +\sum_{y\in\mathcal{Y}^-}((1-\hat{\mathrm{P}}(v_-))\hat{\mathrm{P}}(y|v_-)-\hat{\mathrm{P}}(y,v_+)) \\
=& \sum_{y\in\mathcal{Y}^+}(\hat{\mathrm{P}}(y|v_+)-\hat{\mathrm{P}}(y))+\sum_{y\in\mathcal{Y}^-}(\hat{\mathrm{P}}(y|v_-)-\hat{\mathrm{P}}(y)) \\
& (\because \hat{\mathrm{P}}(y,v_+)+\hat{\mathrm{P}}(y,v_-)=\hat{\mathrm{P}}(y)\ \forall y\in\mathcal{Y}) \\
=& \sum_{y\in\mathcal{Y}^+}\hat{\mathrm{P}}(y)\frac{\hat{\mathrm{P}}(y|v_+)}{\hat{\mathrm{P}}(y)}+\sum_{y\in\mathcal{Y}^-}\hat{\mathrm{P}}(y)\frac{\hat{\mathrm{P}}(y|v_-)}{\hat{\mathrm{P}}(y)}-1. \qquad (3.14)
\end{aligned}
$$

From definition of $\mathcal{Y}^+$ and $\mathcal{Y}^-$,

$$
\hat{\mathrm{P}}(y|v_+)=\max\{\hat{\mathrm{P}}(y|v_+),\hat{\mathrm{P}}(y|v_-)\} \quad \text{if } y\in\mathcal{Y}^+, \text{ and} \qquad (3.15)
$$

$$
\hat{\mathrm{P}}(y|v_-)=\max\{\hat{\mathrm{P}}(y|v_+),\hat{\mathrm{P}}(y|v_-)\} \quad \text{if } y\in\mathcal{Y}^- \qquad (3.16)
$$

hold. By substituting Eqs. 3.15 and 3.16 into Eq. 3.14, we have

$$
\begin{aligned}
&\mathrm{TV}(\hat{\mathrm{P}}(Y|v_+), \hat{\mathrm{P}}(Y|v_-)) \\
=&\sum_{y\in\mathcal{Y}} \hat{\mathrm{P}}(y) \frac{\max\{\hat{\mathrm{P}}(y|v_+), \hat{\mathrm{P}}(y|v_-)\}}{\hat{\mathrm{P}}(y)} - 1 \\
=&\mathbf{E}_Y\left[\frac{\max\{\hat{\mathrm{P}}(y|v_+), \hat{\mathrm{P}}(y|v_-)\}}{\hat{\mathrm{P}}(y)}\right] - 1 \\
=&\mathbf{E}_Y\left[\max\left\{\frac{\hat{\mathrm{P}}(y, v_+)}{\hat{\mathrm{P}}(y)\hat{\mathrm{P}}(v_+)}, \frac{\hat{\mathrm{P}}(y, v_-)}{\hat{\mathrm{P}}(y)\hat{\mathrm{P}}(v_-)}\right\}\right] - 1 \\
=&\mathbf{E}_Y\left[\max_{v\in\{v_+, v_-\}} \frac{\hat{\mathrm{P}}(y, v)}{\hat{\mathrm{P}}(y)\hat{\mathrm{P}}(v)}\right] - 1.
\end{aligned}
$$

□

Table 3.3: Specification of datasets for classification tasks. #Inst., #Attr., "Viewpoint" and "Target" denote the number of sxamples, the number of attributes, the attribute used as the target and the attribute used as the viewpoint, respectively.

dataset	#Inst.	#Attr.	Viewpoint	Target
Adult (Frank et al. 2010)	16281	13	gender	income
Dutch Census (Dutch Central Bureau for Statistics 2001)	60420	10	gender	income
Bank Marketing (Frank et al. 2010)	45211	17	loan	term deposit
Credit Approval (Frank et al. 2010)	690	15	A1	A16
German Credit Data (Frank et al. 2010)	1000	20	foreign worker	credit risk

Table 3.4: Statistics of datasets for classification tasks. $\#y_+$ and $\#v_+$ represent the number of positive target and viewpoint values, respectively. The prediction accuracy of logistic regression for the target (Acc (y)) and viewpoint (Acc (v)) are also shown.

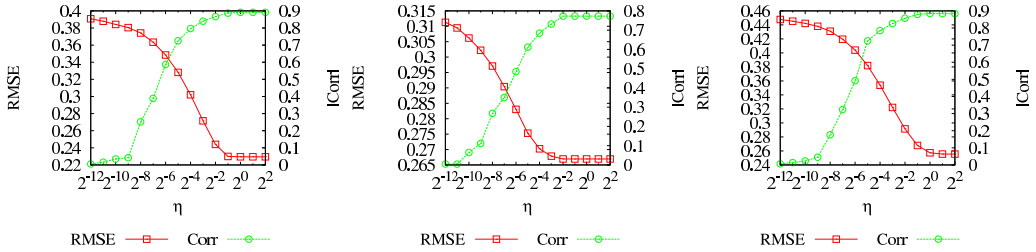
dataset	$\#y_+$	$\#v_+$	Acc (y)	Acc (v)
Adult (Frank et al. 2010)	3846 (23.6%)	10860 (66.7%)	0.850	0.842
Dutch Census (Dutch Central Bureau for Statistics 2001)	31657 (52.4%)	30273 (50.1%)	0.819	0.665
Bank Marketing (Frank et al. 2010)	5289 (11.7%)	7244 (16.0%)	0.900	0.839
Credit Approval (Frank et al. 2010)	307 (44.5%)	480 (69.6%)	0.875	0.676
German Credit Data (Frank et al. 2010)	300 (30.0%)	37 (3.7%)	0.757	0.961

Table 3.5: Summary of the treatment of the viewpoint random variables in two settings.

case	method	learning of f	fairness guarantee	fairness measure
Case 1	others	$\boldsymbol{x}, v$	v	$f(y\|\boldsymbol{x};\boldsymbol{\theta}), v$
	ours	$\boldsymbol{x}, v$	$g(v\|\boldsymbol{x})$	$f(y\|\boldsymbol{x};\boldsymbol{\theta}), v$
Case 2	others	$\boldsymbol{x}, \hat{v}$	$\hat{v}$	$f(y\|\boldsymbol{x};\boldsymbol{\theta}), v$
	ours	$\boldsymbol{x}, \hat{v}$	$g(v\|\boldsymbol{x})$	$f(y\|\boldsymbol{x};\boldsymbol{\theta}), v$

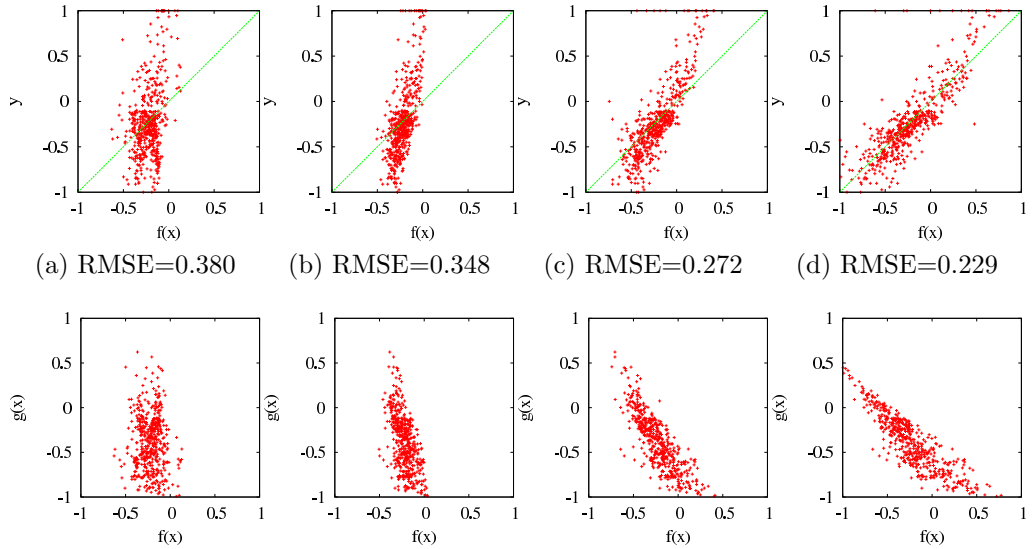
Table 3.6: Specification of datasets for regression task. #Inst., #Attr., "Viewpoint" and "Target" denote the number of samples, the number of attributes, the attribute used as the target and the attribute used as the viewpoint, respectively. "Corr" represents the correlation coefficient between the target and the viewpoint.

dataset	#Inst.	#Attr.	Viewpoint	Target	Corr
Housing (Frank et al. 2010)	506	14	LSTAT	MEDV	-0.738
Wine Quality (Red) (Frank et al. 2010)	1599	12	alcohol	quality	0.476
Communities and Crime (Frank et al. 2010)	1994	123	PctKids-2Par	ViolentCri-mesPerPop	-0.738



(a) Housing (b) Wine Quality (c) Communities and Crime

Figure 3.3: The plots show RMSE and the absolute value of the correlation coefficient between the predicted target value and the viewpoint value corresponding to the neutrality parameter η.



(a) RMSE=0.380 (b) RMSE=0.348 (c) RMSE=0.272 (d) RMSE=0.229

(e) $\eta = 2^{-9}$ (f) $\eta = 2^{-6}$ (g) $\eta = 2^{-3}$ (h) no neutralization

Figure 3.4: Scatter plots with respect to Housing dataset. Top row: scatter plots of target prediction value $\hat{y}$ and true target value y. Bottom row: scatter plots of target prediction value $\hat{y}$ and viewpoint prediction value $\hat{v}$. Correlation in the $\hat{y} - \hat{v}$ plots means that the fair level of the regression model is low.

Chapter 4

Fair Empirical Risk Minimization with populational Fairness Guarantee

In supervised learning, overfitting is a fundamental problem to construct the best accurate predictor. If the learned predictor overfits to the given training samples, even with a low prediction error for the training samples is small, the prediction error for the test samples can be large. Overfitting is also problematic when the learner removes the disparate impact from the learned predictor. Even if the learner constructs a predictor so that the level of empirical unfairness is small, the level of populational unfairness can be large. In this chapter, we deal with this problem of populational fairness and aims to develop a learning algorithm that removes the disparate impact with a guarantee of populational fairness.

In classification, the predictor is usually referred to as *classifier* or *hypothesis*. In this study, let us consider two hypotheses; that is, *target hypothesis* f and *viewpoint hypothesis* g. The target hypothesis is a mapping from the input to the target, and this is what we want in the end. The viewpoint hyothesis is a mapping from the input to the viewpoint and is revealed to the leaner prior to the learning phase as well as Chapter 3. Given a viewpoint hypothesis g, we evaluate the degree of empirical unfairness and populational unfairness as *empirical neutrality risk* and *generalization neutrality risk*, respectively. The goal of the learner is to construct the target hypothesis f that achieves the most efficient trade-off between the generalization risk of classification and the generalization neutrality risk of fairness.

Within the context of removing discrimination from classifiers, the need for a fairness guarantee has already been extensively studied. While many researchers developed fair learning algorithms that ensures empirical fairness, such as (Bolukbasi et al. 2016; Calders et al. 2010; Feldman et al. 2015; Fukuchi et al. 2013; Hardt et al. 2016; Kamiran et al. 2010; Kamishima et al. 2012b; Zafar et al. 2017; Zemel et al. 2013; Zliobaite et al. 2011b), there are a few studies aiming to ensure the populational fairness. For the disparate mistreatment, Hardt et al. (2016) analyzed the error caused by approximation of the probability distribution ρ, and hence we can bound the error caused by the use of the empirical distribution of ρ. Also, Woodworth et al. (2017) developed a two-step algorithm which consists of the interprocessing and postprocessing methods, and proved the guarantee of populational fairness of the algorithm. For the bandit setting or the reinforcement learning setting, Jabbari et al. (2017) and Joseph et al. (2016) showed the populational fairness guarantee for the disparate mistreatment. However, there is no method that removes the disparate impact with the guarantee of populational fairness for supervised learning.

The existing methods incorporate a hypothesis fairness measure into the ob-

jective function in the form of a regularization term or constraint; however, these are all non-convex. One of the reasons why populational fairness is not theoretically guaranteed for these methods is the non-convexity of the objective functions. In this study, we introduce a convex surrogate for a fairness measure to provide a theoretical analysis of populational fairness.

Our Contribution. The contribution of this study is three-fold. First, we introduce our novel neutralized empirical risk minimization (NERM) framework in which, assuming the target hypothesis and viewpoint hypothesis output binary predictions, it is possible to learn a target hypothesis that minimizes empirical and empirical neutral risks. Given samples and a viewpoint hypothesis, NERM is formulated as a convex optimization problem where the objective function is the linear combination of two terms: the empirical risk term penalizing the target hypothesis prediction error and the neutralization term penalizing correlation between the target and the viewpoint. The predictive performance and the fairness performance can be balanced by adjusting a parameter, referred to as the neutralization parameter, that balances the two term. Because of its convexity, the optimality of the resultant target hypothesis is guaranteed (in Section 4.2).

Second, we derive a bound on empirical and generalization neutrality risks for NERM. We also show that the bound on the generalization neutrality risk can be controlled by the neutralization parameter (in Section 4.3). As discussed before, many diverse algorithms targeting the neutralization of supervised classifications have been presented. However, none of these have given theoretical guarantees on generalization neutrality risk. To the best of our knowledge, this is the first study that gives a bound on generalization neutrality risk.

Third, we present a specific NERM learning algorithm for neutralized linear classification. The derived learning algorithm is interpreted as a *support vector machine* (SVM) (Vapnik 1998) variant with a neutralization guarantee (in Section 4.4).

4.1 Empirical Risk Minimization

The most of setups are the same as supervised learning under fairness and Section 2.1.2 described in Chapter 2. We restrict our attention to binary classification, $Y = \{-1, 1\}$, but our method can be expanded to handle multi-valued classification via a straightforward modification. Given the i.i.d. samples, the

supervised learning objective is to construct a target hypothesis $f : X \to \mathbb{R}$ where the hypothesis is chosen from a class of measurable functions $f \in \mathcal{F}$. We assume that classification results are given by $\text{sgn} \circ f(x)$, that is, $y = 1$ if $f(x) > 0$; otherwise $y = -1$. Given a loss function $\ell : Y \times \mathbb{R} \to \mathbb{R}^+$, the generalization risk is defined by

$$R(f) = \mathbf{E}[\ell(Y, f(X))].$$

Our goal is to find $f^* \in \mathcal{F}$ that minimizes the generalization risk $R(f)$. In general, ρ is unknown and the generalization risk cannot be directly evaluated. Instead, we take the approach of empirical risk minimization (ERM) which minimize the empirical loss with respect to the samples D_n

$$R_n(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, f(X_i)).$$

To avoid overfitting, a regularization term $\Omega : \mathcal{F} \to \mathbb{R}^+$ is added to the empirical loss by penalizing complex hypotheses. Minimization of the empirical loss with a regularization term is referred to as *regularized empirical risk minimization (RERM)*.

4.1.1 Generalization risk bound

Rademacher Complexity measures the complexity of a hypothesis class with respect to a probability measure that generates samples. The Rademacher Complexity of class $\mathcal{F}$ is defined as

$$\mathcal{R}_n(\mathcal{F}) = \mathbf{E}[D_n, \boldsymbol{\sigma}] \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(x_i)$$

where $\boldsymbol{\sigma} = (\sigma_1, ..., \sigma_n)^T$ are independent random variables such that $\mathbb{P}\sigma_i = 1 = \mathbb{P}\sigma_i = -1 = 1/2$. Bartlett et al. (2002) derived a generalization loss bound using the Rademacher complexity as follows:

Theorem 4 (Bartlett et al. (2002))**.** *Let ρ be a probability measure on $(Z, \mathfrak{Z})$ and let $\mathcal{F}$ be a set of real-value functions defined on X, with $\sup\{|f(x)| : f \in \mathcal{F}\}$ finite for all $x \in X$. Suppose that $\phi : \mathbb{R} \to [0, c]$ satisfies and is Lipschitz continuous with constant L_ϕ. Then, with probability at least $1 - \delta$, every function in $\mathcal{F}$ satisfies*

$$R(f) \leq R_n(f) + 2L_\phi \mathcal{R}_n(\mathcal{F}) + c\sqrt{\frac{\ln(2/\delta)}{2n}}.$$

4.2 Generalization Neutrality Risk and Empirical Neutrality Risk

In this section, we introduce the viewpoint hypothesis into the ERM framework and define a new principle of supervised learning, neutralized ERM (NERM), with the notion of *generalization neutrality risk*. Convex relaxation of the neutralization measure is also discussed in this section.

4.2.1 $+1/-1$ Generalization neutrality risk

Suppose a measurable function $g : X \to \mathbb{R}$ is given. The prediction of g is referred to as the *viewpoint* and g is referred to as the *viewpoint hypothesis*. We say the target hypothesis f is neutral to the viewpoint hypothesis g if the target predicted by the learned target hypothesis f and the viewpoint predicted by the viewpoint hypothesis g are not mutually correlating. In our setting, we assume the target hypothesis f and viewpoint hypothesis g to give binary predictions by $\mathrm{sgn} \circ f$ and $\mathrm{sgn} \circ g$, respectively. Given a probability measure ρ and a viewpoint hypothesis g, the neutrality of the target hypothesis f is defined by the correlation between $\mathrm{sgn} \circ f$ and $\mathrm{sgn} \circ g$ over ρ. If $f(x)g(x) > 0$ holds for multiple samples, then the classification $\mathrm{sgn} \circ f$ closely correlates to the viewpoint $\mathrm{sgn} \circ g$. On the other hand, if $f(x)g(x) \leq 0$ holds for multiple samples, then the classification $\mathrm{sgn} \circ f$ and the viewpoint $\mathrm{sgn} \circ g$ are inversely correlating. Since we want to suppress both correlations, our neutrality measure is defined as follows:

Definition 4 ($+1$/-1 generalization neutrality risk)**.** Let $f \in \mathcal{F}$ and $g \in \mathcal{G}$ be a target hypothesis and viewpoint hypothesis, respectively. Let ρ be a probability measure over $(Z, \mathfrak{Z})$. Then, the *+1/-1 generalization neutrality risk* of target hypothesis f with respect to viewpoint hypothesis g over ρ is defined by

$$C_{\mathrm{sgn}}(f, g) = \left| \int \mathrm{sgn} \circ (fg) d\rho \right|.$$

When the probability measure ρ cannot be obtained, a $+1/-1$ generalization neutrality risk $C_{\mathrm{sgn}}(f, g)$ can be empirically evaluated with respect to the given samples D_n.

Definition 5 ($+1/-1$ empirical neutrality risk)**.** Suppose that $D_n = \{(x_i, y_i)\}_{i=1}^n \in Z^n$ is a given sample set. Let $f \in \mathcal{F}$ and $g \in \mathcal{G}$ be the target hypothesis and the viewpoint hypothesis, respectively. Then, the *+1/−1*

empirical neutrality risk of target hypothesis f with respect to viewpoint hypothesis g is defined by

$$C_{n,\mathrm{sgn}}(f,g) = \frac{1}{n}\left|\sum_{i=1}^{n} \mathrm{sgn}(f(x_i)g(x_i))\right|. \tag{4.1}$$

4.2.2 Neutralized empirical risk minimization (NERM)

With the definition of neutrality risk, a novel framework, the *Neutralized Empirical Risk Minimization* (NERM) is introduced. NERM is formulated as the minimization of the empirical risk and empirical $+1/-1$ neutrality risk:

$$\min_{f\in\mathcal{F}} \; R_n(f) + \Omega(f) + \eta C_{n,\mathrm{sgn}}(f,g). \tag{4.2}$$

where $\eta > 0$ is the neutralization parameter which determines the trade-off ratio between the empirical risk and the empirical neutrality risk.

4.2.3 Convex relaxation of $+1/-1$ neutrality risk

Unfortunately, the optimization problem defined by Eq (4.2) cannot be efficiently solved due to the nonconvexity of Eq (4.1). Therefore, we must first relax the absolute value function of $C_{\mathrm{sgn}}(f,g)$ into the max function. Then, we introduce a convex surrogate of the sign function, yielding a convex relaxation of the $+1/-1$ neutrality risk.

By letting I be the indicator function, the $+1/-1$ generalization neutrality risk can be decomposed into two terms:

$$\begin{aligned} C_{\mathrm{sgn}}(f,g) &= \Big| \underbrace{\int I(\mathrm{sgn}\, g(x) = \mathrm{sgn}\, f(x))\rho(dx)}_{\text{prob. that } f \text{ agrees with } g} - \underbrace{\int I(\mathrm{sgn}\, g(x) \neq \mathrm{sgn}\, f(x))\rho(dx)}_{\text{prob. that } f \text{ disagrees with } g} \Big| \\ &:= |C^{+}_{\mathrm{sgn}}(f,g) - C^{-}_{\mathrm{sgn}}(f,g)| \end{aligned} \tag{4.3}$$

The upper bound of the $+1/-1$ generalization neutrality risk $C_{\mathrm{sgn}}(f,g)$ is tight if $C^{+}_{\mathrm{sgn}}(f,g)$ and $C^{-}_{\mathrm{sgn}}(f,g)$ are close. Thus, the following property is derived.

Proposition 2. *Let* $C^{+}_{\mathrm{sgn}}(f,g)$ *and* $C^{-}_{\mathrm{sgn}}(f,g)$ *be functions defined in Eq* (4.3). *For any* $\eta \in [0.5, 1]$, *if*

$$C^{\max}_{\mathrm{sgn}}(f,g) \coloneqq \max(C^{+}_{\mathrm{sgn}}(f,g), C^{-}_{\mathrm{sgn}}(f,g)) \leq \eta,$$

then

$$C_{\mathrm{sgn}}(f,g) = |C^{+}_{\mathrm{sgn}}(f,g) - C^{-}_{\mathrm{sgn}}(f,g)| \leq 2\eta - 1.$$

Proposition 2 shows that $C^{\max}_{\mathrm{sgn}}(f,g)$ can be used as the generalization neutrality risk instead of $C_{\mathrm{sgn}}(f,g)$. Next, we relax the indicator function contained in $C^{\pm}_{\mathrm{sgn}}(f,g)$.

Definition 6 (relaxed convex generalization neutrality risk)**.** Let $f \in \mathcal{F}$ and $g \in \mathcal{G}$ be a classification hypothesis and viewpoint hypothesis, respectively. Let ρ be a probability measure over $(Z, \mathfrak{Z})$. Let $\psi : \mathbb{R} \to \mathbb{R}^{+}$ be a convex function and

$$C^{\pm}_{\psi}(f,g) = \int \psi(\pm g(x) f(x)) \rho(dx).$$

Then, the *relaxed convex generalization neutrality risk* of f with respect to g is defined by

$$C_{\psi}(f,g) = \max(C^{+}_{\psi}(f,g), C^{-}_{\psi}(f,g)).$$

The empirical evaluation of relaxed convex generalization neutrality risk is defined in a straightforward manner.

Definition 7 (convex relaxed empirical neutrality risk)**.** Suppose $D_n = \{(x_i, y_i)\}_{i=1}^{n} \in Z^n$ to be a given sample set. Let $f \in \mathcal{F}$ and $g \in \mathcal{G}$ be the target hypothesis and the viewpoint hypothesis, respectively. Let $\psi : \mathbb{R} \to \mathbb{R}^{+}$ be a convex function and

$$C^{\pm}_{n,\psi}(f,g) = \frac{1}{n} \sum_{i=1}^{n} \psi(\pm g(x_i) f(x_i)).$$

Then, *relaxed convex empirical neutrality risk* of f with respect to g is defined by

$$C_{n,\psi}(f,g) = \max(C^{+}_{n,\psi}(f,g), C^{-}_{n,\psi}(f,g)).$$

$C^{\pm}_{n,\psi}(f,g)$ is convex because it is a summation of the convex function ψ. Noting that $\max(f_1(x), f_2(x))$ is convex if f_1 and f_2 are convex, $C_{n,\psi}(f,g)$ is convex as well.

4.2.4 NERM with relaxed convex empirical neutrality risk

Finally, we derive the convex formulation of NERM with the relaxed convex empirical neutrality risk as follows:

$$\min_{f \in \mathcal{F}} \; R_n(f) + \Omega(f) + \eta C_{n,\psi}(f, g). \tag{4.4}$$

If the regularized empirical risk is convex, then this is a convex optimization problem. The neutralization term resembles the regularizer term in the formulation sense. However, whereas the regularizer represents a prior structual information of f, the neutralization term realizes a different role because it depends on samples. Since the neutralization term is dependent on samples, it can be interpreted as a prior information of *data*. The notion of a prior data information is relevant to *transfer learning* (Pan et al. 2010), which aims to achieve learning dataset information from other datasets. However, further research on the relationships between the neutralization and transfer learning will be left as an area of future work.

4.3 Generalization Neutrality Risk Bound

In this section, we show theoretical analyses of NERM generalization neutrality risk and generalization risk. First, we derive a probabilistic uniform bound on the generalization neutrality risk for any $f \in \mathcal{F}$ with respect to the empirical neutrality risk $C_{n,\psi}(f, g)$ and the Rademacher complexity of $\mathcal{F}$. Then, we derive a bound on the generalization neutrality risk of the optimal hypothesis.

For convenience, we introduce the following notations. For a hypothesis class $\mathcal{F}$ and constant $c \in \mathbb{R}$, we denote $-\mathcal{F} = \{-f : f \in \mathcal{F}\}$ and $c\mathcal{F} = \{cf : f \in \mathcal{F}\}$. For any function $\phi : \mathbb{R} \to \mathbb{R}$, let $\phi \circ \mathcal{F} = \{\phi \circ f : f \in \mathcal{F}\}$. Similarly, for any function $g : X \to \mathbb{R}$, let $g\mathcal{F} = \{h : f \in \mathcal{F}, h(x) = g(x)f(x) \; \forall x \in X\}$.

4.3.1 Uniform bound of generalization neutrality risk

A probabilistic uniform bound on $C_\psi(f, g)$ for any hypothesis $f \in \mathcal{F}$ is derived as follows.

Theorem 5. *Let $C_\psi(f, g)$ and $C_{n,\psi}(f, g)$ be the relaxed convex generalization neutrality risk and the relaxed convex empirical neutrality risk of $f \in \mathcal{F}$ w.r.t. $g \in \mathcal{G}$. Suppose that $\psi : \mathbb{R} \to [0, c]$ is Lipschitz continuous with constant L_ψ. Then, with probability at least $1 - \delta$, every function in $\mathcal{F}$ satisfies*

$$C_\psi(f, g) \leq C_{n,\psi}(f, g) + 2L_\psi \mathcal{R}_n(g\mathcal{F}) + c\sqrt{\frac{\ln(2/\delta)}{2n}}.$$

As proved by Theorem 5, $C_\psi(f, g) - C_{n,\psi}(f, g)$, the approximation error of the generalization neutrality risk is uniformly upper-bounded by the Rademacher complexity of hypothesis classes $g\mathcal{F}$ and $O(\sqrt{\ln(1/\delta)/n})$, where δ is the confidence probability and n is the sample size.

4.3.2 Generalization neutrality risk bound for NERM optimal hypothesis

Let $\hat{f} \in \mathcal{F}$ be the optimal hypothesis of NERM. We derive the bounds on the empirical and generalization neutrality risks achieved by $\hat{f}$ under the following conditions:

1. Hypothesis class $\mathcal{F}$ includes a hypothesis f_0 s.t. $f_0(x) = 0$ for $\forall x$, and
2. the regularization term of f_0 is $\Omega(f_0) = 0$.

(A)

The conditions are relatively moderate. For example, consider the linear hypothesis $f(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x}$ and $\Omega(f) = \|\boldsymbol{w}\|_2^2$ (ℓ_2^2 norm) and let $W \subseteq \mathbb{R}^D$ be a class of the linear hypothesis. If $\mathbf{0} \in W$, the two conditions above are satisfied. Assuming that $\mathcal{F}$ satisfies these conditions, the following theorem provides the bound on the generalization neutrality risk.

Theorem 6. *Let $\hat{f}$ be the optimal target hypothesis of NERM, where the viewpoint hypothesis is $g \in \mathcal{G}$ and the neutralization parameter is η. Suppose that $\psi : \mathbb{R} \to [0, c]$ satisfies and is Lipschitz continuous with constant L_ψ. If conditions (A) are satisfied, then with probability at least $1 - \delta$,*

$$C_\psi(\hat{f}, g) \leq \psi(0) + \phi(0)\frac{1}{\eta} + 2L_\psi \mathcal{R}_n(g\mathcal{F}) + c\sqrt{\frac{\ln(2/\delta)}{2n}}.$$

For the proof of Theorem 6, we first derive the upper bound of the empirical neutrality risk of $\hat{f}$.

Corollary 1. *If the conditions (A) are satisfied, then the empirical relaxed convex neutrality risk of $\hat{f}$ is bounded by*

$$C_{n,\psi}(\hat{f}, g) \leq \psi(0) + \phi(0)\frac{1}{\eta}.$$

Theorem 6 is immediately obtained from Theorem 5 and Corollary 1.

4.3.3 Generalization risk bound for NERM

In this section, we compare the generalization risk bound of NERM with that of a regular ERM. Theorem 4 denotes a uniform bound of the generalization risk. This theorem holds with the hypotheses which are optimal in terms of NERM and ERM. However, the hypotheses which are optimal regarding NERM and ERM have different empirical risk values. The empirical risk of NERM is greater than that of ERM since NERM has a term that penalizes less neutrality. More precisely, if we let $\bar{f}$ be the optimal hypothesis in term of ERM, we have

$$R_n(\hat{f}) - R_n(\bar{f}) \geq 0. \tag{4.5}$$

The reason for this is that empirical risk of any other hypothesis is greater than one of $\bar{f}$ since $\bar{f}$ minimizes empirical risk. Furthermore, due to $\hat{f}$ is a minimizer of $R_n(f) + \eta C_{n,\phi}(f, g)$, we have

$$\begin{aligned} R_n(\hat{f}) + \eta C_{n,\phi}(\hat{f}, g) - R_n(\bar{f}) - \eta C_{n,\phi}(\bar{f}, g) &\leq 0 \\ R_n(\hat{f}) - R_n(\bar{f}) &\leq \eta(C_{n,\phi}(\bar{f}, g) - C_{n,\phi}(\hat{f}, g)). \end{aligned} \tag{4.6}$$

Since the left term of this inequality is greater than zero due to Eq (4.5), the empirical risk becomes greater if the empirical neutrality risk becomes lower. Consequently, the generalization risk of NERM is larger than that of ERM to make neutrality risk lower.

4.4 Neutral SVM

SVMs (Vapnik 1998) is a margin-based supervised learning method for binary classification. The algorithm of SVMs can be interpreted as minimization of the empirical risk with regularization term, which follows the RERM principle. In this section, we introduce an SVM variant that follows the NERM principle.

The soft-margin SVM employs the linear classifier $f(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + b$ as the target hypothesis. In the objective function, the hinge loss is used for the loss function, as $\phi(yf(x)) = \max(0, 1 - yf(x))$, and the ℓ_2 norm is used for the regularization term, $\Omega(f) = \lambda\|f\|_2^2/2n$, where $\lambda > 0$ denotes the regularization parameter. In our SVM in NERM, referred to as the neutral SVM, the loss function and regularization term are the same as in the soft-margin SVM. For a surrogate function of the neutralization term, the hinge loss $\psi(\pm g(x)f(x)) = \max(0, 1 \mp g(x)f(x))$ was employed. Any hypothesis can be used for the viewpoint hypothesis. Accordingly, following the NERM principle defined in Eq (4.4), the neutral SVM is formulated by

$$\min_{\boldsymbol{w},b} \sum_{i=1}^{n} \max(0, 1 - y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b)) + \frac{\lambda}{2}\|\boldsymbol{w}\|_2^2 + \eta C_{n,\psi}(\boldsymbol{w}, b, g), \tag{4.7}$$

where

$$\begin{aligned} C_{n,\psi}(\boldsymbol{w}, b, g) &= \max(C^+_{n,\psi}(\boldsymbol{w}, b, g), C^-_{n,\psi}(\boldsymbol{w}, b, g)), \\ C^{\pm}_{n,\psi}(\boldsymbol{w}, b, g) &= \sum_{i=1}^{n} \max(0, 1 \mp g(\boldsymbol{x}_i)(\boldsymbol{w}^T\boldsymbol{x}_i + b)). \end{aligned}$$

Since the risk, regularization, and neutralization terms are all convex, the objective function of the neutral SVM is convex. The primal form can be solved by applying the subgradient method (Shor et al. 1985) to Eq (4.7).

4.5 Experiments

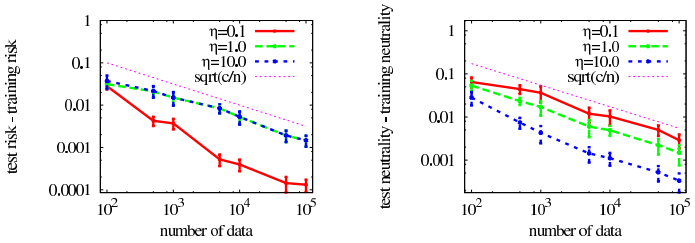
In this section, we present an experimental evaluation of our neutral SVM for synthetic and real datasets. In the experiments with synthetic data, we experimentally evaluate the change of generalization risk and generalization neutrality risk according to the number of samples, in which their relations are described in Theorem 5. In the experiments for real datasets, we compare our method with CV2NB Calders et al. 2010, PR (Kamishima et al. 2012b) and η-neutral logistic regression (ηLR for short) (Fukuchi et al. 2013) regarding risk and neutrality risk. The CV2NB method learns a naíve Bayes model and then modifies the model parameters so that the resultant CV score approaches zero. The PR and ηLR are based on maximum likelihood estimation of a logistic regression (LR) model. These methods have two parameters, the regularizer parameter λ, and the neutralization parameter η. The PR penalizes the objective function of the LR model with mutual information. The ηLR

performs maximum likelihood estimation of the LR model while enforcing η-neutrality as constraints. The neutralization parameter of neutral SVM and PR balances risk minimization and neutrality maximization. Thus, it can be tuned in the same manner used to determine the regularizer parameter. The neutralization parameter of ηLR determines the region of the hypothesis in which the hypotheses are regarded as neutral. The tuning strategy of the regularizer parameter and neutralization parameter are different in all these methods. We determined the neutralization parameter tuning range of these methods via preliminary experiments.

4.5.1 Synthetic dataset

To investigate the change of generalization neutrality risk with sample size n, we performed our neutral SVM experiments for a synthetic dataset. First, we constructed the input $\boldsymbol{x}_i \in \mathbb{R}^{10}$ with the vector being sampled from the uniform distribution over $[-1, 1]^{10}$. The target y_i corresponding to the input $\boldsymbol{x}_i$ is generated as $y_i = \mathrm{sgn}(\boldsymbol{w}_y^T \boldsymbol{x}_i)$ where $\boldsymbol{w}_y \in \mathbb{R}^{10}$ is a random vector drawn from the uniform distribution over $[-1, 1]^{10}$. Noises are added to labels by inverting the label with probability $1/(1 + \exp(-100|\boldsymbol{w}_y^T \boldsymbol{x}_i|))$. The inverting label probability is small if the input $\boldsymbol{x}_i$ is distant from a plane $\boldsymbol{w}_y^T \boldsymbol{x} = 0$. The viewpoint v_i corresponding to the input $\boldsymbol{x}_i$ is generated as $v_i = \mathrm{sgn}(\boldsymbol{w}_v^T \boldsymbol{x}_i)$, where the first element of $\boldsymbol{w}_v$ is set as $w_{v,1} = w_{y,1}$ and the rest of elements are drawn from the uniform distribution over $[-1, 1]^9$. Noises are added in the same manner as the target. The equality of the first element of $\boldsymbol{w}_y$ and $\boldsymbol{w}_v$ leads to correlation between y_i and v_i. Set the regularizer parameter as $\lambda = 0.05n$. The neutralization parameter was varied as $\eta \in \{0.1, 1.0, 10.0\}$. In this situation, we evaluate the approximation error of the generalization risk and the generalization neutrality risk by varying sample size n. The approximation error of generalization risk is the difference of the empirical risk between training and test samples, while that of the generalization neutrality risk is the difference of the empirical neutrality risk between training and test samples. Five fold cross-validation was used for evaluation of the approximation error of the empirical risk and empirical neutrality; the average of ten different folds are shown as the results.

Results. Figure 4.1 shows the change of the approximation error of generalization risk (the difference of the empirical risks w.r.t. test samples and training samples), and the approximation error of generalization neutrality risk (the difference of the empirical neutrality risks w.r.t. test samples and training samples) with changing sample size n. The plots in Figure 4.1 left and right



(a) risk (b) neutrality risk

Figure 4.1: Change of approximation error of generalization risk (left) and approximation error of generalization neutrality risk (right) by neutral SVM (our proposal) according to varying the number of samples n. The horizontal axis shows the number of samples n, and the error bar shows the standard deviation across the change of five-fold division. The line "sqrt(c/n)" denotes the convergence rate of the approximation error of the generalization risk (in Theorem 4) or the generalization neutrality risk (in Theorem 5). Each line indicates the results with the neutralization parameter $\eta \in \{0.1, 1.0, 10.0\}$. The regularizer parameter was set as $\lambda = 0.05n$.

show the approximation error of generalization risk and the approximation error of generalization neutrality risk, respectively.

Recall that the discussions in Section 4.3.3 showed that the approximation error of generalization risk decreases with $O(\sqrt{\ln(1/\delta)/n})$ rate. As indicated by the Theorem 4, Figure 4.1 (left) clearly shows that the approximation error of the generalization risk decreases as sample size n increases. Similarly, discussions in Section 4.3.1 revealed that the approximation error of generalization neutrality risk also decreases with $O(\sqrt{\ln(1/\delta)/n})$ rate, which can be experimentally confirmed in Figure 4.1 (right). The plot clearly shows that the approximation error of the generalization neutrality risk decreases as the sample size n increases.

4.5.2 Real datasets

We compare the classification performance and neutralization performance of neutral SVM with CV2NB, PR, and ηLR for some real datasets specified in Table 4.1. In Table 4.1, #Inst. and #Attr. denote the sample size and the number of attributes, respectively; "Viewpoint" and "Target" denote the attributes used as the target and the viewpoint, respectively. All dataset attributes were discretized by the same procedure described in (Calders et al.

Table 4.1: Specification of Datasets

dataset	#Inst.	#Attr.	Viewpoint	Target
Adult	16281	13	gender	income
Dutch	60420	10	gender	income
Bank	45211	17	loan	term deposit
German	1000	20	foreign worker	credit risk

Table 4.2: Range of neutralization parameter

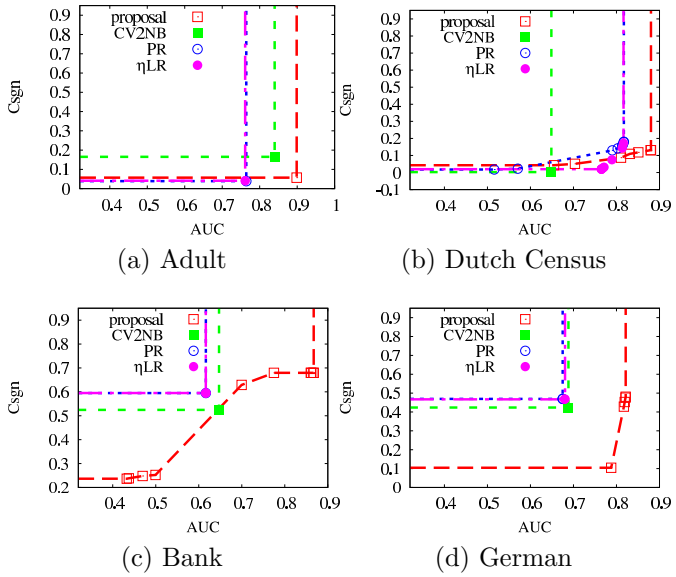
method	range of neutralization parameter
PR	0, 0.01, 0.05, 0.1, ..., 100
ηLR	0, 5×10^{-5}, 1×10^{-4}, 5×10^{-4}, ..., 0.5
neutral SVM	0, 0.01, 0.05, 0.1, ..., 100

2010) and coded by 1-of-K representation for PR, ηLR, and neutral SVM. We used the primal problem of neutral SVM (non-kernelized version) to compare our method with the other methods in the same representation. For PR, ηLR, and neutral SVM, the regularizer parameter was tuned in advance for each dataset in the non-neutralized setting by means of five-fold cross validation, and the tuned parameter was used for the neutralization setting. CV2NB has no regularization parameter to be tuned. Table 4.2 shows the range of the neutralization parameter used for each method.

The classification performance and neutralization performance was evaluated with *Area Under the receiver operating characteristic Curve* (AUC) and $+1/-1$ empirical neutrality risk $C_{n,\mathrm{sgn}}(f, g)$, respectively. Both measures were evaluated with five-fold cross-validation and the average of ten different folds are shown in the plots.

Results. Figure 4.2 shows the classification performance (AUC) and neutralization performance ($C_{n,\mathrm{sgn}}(f, g)$) at different settings of neutralization parameter η. In the graph, the best result is shown at the right bottom. Since the classification performance and neutralization performance are in a trade-off relationship, as indicated by Theorem Eq (4.6), the results dominated by the other parameter settings are omitted in the plot for each method.

CV2NB achieves the best neutrality in Dutch Census, but is less neutral compared to the other methods in the rest of the datasets. In general, the classification performance of CV2NB is lower than those of the other methods due to the poor classification performance of naíve Bayes. PR and ηLR achieve competitive performance to neutral SVM in Adult and Dutch Census



(a) Adult (b) Dutch Census

(c) Bank (d) German

Figure 4.2: Performance of CV2NB, PR, ηLR, and neutral SVM (our proposal). The vertical axis shows the AUC, and horizontal axis shows $C_{n,sgn}(f, g)$. The points in these plots are omitted if they are dominated by others. The bottommost line shows limitations of neutralization performance, and the rightmost line shows limitations of classification performance, which are shown only as guidelines.

in term of the neutrality risk, but the results are dominated in term of AUC. Furthermore, the results of PR and ηLR in Bank and German are dominated. The results of neutral SVM are dominant compared to the other methods in Bank and German dataset, and it is noteworthy that the neutral SVM achieves the best AUC in almost all datasets. This presumably reflects the superiority of SVM in the classification performance, compared to the naíve Bayes and logistic regression.

4.6 Conclusion

We proposed a novel framework, NERM. NERM provides a framework that learns a target hypothesis that minimizes the empirical risk, and that is empirically neutral regarding risk to a given viewpoint hypothesis. Our contributions are as follows: (1) We define NERM as a framework for guaranteeing populational fairness of classification problems. In contrast to existing methods, the NERM can be formulated as a convex optimization problem by using convex

relaxation. (2) We provide a theoretical analysis of the generalization neutrality risk of NERM. The theoretical results show the approximation error of the generalization neutrality risk of NERM is uniformly upper-bounded by the Rademacher complexity of hypothesis class $g\mathcal{F}$ and $O(\sqrt{\ln(1/\delta)/n})$. Moreover, we derive a bound on the generalization neutrality risk for the optimal hypothesis corresponding to the neutralization parameter η. (3) We present a specific learning algorithm for NERM, neutral SVM.

Suppose the viewpoint is set to some private information. Then, noting that neutralization reduces correlation between the target and viewpoint values, outputs obtained from the neutralized target hypothesis do not help to predict the viewpoint values. Thus, neutralization realizes a certain type of privacy preservation. Besides, as already mentioned, NERM can be interpreted as a variant of transfer learning by regarding the neutralization term as data-dependent prior knowledge. Clarifying connection to privacy-preservation and transfer learning is remained as an area of future work.

Appendix 4.A Proof of Proposition 2

Proof. Noting that

$$C^{+}_{\mathrm{sgn}}(f,g) + C^{-}_{\mathrm{sgn}}(f,g) = 1$$

holds, we have

$$|C^{+}_{\mathrm{sgn}}(f,g) - C^{-}_{\mathrm{sgn}}(f,g)| = |2C^{+}_{\mathrm{sgn}}(f,g) - 1|.$$

Since $C^{+}_{\mathrm{sgn}}(f,g) \geq 1-\eta$ and $C^{+}_{\mathrm{sgn}}(f,g) \leq \eta$ by the assumption,

$$|C^{+}_{\mathrm{sgn}}(f,g) - C^{-}_{\mathrm{sgn}}(f,g)| \leq 2\eta - 1.$$

□

Appendix 4.B Proof of Theorem 5

Proof. First, we derive the uniform bound on the relaxed convex empirical neutrality risk. For any $f \in \mathcal{F}$, we have

$$C^{\pm}_{\psi}(f,g) \leq C^{\pm}_{n,\psi}(f,g) + \sup_{f\in\mathcal{F}}(C^{\pm}_{\psi}(f,g) - C^{\pm}_{n,\psi}).$$

Using the McDiarmid inequality, with probability $1-\delta/2$, we have

$$\sup_{f\in\mathcal{F}}(C_\psi^\pm(f,g)-C_{n,\psi}^\pm)\leq \mathbf{E}[D_n]\sup_{f\in\mathcal{F}}(C_\psi^\pm(f,g)-C_{n,\psi}^\pm)+c\sqrt{\frac{\ln(2/\delta)}{2n}}.$$

Application of the symmetrization technique yields the following bound:

$$\begin{aligned}
&\mathbf{E}[D_n]\sup_{f\in\mathcal{F}}(C_\psi^\pm(f,g)-C_{n,\psi}^\pm)\\
=&\mathbf{E}[D_n]\sup_{f\in\mathcal{F}}(\mathbf{E}[D_n']C_{n,\psi}^\pm(f,g)-C_{n,\psi}^\pm) &(4.8)\\
\leq&\mathbf{E}[D_n,D_n']\sup_{h\in\psi\circ\pm g\mathcal{F}}\frac{1}{n}\sum_{i=1}^n(h(x_i)-h(x_i')) &(4.9)\\
\leq&L_\psi\mathbf{E}[D_n,D_n']\sup_{h\in\pm g\mathcal{F}}\frac{1}{n}\sum_{i=1}^n(h(x_i)-h(x_i')) &(4.10)\\
=&L_\psi\mathbf{E}[D_n,D_n',\boldsymbol{\sigma}]\sup_{h\in\pm g\mathcal{F}}\frac{1}{n}\sum_{i=1}^n\sigma_i(h(x_i)-h(x_i'))\\
=&2L_\psi\mathbf{E}[D_n,\boldsymbol{\sigma}]\sup_{h\in\pm g\mathcal{F}}\frac{1}{n}\sum_{i=1}^n\sigma_i h(x_i)\\
=&2L_\psi\mathbf{E}[D_n,\boldsymbol{\sigma}]\sup_{h\in g\mathcal{F}}\frac{1}{n}\sum_{i=1}^n\sigma_i h(x_i)\\
=&2L_\psi\mathcal{R}_n(g\mathcal{F})
\end{aligned}$$

The symmetrization technique (in the Glivenko–Cantelli theorem) was used to derive Eq (4.8). The inequality in Eq (4.9) is derived using the Jensen inequality and the convexity of $\sup(\cdot)$. The inequality in Eq (4.10) holds because $\psi(\cdot)$ is L_ψ-Lipschitz. Hence, with probability at least $1-\delta/2$,

$$C_\psi^\pm(f,g)\leq C_{n,\psi}^\pm(f,g)+2L_\psi\mathcal{R}_n(g\mathcal{F})+c\sqrt{\frac{\ln(2/\delta)}{2n}} \tag{4.11}$$

If $C_{n,\psi}^\pm(f,g)\leq C_{n,\psi}^\mp(f,g)$ holds, we can show that the following bound holds with probability at least $1-\delta/2$ in a similar manner:

$$C_\psi^\pm(f,g)\leq C_{n,\psi}^\mp(f,g)+2L_\psi\mathcal{R}_n(g\mathcal{F})+c\sqrt{\frac{\ln(2/\delta)}{2n}} \tag{4.12}$$

Combining Eq (4.11) and Eq (4.12), with probability at least $1-\delta$,

$$\begin{aligned} C_\psi(f,g) =& \max(C_\psi^+(f,g), C_\psi^-(f,g)) \\ \leq& C_{n,\psi}(f,g) + 2L_\psi \mathcal{R}_n(g\mathcal{F}) + c\sqrt{\frac{\ln(2/\delta)}{2n}}. \end{aligned}$$

□

Appendix 4.C Proof of Corollary 1

Proof. Using conditions of (A), the upper bound of the objective function of NERM with respect to $\hat{f}$ is given as follows:

$$\begin{aligned} R_n(\hat{f}) + \Omega(\hat{f}) + \eta C_{n,\psi}(\hat{f},g) \leq& R_n(f_0) + \Omega(f_0) + \eta C_{n,\psi}(f_0,g) \\ =& \phi(0) + \eta\psi(0). \end{aligned}$$

Since $R_n(f) \geq 0$ and $\Omega(f) \geq 0$, we have

$$\begin{aligned} \eta C_{n,\psi}(\hat{f},g) \leq& \phi(0) + \eta\psi(0) \\ C_{n,\psi}(\hat{f},g) \leq& \psi(0) + \phi(0)\frac{1}{\eta}. \end{aligned}$$

□

Appendix 4.D Optimization of Primal Neutral SVM

Neutral SVM is formulated as the following optimization problem

$$\min_{\boldsymbol{w}\in\mathbb{R}^d, b\in\mathbb{R}} \Psi(\boldsymbol{w},b) = \sum_{i=1}^n \max(0, 1 - y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b)) + \frac{1}{2}\|\boldsymbol{w}\|_2^2 + \eta C_{n,\psi}(\boldsymbol{w},b,g)$$

where

$$\begin{aligned} C_{n,\psi}(\boldsymbol{w},b,g) &= \max(C_{n,\psi}^+(\boldsymbol{w},b,g), C_{n,\psi}^-(\boldsymbol{w},b,g)), \\ C_{n,\psi}^\pm(\boldsymbol{w},b,g) &= \sum_{i=1}^n \max(0, 1 \mp g(\boldsymbol{x}_i)(\boldsymbol{w}^T\boldsymbol{x}_i + b)). \end{aligned}$$

Since the problem of Eq (4.7) can be solved by applying the subgradient method (Shor et al. 1985), we provide the subgradient of the objective function of Eq (4.7).

For convenience, we introduce the following notations. For a set $C \subseteq \mathbb{R}^d \times \mathbb{R}$ and constant $a \in \mathbb{R}$, we denote $aC = \{(a\boldsymbol{w}, ab) | (\boldsymbol{w}, b) \in C\}$. For any sets $C_1 \subseteq \mathbb{R}^d \times \mathbb{R}$ and $C_2 \subseteq \mathbb{R}^d \times \mathbb{R}$, let $C_1 + C_2 = \{(\boldsymbol{w}_1 + \boldsymbol{w}_2, b_1 + b_2) | (\boldsymbol{w}_1, b_1) \in C_1, (\boldsymbol{w}_2, b_2) \in C_2\}$, and let $\mathbf{Co}\ C_1 \cup C_2$ be convex hull of C_1 and C_2, i.e., $\mathbf{Co}\ C_1 \cup C_2 = \{\alpha c_1 + (1-\alpha)c_2 | c_1 \in C_1, c_2 \in C_2, \alpha \in [0,1]\}$.

The subgradient of the objective function of Eq (4.7) is derived as follows:

$$\partial \Psi(\boldsymbol{w}, b) = \sum_{i=1}^{n} \partial \ell(y_i, \boldsymbol{w}^T \boldsymbol{x}_i + b) + \lambda \{(\boldsymbol{w}, 0)\} + \eta \partial C_{n,\psi}(\boldsymbol{w}, b, g)$$

where

$$\partial \ell(t_i, \boldsymbol{w}^T \boldsymbol{x}_i + b) = \begin{cases} \{(-t_i \boldsymbol{x}_i, 1)\} & \text{if } \boldsymbol{w}^T \boldsymbol{x}_i + b < 1, \\ \{(\mathbf{0}, 0)\} & \text{if } \boldsymbol{w}^T \boldsymbol{x}_i + b > 1, \\ \{(-\alpha t_i \boldsymbol{x}_i, \alpha) | \alpha \in [0,1]\} & \text{if } \boldsymbol{w}^T \boldsymbol{x}_i + b = 1, \end{cases} \tag{4.13}$$

$$\partial C_{n,\psi}(\boldsymbol{w}, b, g) = \begin{cases} \partial C^+_{n,\psi}(\boldsymbol{w}, b, g) & \text{if } C^+_{n,\psi}(\boldsymbol{w}, b, g) > C^-_{n,\psi}(\boldsymbol{w}, b, g), \\ \partial C^-_{n,\psi}(\boldsymbol{w}, b, g) & \text{if } C^+_{n,\psi}(\boldsymbol{w}, b, g) < C^-_{n,\psi}(\boldsymbol{w}, b, g), \\ \mathbf{Co}(\partial C^+_{n,\psi}(\boldsymbol{w}, b, g) \cup \partial C^-_{n,\psi}(\boldsymbol{w}, b, g)) & \text{if } C^+_{n,\psi}(\boldsymbol{w}, b, g) = C^-_{n,\psi}(\boldsymbol{w}, b, g), \end{cases} \tag{4.14}$$

$$\partial C^{\pm}_{n,\psi}(\boldsymbol{w}, b, g) = \sum_{i=1}^{n} \partial \ell(\pm v_i, \boldsymbol{w}^T \boldsymbol{x}_i + b). \tag{4.15}$$

Eq (4.13), Eq (4.14) and Eq (4.15) denote the subgradient of hinge loss function, $C_{n,\psi}(\boldsymbol{w}, b, g)$ and $C^{\pm}_{n,\psi}(\boldsymbol{w}, b, g)$, respectively.

Chapter 5

Minimax Optimal Additive Functional Estimation

After constructing a predictor, we want to evaluate the fairness of the predictor using the test samples. However, since the underlying distribution ρ is unknown to us, we cannot evaluate populational fairness directly and need to estimate it by using the test samples. We refer to this problem as populational fairness evaluation. Populational fairness evaluation can be reduced to the entropy estimation problem. For example, suppose we employ the mutual information as a fairness measure, which can be decomposed as

$$I(Y;V) = H(Y|V) - H(V),$$

where H denotes the Shannon entropy. Since the right term $H(V)$ is invariant by a change of the predictor, we can use the conditional entropy $H(Y|V)$ as a fairness measure of the predictor.

The goal of the entropy estimation is to estimate the value of the entropy using the samples efficiently. Eventually, we want to construct the *optimal estimator*, which can most accurately estimate the value of the entropy among all estimators. In this chapter, we deal with a problem of estimating the *additive function*, which is a criterion of the discrete distribution with a certain form and covers the estimation problems of the most entropy measure and aims to construct the optimal estimator of the additive function regarding the minimax optimality.

5.1 Additive Functional Estimation

Let P be a probability measure with alphabet size k, and X be a discrete random variable drawn from P. Without loss of generality, we can assume that the domain of P is $[k]$, where we denote $[m] = \{1, ..., m\}$ for a positive integer m. We use a vector representation of P; $P = (p_1, ..., p_k)$ where $p_i = P\{X = i\}$. Let ϕ be a mapping from $[0, 1]$ to $\mathbb{R}^+$. Given a set of i.i.d. samples $S_n = \{X_1, ..., X_n\}$ from P, we deal with the problem of estimating an *additive functional* of ϕ. The additive functional θ of ϕ is defined as

$$\theta(P;\phi) = \sum_{i=1}^{k} \phi(p_i).$$

We simplify this notation to $\theta(P;\phi) = \theta(P)$. Most entropy-like criteria can be formed in terms of θ. For instance, when $\phi(p) = -p \ln p$, θ is Shannon entropy. For a positive real α, letting $\phi(p) = p^\alpha$, $\ln(\theta(P))/(1-\alpha)$ becomes Rényi entropy. More generally, letting $\phi = f$ where f is a concave function, θ becomes f-entropies (Akaike 1998).

Techniques for the estimation of the entropy-like criteria have been considered in various fields, including physics (Lake et al. 2011), neuroscience (Nemenman et al. 2004), and security (Gu et al. 2005). In machine learning, methods that involve entropy estimation were introduced for decision-trees (Quinlan 1986), feature selection (Peng et al. 2005), and clustering (Dhillon et al. 2003). For example, the decision-tree learning algorithms, i.e., ID3, C4.5, and C5.0 construct a decision tree in which the criteria for the tree splitting are defined based on Shannon entropy (Quinlan 1986). Similarly, information theoretic feature selection algorithms evaluate the relevance between the features and the target using the entropy (Peng et al. 2005).

The goal of this study is to derive the minimax optimal estimator of θ given a function ϕ. For the precise definition of the minimax optimality, we introduce the minimax risk. A sufficient statistic of P is a histogram $N = (N_1, ..., N_k)$, where $N_j = \sum_{i=1}^{n} \mathbf{1}_{\{X_i=j\}}$ and $N \sim \text{Multinomial}(n, P)$. The estimator of θ is defined as a function $\hat{\theta} : [n]^k \to \mathbb{R}$. Then, the quadratic minimax risk is defined as

$$R^*(n, k; \phi) = \inf_{\hat{\theta}} \sup_{P \in \mathcal{M}_k} \mathbf{E}\left[\left(\hat{\theta}(N) - \theta(P)\right)^2\right],$$

where $\mathcal{M}_k$ is the set of all probability measures on $[k]$, and the infimum is taken over all estimators $\hat{\theta}$. With this definition of the minimax risk, an estimator $\hat{\theta}$ is minimax (rate-)optimal if there exists a positive constant C such that

$$\sup_{P \in \mathcal{M}_k} \mathbf{E}\left[\left(\hat{\theta}(N) - \theta(P)\right)^2\right] \leq C R^*(n, k; \phi).$$

A natural estimator of θ is the plugin or the maximum likelihood estimator, in which the estimated value is obtained by substituting the empirical mean of the probabilities P into θ. However, the estimator has a large bias for large k. Indeed, the plugin estimators for $\phi(p) = -p \ln p$ and $\phi(p) = p^\alpha$ have been shown to be suboptimal in the large-k regime in recent studies (Acharya et al. 2015; Jiantao Jiao et al. 2015; Yihong Wu et al. 2016).

Recent studies investigated the minimax optimal estimators for $\phi(p) = -p \ln p$ and $\phi(p) = p^\alpha$ in the large-k regime (Acharya et al. 2015; Jiantao Jiao et al. 2015; Yihong Wu et al. 2016). However, the results of these studies were only derived for these ϕ. Jiantao Jiao et al. (2015) suggested that the estimator is easily extendable to the general additive functional, although they did not prove the minimax optimality.

In this paper, we propose a minimax optimal estimator for the estimation problem of the additive functional θ for general ϕ under certain conditions on the smoothness. Our estimator achieves the minimax optimal rate even in the large-k regime for $\phi \in C^4[0,1]$ such that $\left|\phi^{(4)}(p)\right|$ is finite for $p \in (0,1]$, where $C^4[0,1]$ denotes a class of four times differentiable functions from $[0,1]$ to $\mathbb{R}$. For such ϕ, we reveal a property of ϕ which can substantially influence the minimax optimal rate.

Related work. The simplest way to estimate θ is to use the so-called plugin estimator or the maximum likelihood estimator, in which the empirical probabilities are substituted into θ as P. Letting $\tilde{P} = (\hat{p}_1, ..., \hat{p}_k)$ and $\hat{p}_i = N_i/n$, the plugin estimator is defined as

$$\theta_{\text{plugin}}(N) = \theta(\tilde{P}).$$

The plugin estimator is asymptotically consistent under weak assumptions for fixed k (Antos et al. 2001). However, this is not true for the large-k regime. Indeed, Jiantao Jiao et al. (2015) and Yihong Wu et al. (2016) derived a lower bound for the quadratic risk for the plugin estimator of $\phi(p) = p\ln(1/p)$ and $\phi(p) = p^\alpha$. In the case of Shannon entropy, the lower bound is given as

$$\sup_{P \in \mathcal{M}_k} \mathbf{E}\left[\left(\theta_{\text{plugin}}(N) - \theta(P)\right)^2\right] \geq C\left(\frac{k^2}{n^2} + \frac{\ln^2 k}{n}\right),$$

where C denotes a universal constant. The first term k^2/n^2 comes from the bias and it indicates that if k grows linearly with respect to n, the plugin estimator becomes inconsistent. This means the plugin estimator is suboptimal in the large-k regime. Bias-correction methods, such as (Grassberger 1988; Miller 1955; Zahl 1977), can be applied to the plugin estimator of $\phi(p) = -p\ln p$ to reduce the bias whereas these bias-corrected estimators are still suboptimal. The estimators based on Bayesian approaches in (Holste et al. 1998; Schober 2013; Schürmann et al. 1996) are also suboptimal (Yanjun Han et al. 2015).

Many researchers have studied estimators that can consistently estimate the additive functional with sublinear samples with respect to the alphabet size k to derive the optimal estimator in the large-k regime. The existence of consistent estimators even with sublinear samples were first revealed in Paninski (2004), but an explicit estimator was not provided. Valiant et al. (2011a) introduced an estimator based on linear programming that consistently estimates $\phi(p) = -p\ln p$ with sublinear samples. However, the estimator of (Valiant et al. 2011a) has not been shown to achieve the minimax rate even in a more detailed analysis in (Valiant et al. 2011b). Recently, Acharya et al.

(2015) showed that the bias-corrected estimator of Rényi entropy achieves the minimax optimal rate in regard to the sample complexity if $\alpha > 1$ and $\alpha \in \mathbb{N}$, but they did not show the minimax optimality for other α. Jiantao Jiao et al. (2015) introduced a minimax optimal estimator for $\phi(p) = -p \ln p$ for any $\alpha \in (0, 3/2)$ in the large-k regime. Y. Wu et al. (2015) derived a minimax optimal estimator for $\phi(p) = \mathbf{1}_{p>0}$. For $\phi(p) = -p \ln p$, Jiantao Jiao et al. (2015) and Yihong Wu et al. (2016) independently introduced the minimax optimal estimators in the large-k regime. In the case of Shannon entropy, the optimal rate was obtained as

$$\frac{k^2}{(n \ln n)^2} + \frac{\ln^2 k}{n}.$$

The first term indicates that the introduced estimator can consistently estimate Shannon entropy if $n \geq Ck/\ln k$.

The estimators introduced by Acharya et al. (2015), Jiantao Jiao et al. (2015), and Yihong Wu et al. (2016) are composed of two estimators: the bias-corrected plugin estimator and the best polynomial estimator. The bias-corrected plugin estimator is composed of the sum of the plugin estimator and a bias-correction term which offsets the second-order approximation of the bias as in (Miller 1955). The best polynomial estimator is an unbiased estimator of the polynomial that best approximates ϕ in terms of the uniform error. Specifically, the best approximation for the polynomial of ϕ in an interval $I \subseteq [0, 1]$ is the polynomial g that minimizes $\sup_{x \in I} |\phi(x) - g(x)|$. Jiantao Jiao et al. (2015) suggested that this estimator can be extended for the general additive functional θ. However, the minimax optimality of the estimator was only proved for specific cases of ϕ, including $\phi(p) = -p \ln p$ and $\phi(p) = p^\alpha$. Thus, to prove the minimax optimality for other ϕ, we need to individually analyze the minimax optimality for specific ϕ. Here, we aim to clarify which property of ϕ substantially influences the minimax optimal rate when estimating the additive functional.

Besides, the optimal estimators for divergences with large alphabet size have been investigated in (Bu et al. 2016; Y. Han et al. 2016; J. Jiao et al. 2016). The estimation problems of divergences are much complicated than the additive function, while the similar techniques were applied to derive the minimax optimality.

Our contributions. In this paper, we propose the minimax optimal estimator for $\theta(P; \phi)$. We reveal that the *divergence speed* of the fourth derivative of ϕ plays an important role in characterizing the minimax optimal rate. Informally, for $\beta > 0$, the meaning of "the divergence speed of a function

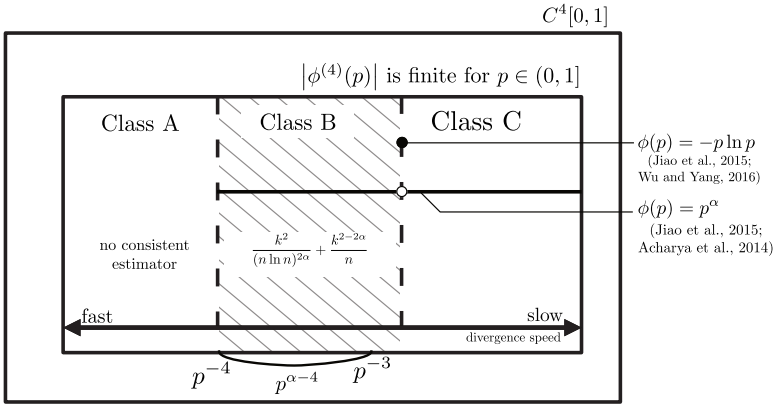


Figure 5.1: Relationship between the divergence speed of the fourth derivative of ϕ and the minimax optimality of the estimation problem of $\theta(P;\phi)$.

$f(p)$ is $p^{-\beta}$" is that $|f(p)|$ goes to infinity at the same speed as $p^{-\beta}$ when p approaches 0. When the divergence speed of the fourth derivative of $\phi(p)$ is $p^{-\beta}$, the fourth derivative of ϕ diverges faster as β increases.

Our results are summarized in Figure 5.1. Figure 5.1 illustrates the relationship between the divergence speed of the fourth derivative of ϕ and the minimax optimality of the estimation problem of $\theta(P;\phi)$. In Figure 5.1, the outermost rectangle represents the space of the four times continuous differentiable functions $C^4[0,1]$. The innermost rectangle denotes the subset class of $C^4[0,1]$ such that the absolute value of its fourth derivative $\left|\phi^{(4)}(p)\right|$ is finite for any $p \in (0,1]$. In this subclass of ϕ, the horizontal direction represents the divergence speed of the fourth derivative of ϕ, in which a faster ϕ is on the left-hand side and a slower ϕ is on the right-hand side. The ϕ with an explicit form and divergence speed is denoted by a point in the rectangle. For example, the black circle denotes $\phi(p) = -p \ln p$ where the divergence speed of the fourth derivative of this ϕ is p^{-3}. Class B denotes a set of any function ϕ such that the divergence speed of the fourth derivative is $p^{\alpha-4}$ where $\alpha \in (0,1)$. As already discussed, existing methods have achieved minimax optimality in the large-k regime for specific ϕ, including $\phi(p) = -p \ln p$ (black circle in Figure 5.1) and $\phi(p) = p^\alpha$ (middle line in Figure 5.1 where the white circle denotes that there is no $\alpha > 0$ such that the divergence speed is p^{-3}).

We investigate the minimax optimality of the estimation problem of θ for ϕ in Class A and Class B. Class A is a class of ϕ such that the divergence speed of the fourth derivative is faster than p^{-4}. Class B is a class of ϕ such

that the divergence speed of the fourth derivative is $p^{\alpha-4}$ where $\alpha \in (0,1)$. In Class A, we show that we cannot construct a consistent estimator of θ for any ϕ in Class A (the leftmost hatched area in Figure 5.1, Proposition 3). In other words, the minimax optimal rate is larger than constant order if the divergence speed of the fourth derivative is faster than p^{-4}. Thus, there is no need to derive the minimax optimal estimator in Class A.

Also, we derive the minimax optimal estimator for any ϕ in Class B (the middle hatched area in Figure 5.1, Theorem 7). For example, $\phi(p) = p^\alpha$ (Réyni entropy case), $\phi(p) = \cos(cp)p^\alpha$, and $\phi(p) = e^{cp}p^\alpha$ for $\alpha \in (0,1)$ include the coverage of our estimator, where c is a universal constant. Intuitively, since the large derivative makes the estimation problem θ more difficult, the minimax rate decreases if the derivative of ϕ diverges faster. Our minimax optimal rate reflects this behavior. For ϕ in Class B, the minimax optimal rate is obtained as

$$\frac{k^2}{(n\ln n)^{2\alpha}} + \frac{k^{2-2\alpha}}{n},$$

where $k \gtrsim \ln^{\frac{4}{3}} n$ if $\alpha \in (0, 1/2]$. We can clearly see that this rate decreases for larger α, i.e., a slower divergence speed.

Currently, the minimax optimality of ϕ in Class C is an open problem. However, we provide a notable discussion in Section 5.3.

5.2 Preliminaries

Notations. We now introduce some additional notations. For any positive real sequences $\{a_n\}$ and $\{b_n\}$, $a_n \gtrsim b_n$ denotes that there exists a positive constant c such that $a_n \geq cb_n$. Similarly, $a_n \lesssim b_n$ denotes that there exists a positive constant c such that $a_n \leq cb_n$. Furthermore, $a_n \asymp b_n$ implies $a_n \gtrsim b_n$ and $a_n \lesssim b_n$. For an event $\mathcal{E}$, we denote its complement by $\mathcal{E}^c$. For two real numbers a and b, $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. For a function $\phi : \mathbb{R} \to \mathbb{R}$, we denote its i-th derivative as $\phi^{(i)}$.

Poisson sampling. We employ the Poisson sampling technique to derive upper and lower bounds for the minimax risk. The Poisson sampling technique models the samples as independent Poisson distributions, while the original samples follow a multinomial distribution. Specifically, the sufficient statistic for P in the Poisson sampling is a histogram $\tilde{N} = (\tilde{N}_i, ..., \tilde{N}_k)$, where $\tilde{N}_1, ..., \tilde{N}_k$ are independent random variables such that $\tilde{N}_i \sim \text{Poi}(np_i)$. The minimax

risk for Poisson sampling is defined as follows:

$$\tilde{R}^*(n,k;\phi) = \inf_{\{\hat{\theta}\}} \sup_{P\in\mathcal{M}_k} \mathbf{E}\left[\left(\hat{\theta}(\tilde{N}) - \theta(P)\right)^2\right].$$

The minimax risk of Poisson sampling well approximates that of the multinomial distribution. Indeed, Jiantao Jiao et al. (2015) presented the following lemma.

Lemma 3 (Jiantao Jiao et al. (2015))**.** *The minimax risk under the Poisson model and the multinomial model are related via the following inequalities:*

$$\tilde{R}^*(2n,k;\phi) - \sup_{P\in\mathcal{M}_k} |\theta(P)|e^{-n/4} \leq R^*(n,k;\phi) \leq 2\tilde{R}^*(n/2,k;\phi).$$

Lemma 3 states $R^*(n,k;\phi) \asymp \tilde{R}^*(n,k;\phi)$, and thus we can derive the minimax rate of the multinomial distribution from that of the Poisson sampling.

Best polynomial approximation. Acharya et al. (2015), Jiantao Jiao et al. (2015), and Yihong Wu et al. (2016) presented a technique of the best polynomial approximation for deriving the minimax optimal estimators and their lower bounds for the risk. Let $\mathcal{P}_L$ be the set of polynomials of degree L. Given a function ϕ, a polynomial p, and an interval $I \subseteq [0,1]$, the uniform error between ϕ and p on I is defined as

$$\sup_{x\in I} |\phi(x) - p(x)|. \tag{5.1}$$

The best polynomial approximation of ϕ by a degree-L polynomial with a uniform error is achieved by the polynomial $p \in \mathcal{P}_L$ that minimizes Eq (5.1). The error of the best polynomial approximation is defined as

$$E_L(\phi, I) = \inf_{p\in\mathcal{P}_L} \sup_{x\in I} |\phi(x) - p(x)|.$$

The error rate with respect to the degree L has been studied since the 1960s (Achieser 2013; Ditzian et al. 2012; Petrushev et al. 2011; Timan et al. 1965). The polynomial that achieves the best polynomial approximation can be obtained, for instance, by the Remez algorithm (Remez 1934) if I is bounded.

5.3 Main results

Suppose ϕ is four times continuously differentiable on $(0,1]$[1]. We reveal that the *divergence speed* of the fourth derivative of ϕ plays an important role for the minimax optimality of the estimation problem of the additive functional. Formally, the divergence speed is defined as follows.

Definition 8 (divergence speed)**.** For an integer $m \geq 1$, let ϕ be an m times continuously differentiable function on $(0,1]$. For $\beta > 0$, the divergence speed of the mth derivative of ϕ is $p^{-\beta}$ if there exist finite constants $W > 0$, c_m, and c'_m such that for all $p \in (0,1]$

$$\left|\phi^{(m)}(p)\right| \leq \beta_{m-1} W p^{-\beta} + c_m, \text{ and } \left|\phi^{(m)}(p)\right| \geq \beta_{m-1} W p^{-\beta} + c'_m,$$

where $\beta_m = \prod_{i=1}^{m}(i - m + \beta)$.

A larger β implies faster divergence. We analyze the minimax optimality for two cases: the divergence speed of the fourth derivative of ϕ is i) larger than p^{-4} (Class A), and ii) $p^{\alpha-4}$ (Class B), for $\alpha \in (0,1)$.

Minimax optimality for Class A. We now demonstrate that we cannot construct a consistent estimator for any n and $k \geq 3$ if the divergence speed of ϕ is larger than p^{-4}.

Proposition 3. *Let ϕ be a continuously differentiable function on $(0,1]$. If there exists finite constants $W > 0$ and c'_1 such that for $p \in (0,1]$*

$$\left|\phi^{(1)}(p)\right| \geq W p^{-1} + c'_1,$$

then there is no consistent estimator, i.e., $R^*(n,k;\phi) \gtrsim 1$.

The proof of Proposition 3 is given in Section 5.D. From Lemma 17, the divergence speed of the first derivative is p^{-1} if that of the fourth derivative is p^{-4}. Thus, if the divergence speed of ϕ is greater than p^{-4}, we cannot construct an estimator that consistently estimates θ for any probability measure $P \in \mathcal{M}_k$. Consequently, there is no need to derive the minimax optimal estimator in this case.

Minimax optimality for Class B. We derive the minimax optimal rate for ϕ in which the divergence speed of its fourth derivative is $p^{\alpha-4}$ for $\alpha \in (0,1)$. Thus, we make the following assumption.

Assumption 1. Suppose ϕ is four times continuously differentiable on $(0,1]$. For $\alpha \in (0,1)$, the divergence speed of the fourth derivative of ϕ is $p^{\alpha-4}$.

[1]We say that a function $\phi : [0,1] \to \mathbb{R}_+$ is differentiable at 1 if $\lim_{h \to -0} \frac{\phi(1+h)-\phi(1)}{h}$ exists.

Note that a set of ϕ satisfying Assumption 1 is Class B depicted in Figure 5.1. The divergence speed increases as α decreases. Under Assumption 1, we derive the minimax optimal estimator of which the minimax rate is given by the following theorems.

Theorem 7. *Under Assumption 1 with $\alpha \in (0, 1/2]$, if $n \gtrsim \frac{k^{1/\alpha}}{\ln k}$ and $k \gtrsim \ln^{\frac{4}{3}} n$,*

$$R^*(n, k; \phi) \asymp \frac{k^2}{(n \ln n)^{2\alpha}}.$$

Otherwise, there is no consistent estimator, i.e., $R^*(n, k; \phi) \gtrsim 1$.

Theorem 8. *Under Assumption 1 with $\alpha \in (1/2, 1)$, if $n \gtrsim \frac{k^{1/\alpha}}{\ln k}$*

$$R^*(n, k; \phi) \asymp \frac{k^2}{(n \ln n)^{2\alpha}} + \frac{k^{2-2\alpha}}{n}.$$

Otherwise, there is no consistent estimator, i.e., $R^*(n, k; \phi) \gtrsim 1$.

Theorems 7 and 8 are proved by combining the results in Sections 5.6 and 5.7. The minimax optimal rate in Theorems 7 and 8 are characterized by the parameter for the divergence speed α from Assumption 1. From Theorems 7 and 8, we can conclude that the minimax optimal rate decreases as the divergence speed increases.

The explicit estimator that achieves the optimal minimax rate shown in Theorems 7 and 8 are described in the next section.

Remark. Assumption 1 covers $\phi(p) = p^\alpha$ for $\alpha \in (0, 1)$, but does not for all existing works. For $\phi(p) = -p \ln(p)$ and $\phi(p) = p^\alpha$ with $\alpha \geq 1$, the divergence speed of these ϕ is lower than $p^{\alpha-4}$ for $\alpha \in (0, 1)$. Indeed, the divergence speed of $\phi(p) = -p \ln(p)$ and $\phi(p) = p^\alpha$ for $\alpha \geq 1$ are p^{-3} and $p^{\alpha-4}$, respectively. We can expect that the corresponding minimax rate is characterized by the divergence speed even when the divergence speed is lower than $p^{\alpha-4}$ for $\alpha \in (0, 1)$. The analysis of the minimax rate for lower divergence speeds remains an open problem.

5.4 Estimator for θ

In this section, we describe our estimator for θ in detail. Our estimator is composed of the bias-corrected plugin estimator and the best polynomial estimator. We first describe the overall estimation procedure on the supposition that the bias-corrected plugin estimator and the best polynomial estimator

are black boxes. Then, we describe the bias-corrected plugin estimator and the best polynomial estimator in detail.

For simplicity, we assume the samples are drawn from the Poisson sampling model, where we first draw $n' \sim \mathrm{Poi}(2n)$, and then draw n' i.i.d. samples $S_{n'} = \{X_1, ..., X_{n'}\}$. Given the samples $S_{n'}$, we first partition the samples into two sets. We use one set of the samples to determine whether the bias-corrected plugin estimator or the best polynomial estimator should be employed, and the other set to estimate θ. Let $\{B_i\}_{i=1}^{n'}$ be i.i.d. random variables drawn from the Bernoulli distribution with the parameter $1/2$, i.e., $\mathbb{P}\{B_i = 0\} = \mathbb{P}\{B_i = 1\} = 1/2$ for $i = 1, ..., n'$. We partition $(X_1, ..., X_{n'})$ according to $(B_1, ..., B_{n'})$, and construct the histograms $\tilde{N}$ and $\tilde{N}'$, which are defined as

$$\tilde{N}_i = \sum_{j=1}^{n'} \mathbf{1}_{X_j=i}\mathbf{1}_{B_j=0}, \quad \tilde{N}'_i = \sum_{j=1}^{n'} \mathbf{1}_{X_j=i}\mathbf{1}_{B_j=1}, \text{ for } i \in [n'].$$

Then, $\tilde{N}$ and $\tilde{N}'$ are independent histograms, and $\tilde{N}_i, \tilde{N}'_i \sim \mathrm{Poi}(np_i)$.

Given $\tilde{N}'$, we determine whether the bias-corrected plugin estimator or the best polynomial estimator should be employed for each alphabet. Let $\Delta_{n,k}$ be a threshold depending on n and k to determine which estimator is employed, which will be specified as in Theorem 11 on page 87. We apply the best polynomial estimator if $\tilde{N}'_i < 2\Delta_{n,k}$, and otherwise, i.e., $\tilde{N}'_i \geq 2\Delta_{n,k}$, we apply the bias-corrected plugin estimator. Let ϕ_{poly} and ϕ_{plugin} be the best polynomial estimator and the bias-corrected plugin estimator for ϕ, respectively. Then, the estimator of θ is written as

$$\tilde{\theta}(\tilde{N}) = \sum_{i=1}^{k} \left(\mathbf{1}_{\tilde{N}'_i \geq 2\Delta_{n,k}} \phi_{\mathrm{plugin}}(\tilde{N}_i) + \mathbf{1}_{\tilde{N}'_i < 2\Delta_{n,k}} \phi_{\mathrm{poly}}(\tilde{N}_i) \right).$$

Finally, we truncate $\tilde{\theta}$ so that the final estimate is not outside of the domain of θ.

$$\hat{\theta}(\tilde{N}) = (\tilde{\theta}(\tilde{N}) \wedge \theta_{\mathrm{sup}}) \vee \theta_{\mathrm{inf}},$$

where $\theta_{\mathrm{inf}} = \inf_{P \in \mathcal{M}_k} \theta(P)$ and $\theta_{\mathrm{sup}} = \sup_{P \in \mathcal{M}_k} \theta(P)$. Next, we describe the details of the best polynomial estimator ϕ_{poly} and the bias-corrected plugin estimator ϕ_{plugin}.

Best polynomial estimator. The best polynomial estimator is an unbiased estimator of the polynomial that provides the best approximation of ϕ. Let $\{a_m\}_{m=0}^{L}$ be coefficients of the polynomial that achieves the best approximation of ϕ by a degree-L polynomial with range $I = [0, \frac{4\Delta_{n,k}}{n}]$, where L is as

specified in Theorem 11 on page 87. Then, the approximation of ϕ by the polynomial at point p_i is written as

$$\phi_L(p_i) = \sum_{m=0}^{L} a_m p_i^m. \tag{5.2}$$

From Eq (5.2), an unbiased estimator of ϕ_L can be derived from an unbiased estimator of p_i^m. For the random variable $\tilde{N}_i$ drawn from the Poisson distribution with the mean parameter np_i, the expectation of the mth factorial moment $(\tilde{N}_i)_m = \frac{\tilde{N}_i!}{(\tilde{N}_i - m)!}$ becomes $(np_i)^m$. Thus, $\frac{(\tilde{N}_i)_m}{n^m}$ is an unbiased estimator of p_i^m. Substituting this into Eq (5.2) gives the unbiased estimator of $\phi_L(p_i)$ as

$$\bar{\phi}_{\mathrm{poly}}(\tilde{N}_i) = \sum_{m=0}^{L} \frac{a_m}{n^j} (\tilde{N}_i)_m.$$

Next, we truncate $\bar{\phi}_{\mathrm{poly}}$ so that it is not outside of the domain of $\phi(p)$. Let $\phi_{\mathrm{inf}, \frac{\Delta_{n,k}}{n}} = \inf_{p \in [0, \frac{\Delta_{n,k}}{n}]} \phi(p)$ and $\phi_{\mathrm{sup}, \frac{\Delta_{n,k}}{n}} = \sup_{p \in [0, \frac{\Delta_{n,k}}{n}]} \phi(p)$. Then, the best polynomial estimator is defined as

$$\phi_{\mathrm{poly}}(\tilde{N}_i) = (\bar{\phi}_{\mathrm{poly}}(\tilde{N}_i) \wedge \phi_{\mathrm{sup}, \frac{\Delta_{n,k}}{n}}) \vee \phi_{\mathrm{inf}, \frac{\Delta_{n,k}}{n}}.$$

Bias-corrected plugin estimator. In the bias-corrected plugin estimator, we apply the bias correction of (Miller 1955). Applying the second-order Taylor expansion to the bias of the plugin estimator gives

$$\begin{aligned} \mathbf{E}\left[\phi\left(\frac{\tilde{N}_i}{n}\right) - \phi(p_i)\right] \approx & \mathbf{E}\left[\phi^{(1)}(p_i)\left(\frac{\tilde{N}_i}{n} - p_i\right) + \frac{\phi^{(2)}(p_i)}{2}\left(\frac{\tilde{N}_i}{n} - p_i\right)^2\right] \\ = & \frac{p_i \phi^{(2)}(p_i)}{2n}. \end{aligned}$$

Thus, we include $-\frac{\tilde{N}_i \phi^{(2)}(\tilde{N}_i/n)}{2n^2}$ as a bias-correction term in the plugin estimator $\phi(\tilde{N}_i/n)$, which offsets the second-order approximation of the bias. However, we do not directly apply the bias-corrected plugin estimator to estimate $\phi(p_i)$ for two reasons. First, the derivative of $\phi(p)$ is large when p approches 0, which results in a large bias. Second, $\phi(p)$ for $p > 1$ is undefined even though $\tilde{N}_i/n$ can exceed 1. Thus, we apply the bias-corrected plugin estimator to the

function $\bar{\phi}_{\frac{\Delta_{n,k}}{n}}$ defined below instead of ϕ. Define

$$
\begin{aligned}
&H_L(p;\phi,a,b)\\
=&\phi(a)+\sum_{m=1}^{L}\frac{\phi^{(m)}(a)}{m!}(p-a)^m(p-b)^{L+1}\sum_{\ell=0}^{L-m}\frac{(-1)^\ell(L+\ell)!}{\ell!L!}(a-b)^{-L-1-\ell}(p-a)^\ell\\
=&\phi(a)+\sum_{m=1}^{L}\frac{\phi^{(m)}(a)}{m!}(p-a)^m\sum_{\ell=0}^{L-m}\frac{L+1}{L+\ell+1}\,\mathrm{B}_{\ell,L+\ell+1}\left(\frac{p-a}{b-a}\right),
\end{aligned}
$$

where $\mathrm{B}_{\nu,n}(x)=\binom{n}{\nu}x^\nu(1-x)^{n-\nu}$ denotes the Bernstein basis polynomial. Then, $H_L(p;\phi,a,b)$ denotes a function that interpolates between $\phi(a)$ and $\phi(b)$ using Hermite interpolation. From generalized Hermite interpolation (Spitzbart 1960), $H_L^{(i)}(a;\phi,a,b)=\phi^{(i)}(a)$ for $i=0,...,L$ and $H_L^{(i)}(b;\phi,a,b)=0$ for $i=1,...,L$. The function $\bar{\phi}_{\frac{\Delta_{n,k}}{n}}$ is defined as

$$
\bar{\phi}_{\frac{\Delta_{n,k}}{n}}(p)=\begin{cases}
H_4\left(\frac{\Delta_{n,k}}{2n};\phi,\frac{\Delta_{n,k}}{n},\frac{\Delta_{n,k}}{2n}\right) & \text{if } p\le\frac{\Delta_{n,k}}{2n},\\
H_4\left(p;\phi,\frac{\Delta_{n,k}}{n},\frac{\Delta_{n,k}}{2n}\right) & \text{if } \frac{\Delta_{n,k}}{2n}<p<\frac{\Delta_{n,k}}{n},\\
H_4(p;\phi,1,2) & \text{if } 1<p<2,\\
H_4(2;\phi,1,2) & \text{if } p\ge 2,\\
\phi(p) & \text{otherwise .}
\end{cases}
$$

From this definition, $\bar{\phi}_{\frac{\Delta_{n,k}}{n}}=\phi$ if $p\in[\frac{\Delta_{n,k}}{n},1]$. From Hermite interpolation, the function $\bar{\phi}_{\frac{\Delta_{n,k}}{n}}$ is four times differentiable on $\mathbb{R}_+$ and $\bar{\phi}^{(1)}_{\frac{\Delta_{n,k}}{n}}(p)=...=\bar{\phi}^{(4)}_{\frac{\Delta_{n,k}}{n}}(p)=0$ for $p\le\frac{\Delta_{n,k}}{2n}$ and $p\ge 2$. By introducing $\bar{\phi}_{\frac{\Delta_{n,k}}{n}}$, we can bound the fourth derivative of $\bar{\phi}_{\frac{\Delta_{n,k}}{n}}$ using $\Delta_{n,k}$, and this enables us to control the bias with the threshold parameter $\Delta_{n,k}$. Using $\bar{\phi}_{\frac{\Delta_{n,k}}{n}}$ instead of ϕ yields the bias-corrected plugin estimator

$$
\phi_{\text{plugin}}(\tilde{N}_i)=\bar{\phi}_{\frac{\Delta_{n,k}}{n}}\left(\frac{\bar{N}_i}{n}\right)-\frac{\tilde{N}_i}{2n^2}\bar{\phi}^{(2)}_{\frac{\Delta_{n,k}}{n}}\left(\frac{\bar{N}_i}{n}\right). \tag{5.3}
$$

5.5 Remark about Differentiability for Analysis

Why is the minimax rate characterized by the divergence speed of the *fourth* derivative? Indeed, most of the results can be obtained with a weaker assumption compared to Assumption 1 regarding differentiability, which is formally defined as follows.

Assumption 2. Suppose ϕ is two times continuously differentiable on $(0, 1]$. For $\alpha \in (0, 1)$, the divergence speed of the second derivative of ϕ is $p^{\alpha-2}$.

Assumption 2 only requires two times continuous differentiability, whereas Assumption 1 requires four times. Only the analysis of the bias-corrected plugin estimator requires Assumption 1 to achieve the minimax rate due to the bias-correction term in Eq (5.3). The bias-correction term is formed as the plugin estimator of the second derivative of ϕ, and its convergence rate is highly dependent on the smoothness of the second derivative. The smoothness of the second derivative of ϕ is characterized by the fourth derivative of ϕ, and thus Assumption 1 is required to derive the error bound of the bias-corrected plugin estimator. Another bias-correction method might weaken the assumption as in Assumption 2, but it remains as an future work.

5.6 Analysis of Lower Bound

In this section, we derive a lower bound for the minimax rate of θ. Under Assumption 2, we can derive the lower bound of the minimax risk as in the following theorem.

Theorem 9. *Under Assumption 2, for $k \geq 3$, we have*

$$R^*(n, k; \phi) \gtrsim \frac{k^{2-2\alpha}}{n}.$$

The lower bound is obtained by applying Le Cam's two-point method (see (Tsybakov 2009)). The details of the proof of Theorem 9 can be found in Section 5.B. Next, we derive another lower bound for the minimax rate.

Theorem 10. *Under Assumption 2, if $n \gtrsim \frac{k^{1/\alpha}}{\ln k}$, we have*

$$R^*(n, k; \phi) \gtrsim \frac{k^2}{(n \ln n)^{2\alpha}},$$

where we need $k \gtrsim \ln^{\frac{4}{3}} n$ if $\alpha \in (0, 1/2]$.

The proof is accomplished in the same manner as (Yihong Wu et al. 2016, Proposition 3). The details of the proof of Theorem 10 are also found in Section 5.B. Combining Theorems 9 and 10, we get the lower bounds in Theorems 7 and 8 as $R^*(n,k;\phi) \gtrsim \frac{k^2}{(n\ln n)^{2\alpha}} \vee \frac{k^{2-2\alpha}}{n} \gtrsim \frac{k^2}{(n\ln n)^{2\alpha}} + \frac{k^{2-2\alpha}}{n}$.

5.7 Analysis of Upper Bound

Here, we derive the upper bound for the worst-case risk of the estimator.

Theorem 11. *Suppose* $\Delta_{n,k} = C_2 \ln n$ *and* $L = \lfloor C_1 \ln n \rfloor$ *where* C_1 *and* C_2 *are universal constants such that* $6C_1 \ln 2 + 4\sqrt{C_1 C_2}(1+\ln 2) < 1$ *and* $C_2 > 16$. *Under Assumption 1, the worst-case risk of* $\hat{\theta}$ *is bounded above by*

$$\sup_{P\in\mathcal{M}_k} \mathbf{E}\left[\left(\hat{\theta}(\tilde{N}) - \theta(P)\right)^2\right] \lesssim \frac{k^2}{(n\ln n)^{2\alpha}} + \frac{k^{2-2\alpha}}{n},$$

where we need $k \gtrsim \ln^{\frac{4}{3}} n$ *if* $\alpha \in (0, 1/2]$.

To prove Theorem 11, we derive the bias and the variance of $\hat{\theta}$.

Lemma 4. *Given* $P \in \mathcal{M}_k$, *for* $1 \lesssim \Delta_{n,k} \le n$, *the bias of* $\hat{\theta}$ *is bounded above by*

$$\mathbf{Bias}\left[\tilde{\theta}(\tilde{N}) - \theta(P)\right] \lesssim \sum_{i=1}^{k}\Bigg((e/4)^{\Delta_{n,k}} + \mathbf{Bias}\left[\phi_{\text{plugin}}(\tilde{N}_i) - \phi(p_i)\right]\mathbf{1}_{np_i > \Delta_{n,k}} + \mathbf{Bias}\left[\phi_{\text{poly}}(\tilde{N}_i) - \phi(p_i)\right]\mathbf{1}_{np_i \le 4\Delta_{n,k}} + e^{-\Delta_{n,k}/8}\Bigg).$$

Lemma 5. *Given* $P \in \mathcal{M}_k$, *for* $1 \lesssim \Delta_{n,k} \le n$, *the variance of* $\hat{\theta}$ *is bounded above by*

$$\mathbf{Var}\left[\tilde{\theta}(\tilde{N}) - \theta(P)\right] \lesssim \sum_{i=1}^{k}\Bigg((e/4)^{\Delta_{n,k}} + \mathbf{Var}\left[\phi_{\text{plugin}}(\tilde{N}_i) - \phi(p_i)\right]\mathbf{1}_{np_i > \Delta_{n,k}} + \mathbf{Var}\left[\phi_{\text{poly}}(\tilde{N}_i) - \phi(p_i)\right]\mathbf{1}_{np_i \le 4\Delta_{n,k}} + e^{-\Delta_{n,k}/8} + \left(\mathbf{Bias}\left[\phi_{\text{plugin}}(\tilde{N}_i) - \phi(p_i)\right] + \mathbf{Bias}\left[\phi_{\text{poly}}(\tilde{N}_i) - \phi(p_i)\right]\right)^2 \mathbf{1}_{\Delta_{n,k} \le p_i \le 4\Delta_{n,k}}\Bigg).$$

The proofs of Lemmas 4 and 5 are left to Section 5.C. As proved in Lemmas 4 and 5, the bounds on the bias and the variance of our estimator are obtained

with the bias and the variance of the plugin and the best polynomial estimators for each individual alphabet. Thus, we next analyze the bias and the variance of the plugin and the best polynomial estimators.

Analysis of the best polynomial estimator. The following lemmas provide the upper bounds on the bias and the variance of the best polynomial estimator.

Lemma 6. *Let $\tilde{N} \sim \mathrm{Poi}(np)$. Given an integer L and a positive real Δ, let $\phi_L(p) = \sum_{m=0}^{L} a_m p^m$ be the optimal uniform approximation of ϕ by degree-L polynomials on $[0, \Delta]$, and $g_L(\tilde{N}) = \sum_{m=0}^{L} a_m (\tilde{N})_m / n^m$ be an unbiased estimator of $\phi_L(p)$. Under Assumption 2, we have*

$$\mathbf{Bias}\left[(g_L(\tilde{N}) \wedge \phi_{\sup,\Delta}) \vee \phi_{\inf,\Delta} - \phi(p)\right] \lesssim \sqrt{\mathbf{Var}\left[g_L(\tilde{N}) - \phi_L(p)\right]} + \left(\frac{\Delta}{L^2}\right)^{\alpha}.$$

Lemma 7. *Let $\tilde{N} \sim \mathrm{Poi}(np)$. Given an integer L and a positive real $\Delta \gtrsim \frac{1}{n}$, let $\phi_L(p) = \sum_{m=0}^{L} a_m p^m$ be the optimal uniform approximation of ϕ by degree-L polynomials on $[0, \Delta]$, and $g_L(\tilde{N}) = \sum_{m=0}^{L} a_m (\tilde{N})_m / n^m$ be an unbiased estimator of $\phi_L(p)$. Assume Assumption 2. If $p \le \Delta$ and $2\Delta^3 L \le n$, we have*

$$\mathbf{Var}\left[(g_L(\tilde{N}) \wedge \phi_{\sup,\Delta}) \vee \phi_{\inf,\Delta} - \phi(p)\right] \lesssim \frac{\Delta^3 L 64^L (2e)^{2\sqrt{\Delta n L}}}{n}.$$

The proofs of Lemmas 6 and 7 can be found in Section 5.C.

Analysis of the plugin estimator. The following lemmas provide the upper bounds for the bias and the variance of the plugin estimator.

Lemma 8. *Assume Assumption 1 and $\frac{1}{n} \lesssim \Delta < p \le 1$. Let $\tilde{N} \sim \mathrm{Poi}(np)$. Then, we have*

$$\mathbf{Bias}\left[\bar{\phi}_\Delta\left(\frac{\tilde{N}}{n}\right) - \frac{\tilde{N}}{2n^2}\bar{\phi}_\Delta^{(2)}\left(\frac{\tilde{N}}{n}\right) - \phi(p)\right] \lesssim \frac{1}{n^2\Delta^{2-\alpha}} + \frac{p}{n^2}.$$

Lemma 9. *Assume Assumption 1 and $\frac{1}{n} \lesssim \Delta < p \le 1$. Let $\tilde{N} \sim \mathrm{Poi}(np)$. Then, we have*

$$\mathbf{Var}\left[\bar{\phi}_\Delta\left(\frac{\tilde{N}}{n}\right) - \frac{\tilde{N}}{2n^2}\bar{\phi}_\Delta^{(2)}\left(\frac{\tilde{N}}{n}\right) - \phi(p) + \frac{p\phi^{(2)}(p)}{2n}\right] \lesssim \frac{p^{2\alpha-1}}{n} + \frac{1}{n^4\Delta^{4-2\alpha}} + \frac{p}{n}.$$

The proofs of Lemmas 8 and 9 are left to Section 5.C.

Proof for the Upper Bound. Combining Lemmas 4 to 9, we prove Theorem 11.

Proof of Theorem 11. Set $L = \lfloor C_1 \ln n \rfloor$ and $\Delta_{n,k} = C_2 \ln n$ where C_1 and C_2 are some positive constants. Substituting Lemmas 6 to 9 into Lemmas 4 and 5 yields

$$\begin{aligned}
&\mathbf{Bias}\left[\hat{\theta}(\tilde{N}) - \theta(P)\right] \\
&\lesssim \sum_{i=1}^{k}\left(\frac{1}{n^{C_2(\ln 4-1)}} + \frac{1}{n^{\alpha}(\ln n)^{2-\alpha}} + \frac{p_i}{n^2} + \frac{(\ln n)^2 n^{3C_1 \ln 2 + 2\sqrt{C_1 C_2}(\ln 2+1)}}{n^2} \right. \\
&\qquad\qquad \left. + \frac{1}{(n \ln n)^{\alpha}} + \frac{1}{n^{C_2/8}}\right) \\
&\leq \frac{k}{n^{C_2(\ln 4-1)}} + \frac{k}{n^{\alpha}(\ln n)^{2-\alpha}} + \frac{1}{n^2} + \frac{k(\ln n)^2 n^{3C_1 \ln 2 + 2\sqrt{C_1 C_2}(\ln 2+1)}}{n^2} \\
&\qquad\qquad + \frac{k}{(n \ln n)^{\alpha}} + \frac{k}{n^{C_2/8}},
\end{aligned}$$

and

$$\begin{aligned}
&\mathbf{Var}\left[\hat{\theta}(\tilde{N}) - \theta(P)\right] \\
&\lesssim \sum_{i=1}^{k}\left(\frac{1}{n^{C_2(\ln 4-1)}} + \mathbf{1}_{p_i \geq C_2 \ln n/n}\frac{p_i^{2\alpha-1}}{n} + \frac{1}{n^{2\alpha}(\ln n)^{4-2\alpha}} + \frac{p_i}{n} + \frac{(\ln n)^4 n^{6C_1 \ln 2 + 4\sqrt{C_1 C_2}(\ln 2+1)}}{n^4} + \right. \\
&\qquad\qquad \left. \frac{1}{n^{C_2/8}} + \left(\frac{1}{n^{\alpha}(\ln n)^{2-\alpha}} + \frac{p_i}{n^2} + \frac{(\ln n)^2 n^{3C_1 \ln 2 + 2\sqrt{C_1 C_2}(\ln 2+1)}}{n^2} + \frac{1}{(n \ln n)^{\alpha}}\right)^2\right) \\
&\lesssim \frac{k}{n^{C_2(\ln 4-1)}} + \frac{k^{2-2\alpha}}{n} \vee \frac{k}{n^{2\alpha}\ln^{1-2\alpha} n} + \frac{k}{n^{2\alpha}(\ln n)^{4-2\alpha}} + \frac{1}{n} \\
&\quad + \frac{k(\ln n)^4 n^{6C_1 \ln 2 + 4\sqrt{C_1 C_2}(\ln 2+1)}}{n^4} + \frac{k}{n^{C_2/8}} + \frac{k}{n^{2\alpha}(\ln n)^{4-2\alpha}} + \frac{1}{n^4} \\
&\qquad\qquad + \frac{k(\ln n)^4 n^{6C_1 \ln 2 + 4\sqrt{C_1 C_2}(\ln 2+1)}}{n^4} + \frac{k}{(n \ln n)^{2\alpha}},
\end{aligned}$$

where we use Lemmas 19 and 20. For $\delta > 0$, as long as $C_2(\ln 4 - 1) \geq 2\alpha + \delta$, $6C_1 \ln 2 + 4\sqrt{C_1 C_2}(\ln 2 + 1) \leq 3 - 2\alpha - \delta$, and $C_2/8 \geq 2\alpha + \delta$, we have

$$\begin{aligned}
\mathbf{Bias}\left[\hat{\theta}(\tilde{N}) - \theta(P)\right]^2 &\lesssim \frac{1}{n^4} + \frac{k^2}{n^{2\alpha+\delta}} + \frac{k^2}{(n \ln n)^{2\alpha}} \lesssim \frac{1}{n^4} + \frac{k^2}{(n \ln n)^{2\alpha}} \qquad (5.4) \\
\mathbf{Var}\left[\hat{\theta}(\tilde{N}) - \theta(P)\right] &\lesssim \frac{k^{2-2\alpha}}{n} \vee \frac{k}{n^{2\alpha}\ln^{1-2\alpha} n} + \frac{k}{n^{2\alpha+\delta}} + \frac{k}{(n \ln n)^{2\alpha}} \\
&\lesssim \frac{k^{2-2\alpha}}{n} \vee \frac{k}{n^{2\alpha}\ln^{1-2\alpha} n} + \frac{k}{(n \ln n)^{2\alpha}} \qquad (5.5)
\end{aligned}$$

There exist the constants C_1 and C_2 that satisfies these conditions, for example, $C_1 < 1/6 \ln 2$ and $C_2 > 16$. Since $\hat{\theta}(\tilde{N}), \theta(P) \in [\theta_{\inf}, \theta_{\sup}]$, the bias-variance decomposition gives

$$\begin{aligned}\sup_{P \in \mathcal{M}_k} \mathbf{E}\left[\left(\hat{\theta}(\tilde{N}) - \theta(P)\right)^2\right] &\leq \sup_{P \in \mathcal{M}_k} \mathbf{E}\left[\left(\tilde{\theta}(\tilde{N}) - \theta(P)\right)^2\right] \\ &\leq \left(\mathbf{Bias}\left[\tilde{\theta}(\tilde{N}) - \theta(P)\right]\right)^2 + \mathbf{Var}\left[\tilde{\theta}(\tilde{N}) - \theta(P)\right]. \end{aligned} \tag{5.6}$$

Substituting Eqs (5.4) and (5.5) into Eq (5.6) yields

$$\sup_{P \in \mathcal{M}_k} \mathbf{E}\left[\left(\hat{\theta}(\tilde{N}) - \theta(P)\right)^2\right] \lesssim \frac{k^{2-2\alpha}}{n} \vee \frac{k}{n^{2\alpha} \ln^{1-2\alpha} n} + \frac{k^2}{(n \ln n)^{2\alpha}}.$$

If $\alpha \in (0, 1/2]$ and $k \gtrsim \ln^{\frac{4}{3}}$, the last term is dominated. If $\alpha \in (1/2, 1)$, the term $\frac{k}{n^{2\alpha} \ln^{1-2\alpha} n}$ is dominated by $\frac{k^{2-2\alpha}}{n}$. □

By Theorem 11, we prove that the presented estimator achieves the minimax optimal rate Theorems 7 and 8. The condition $k \gtrsim \ln^{\frac{4}{3}} n$ for $\alpha \in (0, 1/2]$ comes from the variance of the bias-corrected plugin estimator shown in Lemma 9. Removing this condition remains as an open problem.

Appendix 5.A Error Rate of Best Polynomial Approximation

Here, we analyze the upper bound and the lower bound of the best polynomial approximation error $E_L(\phi, [0, \Delta])$. The upper bound and the lower bound are derived as follows.

Lemma 10. *Under Assumption 2, for $\Delta \in (0, 1]$, we have*

$$E_L(\phi, [0, \Delta]) \lesssim \left(\frac{\Delta}{L^2}\right)^{\alpha}.$$

Lemma 11. *Under Assumption 2, for $\Delta \in (0, 1]$ there is a positive constant c such that*

$$\liminf_{L \to \infty} \left(\frac{L^2}{\Delta}\right)^{\alpha} E_L(\phi, [0, \Delta]) > c.$$

Combining Lemmas 10 and 11, we can conclude $E_(\phi, [0, \Delta]) \asymp \left(\frac{\Delta}{L^2}\right)^{\alpha}$. The proofs of these lemmas are given as follows.

Proof of Lemma 10. Letting $\phi_\Delta(p) = \phi(\Delta x^2)$, we have $E_L(\phi, [0, \Delta]) = E_L(\phi_\Delta, [-1, 1])$. We utilize the Jackson's inequality to upper bound the best polynomial approximation error E_L by using the modulus of continuity defined as

$$\omega(f, \delta) = \sup_{x,y \in [-1,1]} \{|f(x) - f(y)| : |x - y| \leq \delta\}.$$

To derive the upper bound of E_L, we divide into two cases: $\alpha \in (0, 1/2]$ and $\alpha \in (1/2, 1)$.

Case $\alpha \in (0, 1/2]$**.** From the Jackson's inequality (Achieser 2013), there is a trigonometric polynomial T_L with degree-L such that

$$\sup_{x \in [0,2\pi]} |f(x) - T_L(x)| \lesssim \sup_{x,y \in [0,2\pi]} \left\{ |f(x) - f(y)| : |x - y| \leq \frac{1}{L} \right\}.$$

By the definition of E_L, we have

$$\begin{aligned}
E_L(f, [-1, 1]) &= \inf_{g \in \mathcal{P}_L} \sup_{x \in [-1,1]} |f(x) - g(x)| \\
&= \inf_{g \in \mathcal{P}_L} \sup_{x \in [0,2\pi]} |f(\cos(x)) - g(\cos(x))| \\
&\lesssim \sup_{x,y \in [0,2\pi]} \left\{ |f(\cos(x)) - f(\cos(y))| : |x - y| \leq \frac{1}{L} \right\} \\
&= \sup_{x,y \in [-1,1]} \left\{ |f(x) - f(y)| : \left|\cos^{-1}(x) - \cos^{-1}(y)\right| \leq \frac{1}{L} \right\} \\
&\leq \sup_{x,y \in [-1,1]} \left\{ |f(x) - f(y)| : |x - y| \leq \frac{1}{L} \right\} = \omega\left(f, \frac{1}{L}\right), \quad (5.7)
\end{aligned}$$

where we use the fact that $|\cos^{-1}(x) - \cos^{-1}(y)| \geq |x - y|$ for $x, y \in [-1, 1]$ to derive the last line. From Lemma 17 and the fact that $p^{\alpha-1} \geq 1$ for $p \in (0, 1]$, we have $\left|\phi^{(1)}(p)\right| \leq (W + |c_1|)p^{\alpha-1}$ for $p \in (0, 1]$. From the absolute

continuousness of ϕ on $(0,1]$, for $x, y \in (-1,1]$ where $x \leq y$ we have

$$\begin{aligned}|\phi_\Delta(x) - \phi_\Delta(y)| \leq& \int_x^y \left|2\Delta t \phi^{(1)}\left(\Delta t^2\right)\right| dt \\ \leq& 2\Delta^\alpha (W + |c_1|) \int_x^y t^{2\alpha - 1} dt \\ =& \frac{\Delta^\alpha (W + |c_1|)}{\alpha}\left(y^{2\alpha} - x^{2\alpha}\right) \\ \leq& \frac{\Delta^\alpha (W + |c_1|)}{\alpha}(y - x)^{2\alpha},\end{aligned}$$

where the last line is obtained since x^β for $\beta \in (0,1]$ is β-Holder continuous. This is valid for the case $x = 0$ since $|\phi_\Delta(0) - \phi_\Delta(y)| = \lim_{x \to 0} |\phi_\Delta(x) - \phi_\Delta(y)|$. Thus, we have

$$\omega(\phi_\Delta, \delta) \leq \frac{\Delta^\alpha (W + |c_1|)}{\alpha} \delta^{2\alpha}.$$

Substituting this into Eq (5.7), we have

$$E_L(\phi_\Delta, [-1,1]) \lesssim \frac{\Delta^\alpha (W + |c_1|)}{\alpha} \frac{1}{L^{2\alpha}} \lesssim \left(\frac{\Delta}{L^2}\right)^\alpha.$$

Case $\alpha \in (1/2, 1)$**.** From the Jackson's inequality (Achieser 2013), there is a trigonometric polynomial T_L with degree-L such that

$$\sup_{x \in [0, 2\pi]} |f(x) - T_L(x)| \lesssim \frac{1}{L} \sup_{x, y \in [0, 2\pi]} \left\{ \left|f^{(1)}(x) - f^{(1)}(y)\right| : |x - y| \leq \frac{1}{L} \right\}.$$

In the similar manner of the case $\alpha \in (0, 1/2]$, we have

$$\begin{aligned}E_L(\phi_\Delta, [-1,1]) =& \inf_{g \in \mathcal{P}_L} \sup_{x \in [0, 2\pi]} |\phi_\Delta(\cos(x)) - g(\cos(x))| \\ \lesssim& \frac{1}{L} \omega\left(\phi_\Delta^{(1)}, \frac{1}{L}\right). \end{aligned} \tag{5.8}$$

Since $p^{\alpha - 2} \geq 1$ for $p \in (0,1]$ and Assumption 2, we have $\left|\phi^{(2)}(p)\right| \leq (\alpha_1 W + |c_2|) p^{\alpha - 2}$ for $p \in (0,1]$. From the absolute continuousness of $\phi^{(1)}$ on $(0,1]$, for

$x, y \in (-1, 1]$ where $x \leq y$ we have

$$\begin{aligned}\left|\phi_\Delta^{(1)}(x) - \phi_\Delta^{(1)}(y)\right| \leq& \int_x^y \left|2\Delta\phi^{(1)}\left(\Delta t^2\right) + 4\Delta^2 t^2 \phi^{(2)}\left(\Delta t^2\right)\right| dt \\ \leq& \int_x^y \left(2\Delta^\alpha (W + |c_1|)t^{2\alpha-2} + 4\Delta^\alpha(\alpha_1 W + |c_2|)t^{2\alpha-2}\right) dt \\ =& \Delta^\alpha \frac{2(W + |c_1|) + 4(\alpha_1 W + |c_2|)}{2\alpha - 1}\left(y^{2\alpha-1} - x^{2\alpha-1}\right) \\ \leq& \Delta^\alpha \frac{2(W + |c_1|) + 4(\alpha_1 W + |c_2|)}{2\alpha - 1}(y - x)^{2\alpha-1}.\end{aligned}$$

Also, we use the fact that x^β for $\beta \in (0, 1]$ is β-Holder continuous. Thus, we have

$$\omega\left(\phi_\Delta^{(1)}, \delta\right) \leq \Delta^\alpha \frac{2(W + |c_1|) + 4(\alpha_1 W + |c_2|)}{2\alpha - 1}\delta^{2\alpha-1}.$$

Substituting this into Eq (5.8), we have

$$E_L(\phi_\Delta, [-1, 1]) \lesssim \frac{1}{L}\Delta^\alpha \frac{2(W + |c_1|) + 4(\alpha_1 W + |c_2|)}{2\alpha - 1}\frac{1}{L^{1-2\alpha}} \lesssim \left(\frac{\Delta}{L^2}\right)^\alpha.$$

□

Proof of Lemma 11. Let $\phi_\Delta(x) = \phi\left(\Delta\frac{x+1}{2}\right)$. Then, we have $E_L(\phi, [0, \Delta]) = E_L(\phi_\Delta, [-1, 1])$. To derive the lower bound of $E_L(\phi_\Delta, [-1, 1])$, we introduce the second-order Ditzian-Totik modulus of smoothness (Ditzian et al. 2012) defined as

$$\omega_\varphi^2(f, t) = \sup_{x,y\in[-1,1]}\left\{\left|f(x) + f(y) - 2f\left(\frac{x+y}{2}\right)\right| : |x - y| \leq 2t\varphi\left(\frac{x+y}{2}\right)\right\},$$

where $\varphi(x) = \sqrt{1-x^2}$. Fix $y = -1$, for $t > 0$ we have

$$
\begin{aligned}
|x-y| \le 2t\varphi\left(\frac{x+y}{2}\right) &\iff \\
x+1 \le 2t\sqrt{1-\frac{(x-1)^2}{4}} &\iff \\
\frac{(x+1)^2}{4t^2} \le 1-\frac{(x-1)^2}{4} &\iff \\
t^{-2}(x+1)^2 + (x-1)^2 - 4 \le 0 &\iff \\
(t^{-2}+1)x^2 + 2(t^{-2}-1)x + (t^{-2}+1) - 4 \le 0 &\iff \\
\left(x + \frac{t^{-2}-1}{t^{-2}+1}\right)^2 + 1 - \frac{4}{t^{-2}+1} - \frac{(t^{-2}-1)^2}{(t^{-2}+1)^2} \le 0 &\iff \\
\left(x+1-\frac{2}{t^{-2}+1}\right)^2 \le \frac{4}{(t^{-2}+1)^2} &\iff \\
-1 \le x \le -1 + \frac{4}{t^{-2}+1}.
\end{aligned}
$$

Thus, we have

$$
\begin{aligned}
\omega_\varphi^2(\phi_\Delta, t) \ge & \sup_x \left\{ \left| \phi_\Delta(x) + \phi_\Delta(-1) - 2\phi_\Delta\left(\frac{x-1}{2}\right) \right| : -1 \le x \le -1 + \frac{4}{t^{-2}+1} \right\} \\
= & \sup_x \left\{ \left| \phi(\Delta x) + \phi(0) - 2\phi\left(\frac{\Delta x}{2}\right) \right| : 0 \le x \le \frac{2}{t^{-2}+1} \right\}
\end{aligned}
$$

Application of the Taylor theorem gives

$$
\begin{aligned}
\phi(\Delta x) + \phi(0) - 2\phi\left(\frac{\Delta x}{2}\right) = & \lambda\phi^{(1)}\left(\frac{\Delta x}{2}\right)\left(0 - \frac{x}{2}\right) + \lambda\phi^{(1)}\left(\frac{\Delta x}{2}\right)\left(x - \frac{x}{2}\right) \\
& - \int_0^{\frac{x}{2}} \Delta^2\phi^{(2)}(\Delta t)(0-t)dt + \int_{\frac{x}{2}}^{x} \Delta^2\phi^{(2)}(\Delta t)(x-t)dt \\
= & \int_0^{\frac{x}{2}} \Delta^2\phi^{(2)}(\Delta t)t dt + \int_{\frac{x}{2}}^{x} \Delta^2\phi^{(2)}(\Delta t)(x-t)dt.
\end{aligned}
$$

Letting $p_0 = (\alpha_1 W/(\alpha_1 W \vee -c_2'))^{1/(2-\alpha)}$, $\left|\phi^{(2)}(p)\right| \ge \alpha_1 W p^{\alpha-2} + c_2' \ge 0$ for $(0, p_0]$. From continuousness of $\phi^{(2)}$, $\phi^{(2)}(x)$ has the same sign in $x \in (0, p_0]$.

Since $t \geq 0$ for $t \in [0, \frac{x}{2}]$ and $x - t \geq 0$ for $t \in [\frac{x}{2}, x]$, we have for $x \in (0, p_0]$

$$
\begin{aligned}
&\left|\phi(\Delta x) + \phi(0) - 2\phi\left(\frac{\Delta x}{2}\right)\right| \\
\geq &\Delta^\alpha \alpha_1 W \left(\int_0^{\frac{x}{2}} t^{\alpha-2} t dt + \int_{\frac{x}{2}}^{x} t^{\alpha-2}(x-t)dt\right) + c_2' \Delta^2 \left(\int_0^{\frac{x}{2}} t dt + \int_{\frac{x}{2}}^{x} (x-t)dt\right) \\
= &\Delta^\alpha \alpha_1 W \left(\frac{x^\alpha}{\alpha 2^\alpha} + \frac{x}{1-\alpha}\left(\frac{x^{\alpha-1}}{2^{\alpha-1}} - x^{\alpha-1}\right) + \frac{1}{\alpha}\left(\frac{x^\alpha}{2^\alpha} - x^\alpha\right)\right) + \frac{c_2' \Delta^2 x^2}{4} \\
= &\Delta^\alpha x^\alpha \left(W(2^{-\alpha} - 1) + \frac{\alpha_1 W}{\alpha}(2^{1-\alpha} - 1) + \frac{c_2' \Delta^{2-\alpha}}{4} x^{2-\alpha}\right) \gtrsim \Delta^\alpha x^\alpha.
\end{aligned}
$$

Thus, we have for sufficiently small t

$$
\omega_\varphi^2(\phi_\Delta, t) \gtrsim \Delta^\alpha \left(\frac{2}{t^{-2} + 1}\right)^\alpha \gtrsim \Delta^\alpha t^{2\alpha}. \tag{5.9}
$$

With the definition of $\omega_\varphi^2(f, t)$, we have the converse result $\frac{1}{L^2}\sum_{m=1}^{L}(m+1)E_m(f, [-1, 1]) \gtrsim \omega_\varphi^2(f, L^{-1})$ (Ditzian et al. 2012). Let L' be an integer such that $L' = c_\ell L$ where $c_\ell > 1$. Then, we have

$$
\begin{aligned}
&E_L(\phi, [0, \Delta]) \\
\geq &\frac{1}{L' - L} \sum_{m=L+1}^{L'} E_m(\phi, [0, \Delta]) \\
\geq &\frac{1}{L'^2} \sum_{m=L+1}^{L'} (m+1) E_m(\phi, [0, \Delta]) \\
\geq &\frac{1}{L'^2} \sum_{m=0}^{L'} (m+1) E_m(\phi, [0, \Delta]) - \frac{1}{L'^2} E_0(\phi, [0, \Delta]) - \frac{1}{L'^2} \sum_{m=1}^{L} (m+1) E_m(\phi, [0, \Delta]).
\end{aligned} \tag{5.10}
$$

From Lemma 18, we have $|\phi(x) - \phi(y)| \leq \frac{W}{\alpha}\Delta^\alpha + |c_1|\Delta$ for $x, y \in [0, \Delta]$. Substituting it and Eq (5.9) into Eq (5.10) and applying the converse result

and Lemma 10 yields that there are constants $C > 0$ and $C' > 0$ such that

$$\begin{aligned}
E_L(\phi, [0, \Delta]) \geq & C\omega_\varphi^2(\phi_\Delta, L'^{-1}) - \frac{W}{L'^2\alpha}\Delta^\alpha - \frac{|c_1|}{L'^2}\Delta - \frac{C'}{L'^2}\sum_{m=1}^{L}(m+1)\left(\frac{\Delta}{m^2}\right)^\alpha \\
\geq & C\frac{\Delta^\alpha}{L'^{2\alpha}} - \frac{W}{L'^2\alpha}\Delta^\alpha - \frac{|c_1|}{L'^2}\Delta - \frac{C'}{L'^2}\sum_{m=1}^{L}(m+1)\left(\frac{\Delta}{m^2}\right)^\alpha \\
\geq & C\frac{\Delta^\alpha}{L'^{2\alpha}} - \frac{W}{\alpha c_\ell^2 L^{2\alpha}}\Delta^\alpha - \frac{|c_1|}{c_\ell^2 L^{2\alpha}}\Delta - \frac{2C'\Delta^\alpha}{L'^2}\sum_{m=1}^{L} m^{1-2\alpha} \\
\geq & C\frac{\Delta^\alpha}{L'^{2\alpha}} - \frac{W}{\alpha c_\ell^2 L^{2\alpha}}\Delta^\alpha - \frac{|c_1|}{c_\ell^2 L^{2\alpha}}\Delta - \frac{2C'\Delta^\alpha}{L'^2}\left(L^{2-2\alpha} \vee \int_0^L x^{1-2\alpha}dx\right) \\
\geq & C\frac{\Delta^\alpha}{c_\ell^{2\alpha}L^{2\alpha}} - \frac{W}{\alpha c_\ell^2 L^{2\alpha}}\Delta^\alpha - \frac{|c_1|}{c_\ell^2 L^{2\alpha}}\Delta - \frac{2C'\Delta^\alpha}{((2-2\alpha)\wedge 1)c_\ell^2 L^{2\alpha}} \\
= & \frac{1}{c_\ell^{2\alpha}}\left(\frac{\Delta}{L^2}\right)^\alpha\left(C - \frac{W}{\alpha c_\ell^{2-2\alpha}} - \frac{|c_1|\Delta^{-\alpha}}{c_\ell^{2-2\alpha}} - \frac{2C'}{((2-2\alpha)\wedge 1)c_\ell^{2-2\alpha}}\right).
\end{aligned}$$

Thus, by taking sufficiently large c_ℓ, there is $c > 0$ such that

$$\limsup_{L\to\infty}\left(\frac{L^2}{\Delta}\right)^\alpha E_L(\phi, [0, \Delta]) > c.$$

□

Appendix 5.B Proofs for Lower Bounds

To prove Theorem 9, the Le Cam's two-point method (See, e.g., (Tsybakov 2009)). The consequent corollary of the Le Cam's two-point method is as follows.

Corollary 2. *For any two probability measures $P, Q \in \mathcal{M}_k$, we have*

$$\tilde{R}^*(n, k; \phi) \geq \frac{1}{4}(\theta(P) - \theta(Q))^2 \exp(-nD_{\mathrm{KL}}(P, Q)),$$

where $D_{\mathrm{KL}}(P, Q)$ denotes the KL-divergence between P and Q.

We provide the proof of Theorem 9.

Proof of Theorem 9. For $\epsilon \in (0, 1/2)$. Define two probability measures on $[k]$

as

$$
\begin{aligned}
P =& \left(\frac{1}{2}, \frac{1}{2(k-1)}, ..., \frac{1}{2(k-1)}\right), \\
Q =& \left(\frac{1}{2}(1+\epsilon), \frac{1}{2(k-1)}(1-\epsilon), ..., \frac{1}{2(k-1)}(1-\epsilon)\right).
\end{aligned}
$$

Then, the KL-divergence between P and Q is obtained as

$$
D_{\mathrm{KL}}(P, Q) = -\frac{1}{2}\ln(1+\epsilon) - \frac{1}{2}\ln(1-\epsilon) = -\frac{1}{2}\ln\left(1-\epsilon^2\right) \leq \epsilon^2.
$$

Applying the Taylor theorem gives that there exist $\xi_1 \in [1/2, (1+\epsilon)/2]$ and $\xi_2 \in [(1-\epsilon)/2(k-1), 1/2(k-1)]$ such that

$$
\begin{aligned}
&\theta(Q) - \theta(P) \\
=&\frac{1}{2}\phi^{(1)}\left(\frac{1}{2}\right)\epsilon - \frac{1}{2}\phi^{(1)}\left(\frac{1}{2(k-1)}\right)\epsilon + \frac{\phi^{(2)}(\xi_1)}{8}\epsilon^2 + \frac{\phi^{(2)}(\xi_2)}{8(k-1)}\epsilon^2.
\end{aligned}
$$

From the reverse triangle inequality, we have

$$
\begin{aligned}
&|\theta(Q) - \theta(P)| \\
\geq&\frac{1}{2}\left|\phi^{(1)}\left(\frac{1}{2(k-1)}\right)\right|\epsilon - \left|\frac{1}{2}\phi^{(1)}\left(\frac{1}{2}\right)\epsilon + \frac{\phi^{(2)}(\xi_1)}{8}\epsilon^2 + \frac{\phi^{(2)}(\xi_2)}{8(k-1)}\epsilon^2\right| \\
\geq&\frac{1}{2}\left|\phi^{(1)}\left(\frac{1}{2(k-1)}\right)\right|\epsilon - \left|\frac{1}{2}\phi^{(1)}\left(\frac{1}{2}\right)\right|\epsilon - \left|\frac{\phi^{(2)}(\xi_1)}{8}\right|\epsilon^2 - \left|\frac{\phi^{(2)}(\xi_2)}{8(k-1)}\right|\epsilon^2.
\end{aligned}
$$

Combining Assumption 2, Lemma 17, and the fact that $\xi_1 \geq 1/2$ and $\xi_2 \geq 1/4(k-1)$ yields

$$
\begin{aligned}
\left|\phi^{(1)}\left(\frac{1}{2(k-1)}\right)\right| \geq& W 2^{1-\alpha}(k-1)^{1-\alpha} + c_1', \\
\left|\phi^{(1)}\left(\frac{1}{2}\right)\right| \leq& W 2^{1-\alpha} + c_1, \\
\left|\phi^{(2)}(\xi_1)\right| \leq& \alpha_1 W 2^{2-\alpha} + c_2, \\
\left|\phi^{(2)}(\xi_2)\right| \leq& \alpha_1 W 4^{2-\alpha}(k-1)^{2-\alpha} + c_2.
\end{aligned}
$$

Consequently, we have

$$
\begin{aligned}
|\theta(Q) - \theta(P)| \geq W 2^{-\alpha}\epsilon\big((k-1)^{1-\alpha} - 1 - \alpha_1(2^{-1} + 2^{1-\alpha}(k-1)^{1-\alpha})\epsilon\big) \\
- 2^{-1}(c_1 - c_1')\epsilon - c_2(2^{-3} + 2^{-3}(k-1)^{-1})\epsilon^2.
\end{aligned}
$$

Set $\epsilon = 1/\sqrt{n}$. Applying Corollary 2, we have

$$\begin{aligned}
&\tilde{R}^*(n,k;\phi) \\
&\geq \frac{W^2(k-1)^{2-2\alpha}}{2^{-2\alpha}n}\left(1 - \frac{1}{(k-1)^{1-\alpha}} - \frac{\alpha_1}{2(k-1)^{\alpha}\sqrt{n}} - \frac{\alpha_1 2^{1-\alpha}}{\sqrt{n}}\right. \\
&\qquad \left. - \frac{c_1 - c_1'}{2^{1-\alpha}W(k-1)^{1-\alpha}} - \frac{2^{\alpha-3}c_2}{W(k-1)^{1-\alpha}\sqrt{n}} - \frac{2^{\alpha-3}c_2}{W(k-1)^{2-\alpha}\sqrt{n}}\right)^2 \\
&\gtrsim \frac{k^{2-2\alpha}}{n}.
\end{aligned}$$

From Lemma 3, this lower bound is valid for $R^*(n,k;\phi)$. □

The proof of Theorem 10 is following the proof of (Yihong Wu et al. 2016). For $\epsilon \in (0,1)$, define the approximate probabilities by

$$\mathcal{M}_k(\epsilon) = \left\{ \{p_i\}_{i=1}^k \in \mathbb{R}_+^k : \sum_{i=1}^k p_i \leq 1 - \epsilon \right\}.$$

With this definition, we define the minimax risk for $\mathcal{M}_k(\epsilon)$ as

$$\tilde{R}^*(n,k,\epsilon;\phi) = \inf_{\hat{\theta}} \sup_{P \in \mathcal{M}_k(\epsilon)} \mathbf{E}\Big(\hat{\theta}(\tilde{N}) - \theta(P)\Big)^2. \tag{5.11}$$

The minimax risk of Poisson sampling can be bounded below by Eq (5.11) as

Lemma 12. *Under Assumption 2, for any $k, n \in \mathbb{N}$ and any $\epsilon < 1/3$,*

$$\tilde{R}^*(n/2,k;\phi) \geq \frac{1}{3}\tilde{R}^*(n,k,\epsilon;\phi) - 4\left(\frac{W}{\alpha}k^{1-\alpha} + |c_1|\right)^2 e^{-n/32} - \frac{W^2}{\alpha^2}k^{2-2\alpha}\epsilon^{2\alpha} - c_1^2\epsilon^2.$$

Proof of Lemma 12. This proof is following the same manner of the proof of (Yihong Wu et al. 2016, Lemma 1). Fix $\delta > 0$. Let $\hat{\theta}(\cdot, n)$ be a near-minimax optimal estimator for fixed sample size n, i.e.,

$$\sup_{P \in \mathcal{M}_k} \mathbf{E}\Big[(\hat{\theta}(N,n) - \theta(P))^2\Big] \leq \delta + R^*(k,n;\phi).$$

For an arbitrary approximate distribution $P \in \mathcal{M}_k(\epsilon)$, we construct an estimator

$$\tilde{\theta}(\tilde{N}) = \hat{\theta}(\tilde{N}, n'),$$

where $\tilde{N}_i \sim \mathrm{Poi}(np_i)$ and $n' = \sum_i N_i$. From the triangle inequality, Lemma 18 and Lemma 19, we have

$$
\begin{aligned}
&\frac{1}{3}(\tilde{\theta}(\tilde{N}) - \theta(P))^2 \\
\leq &\frac{1}{3}\left(\left|\tilde{\theta}(\tilde{N}) - \theta\left(\frac{P}{\sum_{i=1}^k p_i}\right)\right| + \left|\theta\left(\frac{P}{\sum_{i=1}^k p_i}\right) - \theta(P)\right|\right)^2 \\
\leq &\frac{1}{3}\left(\left|\tilde{\theta}(\tilde{N}) - \theta\left(\frac{P}{\sum_{i=1}^k p_i}\right)\right| + \frac{W}{\alpha}\sum_{i=1}^k \left|\frac{p_i}{\sum_{j=1}^k p_j} - p_i\right|^\alpha + |c_1|\sum_{i=1}^k \left|\frac{p_i}{\sum_{j=1}^k p_j} - p_i\right|\right)^2 \\
\leq &\frac{1}{3}\left(\left|\tilde{\theta}(\tilde{N}) - \theta\left(\frac{P}{\sum_{i=1}^k p_i}\right)\right| + \frac{W}{\alpha}\sum_{i=1}^k \left(\frac{p_i}{\sum_{j=1}^k p_j}\left|\sum_{j=1}^k p_j - 1\right|\right)^\alpha \right. \\
&\qquad\qquad \left. + |c_1|\sum_{i=1}^k \frac{p_i}{\sum_{j=1}^k p_j}\left|\sum_{j=1}^k p_j - 1\right|\right)^2 \\
\leq &\frac{1}{3}\left(\left|\tilde{\theta}(\tilde{N}) - \theta\left(\frac{P}{\sum_{i=1}^k p_i}\right)\right| + \frac{W}{\alpha}\epsilon^\alpha \sum_{i=1}^k \left(\frac{p_i}{\sum_{j=1}^k p_j}\right)^\alpha + |c_1|\epsilon\sum_{i=1}^k \frac{p_i}{\sum_{j=1}^k p_j}\right)^2 \\
\leq &\frac{1}{3}\left(\left|\tilde{\theta}(\tilde{N}) - \theta\left(\frac{P}{\sum_{i=1}^k p_i}\right)\right| + \frac{W}{\alpha}k^{1-\alpha}\epsilon^\alpha + |c_1|\epsilon\right)^2 \\
\leq &\left(\left|\tilde{\theta}(\tilde{N}) - \theta\left(\frac{P}{\sum_{i=1}^k p_i}\right)\right|\right)^2 + \frac{W^2}{\alpha^2}k^{2-2\alpha}\epsilon^{2\alpha} + c_1^2\epsilon^2.
\end{aligned}
$$

For the first term, we observe that $\tilde{N} \sim \mathrm{Multinomial}(m, \frac{P}{\sum p_i})$ conditioned on $n' = m$. Therefore, we have

$$
\begin{aligned}
\mathbf{E}\left(\tilde{\theta}(\tilde{N}) - \theta\left(\frac{P}{\sum_{i=1}^k p_i}\right)\right)^2 &= \sum_{m=0}^\infty \mathbf{E}\left[\left(\tilde{\theta}(\tilde{N}, m) - \theta\left(\frac{P}{\sum_{i=1}^k p_i}\right)\right)^2 \middle| n' = m\right]\mathbb{P}\{n' = m\} \\
&\leq \sum_{m=0}^\infty \tilde{R}^*(m, k; \phi)\mathbb{P}\{n' = m\} + \delta.
\end{aligned}
$$

From Lemma 18 and Lemma 19, we have

$$\begin{aligned}\tilde{R}^*(m,k;\phi) \leq & \sup_{P,P'\in\mathcal{M}_k} (\theta(P)-\theta(P'))^2 \\ \leq & \sup_{P,P'\in\mathcal{M}_k} \left(\frac{W}{\alpha}\sum_{i=1}^k |p_i-p_i'|^\alpha + |c_1|\sum_{i=1}^k |p_i-p_i'|\right)^2 \\ \leq & 4\sup_{P\in\mathcal{M}_k}\left(\frac{W}{\alpha}\sum_{i=1}^k p_i^\alpha + |c_1|\sum_{i=1}^k p_i\right)^2 \\ \leq & 4\left(\frac{W}{\alpha}k^{1-\alpha}+|c_1|\right)^2.\end{aligned}$$

Note that $\tilde{R}^*(m,k;\phi)$ is a decreasing function with respect to m. Since $n' \sim \mathrm{Poi}(n\sum_i p_i)$ and $|\sum_i p_i - 1| \leq \epsilon \leq 1/3$, applying Chernoff bound yields $\mathbb{P}\{n' \leq n/2\} \leq e^{-n/32}$. Thus, we have

$$\begin{aligned}&\mathbf{E}\left(\tilde{\theta}(\tilde{N}) - \theta\left(\frac{P}{\sum_{i=1}^k p_i}\right)\right)^2 \\ \leq & \sum_{m\geq n/K} \tilde{R}^*(m,k;\phi)\mathbb{P}\{n'=m\} + 4\left(\frac{W}{\alpha}k^{1-\alpha}+|c_1|\right)^2\mathbb{P}\{n'\leq n/K\} + \delta \\ \leq & \tilde{R}^*(n/K,k;\phi) + 4\left(\frac{W}{\alpha}k^{1-\alpha}+|c_1|\right)^2 e^{-n/32} + \delta.\end{aligned}$$

The arbitrariness of δ gives the desired result. □

The lower bound of $\tilde{R}^*(n,k,\epsilon;\phi)$ is given by the following lemma.

Lemma 13. *Let U and U' be random variables such that $U,U' \in [0,\lambda]$ and $\mathbf{E}[U] = \mathbf{E}[U'] \leq 1$ and $|\mathbf{E}[\theta(U)-\theta(U')]| \geq d$, where $\lambda \leq k$. Let $\epsilon = 4\lambda/\sqrt{k}$. Then*

$$\tilde{R}^*(n,k,\epsilon;\phi) \geq \frac{d^2}{16}\left(\frac{7}{8} - k\mathrm{TV}(\mathbf{E}[\mathrm{Poi}(nU/k)],\mathbf{E}[\mathrm{Poi}(nU'/k)]) - \frac{64W^2\lambda^{2\alpha}}{\alpha^2 k^{2\alpha-1}d^2} - \frac{64c_1^2\lambda^2}{kd^2}\right).$$

Proof of Lemma 13. The proof follows the same manner of the proof of (Yihong Wu et al. 2016, Lemma 2) expect Eq (5.12) below. Let $\beta = \mathbf{E}[U] = \mathbf{E}[U'] \leq 1$. Define two random vectors

$$P = \left(\frac{U_1}{k},...,\frac{U_k}{k},1-\beta\right), P' = \left(\frac{U_1'}{k},...,\frac{U_k'}{k},1-\beta\right),$$

where U_ia nd U_i' are independent copies of U and U', respectively. Put $\epsilon = 4\lambda/\sqrt{k}$. Define the two events:

$$\mathcal{E} = \left[\left|\sum_i \frac{U_i}{k} - \beta\right| \le \epsilon, |\theta(P) - \mathbf{E}[\theta(P)]| \le d/4\right],$$
$$\mathcal{E}' = \left[\left|\sum_i \frac{U_i'}{k} - \beta\right| \le \epsilon, |\theta(P') - \mathbf{E}[\theta(P')]| \le d/4\right].$$

Applying Chebyshev's inequality, the union bound, the triangle inequality and Lemma 18 gives

$$\begin{aligned}
\mathbb{P}\mathcal{E}^c \le& \mathbb{P}\left\{\left|\sum_i \frac{U_i}{k} - \beta\right| > \epsilon\right\} + \mathbb{P}\{|\theta(P) - \mathbf{E}[\theta(P)]| > d/4\} \\
\le& \frac{\mathbf{Var}[U]}{k\epsilon^2} + \frac{16\sum_i \mathbf{Var}[\phi(U_i/k)]}{d^2} \\
\le& \frac{1}{16} + \frac{16\sum_i \mathbf{E}[(\phi(U_i/k) - \phi(\beta/k))^2]}{d^2} \\
\le& \frac{1}{16} + \frac{32\sum_i \mathbf{E}[W^2(U_i - \beta)^{2\alpha}]}{\alpha^2 k^{2\alpha} d^2} + \frac{32\sum_i \mathbf{E}[c_1^2(U_i - \beta)^2]}{k^2 d^2} \\
\le& \frac{1}{16} + \frac{32W^2\lambda^{2\alpha}}{\alpha^2 k^{2\alpha-1} d^2} + \frac{32c_1^2\lambda^2}{kd^2}
\end{aligned} \tag{5.12}$$

By the same manner, we have

$$\mathbb{P}\mathcal{E}'^c \le \frac{1}{16} + \frac{32W^2\lambda^{2\alpha}}{\alpha^2 k^{2\alpha-1} d^2} + \frac{32c_1^2\lambda^2}{kd^2}.$$

We define two priors on the set $\mathcal{M}_k(\epsilon)$, the conditional distributions $\pi = P_{U|\mathcal{E}}$ and $\pi' = P_{U'|\mathcal{E}'}$. By the definition of events $\mathcal{E}, \mathcal{E}'$ and triangle inequality, we obtain that under π, π',

$$|\theta(P) - \theta(P')| \ge \frac{d}{2}.$$

By triangle inequality, we have the total variation of observations under π, π' as

$$\begin{aligned}
\mathrm{TV}(P_{\tilde{N}|\mathcal{E}}, P_{\tilde{N}'|\mathcal{E}'}) \le& \mathrm{TV}(P_{\tilde{N}|\mathcal{E}}, P_{\tilde{N}}) + \mathrm{TV}(P_{\tilde{N}}, P_{\tilde{N}'}) + \mathrm{TV}(P_{\tilde{N}'}, P_{\tilde{N}'|\mathcal{E}'}) \\
=& \mathbb{P}\mathcal{E}^c + \mathrm{TV}(P_{\tilde{N}}, P_{\tilde{N}'}) + \mathbb{P}\mathcal{E}'^c \\
\le& \mathrm{TV}(P_{\tilde{N}}, P_{\tilde{N}'}) + \frac{1}{8} + \frac{64W^2\lambda^{2\alpha}}{\alpha^2 k^{2\alpha-1} d^2} + \frac{64c_1^2\lambda^2}{kd^2}.
\end{aligned}$$

From the fact that total variation of product distribution can be upper bounded by the summation of individual ones, we obtain

$$\begin{aligned}\mathrm{TV}(P_{\tilde{N}}, P_{\tilde{N}'}) \leq & \sum_{i=1}^{k} \mathrm{TV}(P_{\tilde{N}_i}, P_{\tilde{N}'_i}) + \mathrm{TV}(n(1-\beta), n(1-\beta)) \\ = & k\mathrm{TV}(\mathbf{E}[\mathrm{Poi}(nU/k)], \mathbf{E}[\mathrm{Poi}(nU'/k)]).\end{aligned}$$

Then, applying Le Cam's lemma (Le Cam 1986) yields that

$$\tilde{R}^*(n,k,\epsilon;\phi) \geq \frac{d^2}{16}\left(\frac{7}{8} - k\mathrm{TV}(\mathbf{E}[\mathrm{Poi}(nU/k)], \mathbf{E}[\mathrm{Poi}(nU'/k)]) - \frac{64W^2\lambda^{2\alpha}}{\alpha^2 k^{2\alpha-1} d^2} - \frac{64c_1^2\lambda^2}{kd^2}\right).$$

□

To derive the upper bound of $\mathrm{TV}(\mathbf{E}[\mathrm{Poi}(nU/k)], \mathbf{E}[\mathrm{Poi}(nU'/k)])$, we apply the following lemma proved by Yihong Wu et al. (2016).

Lemma 14 (Yihong Wu et al. (2016, Lemma 3))**.** *Let V and V' be random variables on $[0, M]$. If $\mathbf{E}[V^j] = \mathbf{E}[V'^j]$, $j = 1, ..., L$ and $L > 2eM$, then*

$$\mathrm{TV}(\mathbf{E}[\mathrm{Poi}(V)], \mathbf{E}[\mathrm{Poi}(V')]) \leq \left(\frac{2eM}{L}\right)^L.$$

Under the condition of Lemma 14, the following lemmas provides the lower bound of d.

Lemma 15. *For any given integer $L > 0$, there exists two probability measures ν_0 and ν_1 on $[0, \lambda]$ such that*

$$\begin{aligned}\mathbf{E}_{X\sim\nu_0}[X^m] &= \mathbf{E}_{X\sim\nu_1}[X^m], \ \textit{for } m = 0, ..., L, \\ \mathbf{E}_{X\sim\nu_0}[\phi(X)] - \mathbf{E}_{X\sim\nu_1}[\phi(X)] &= 2E_L(\phi, [0, \lambda]).\end{aligned}$$

Lemma 15. The proof is almost same as the proof of Jiantao Jiao et al. (2015, Lemma 10). It follows directly from a standard functional analysis argument proposed by Lepski et al. (1999). It suffices to replace x^α with $\phi(x)$ and $[0, 1]$ with $[0, \lambda]$ in the proof of (Cai et al. 2011, Lemma 1). □

As proved Lemma 15, we can choose the probability measures of U and U' in Lemma 12 so that d in Lemma 12 becomes the uniform approximation error of the best polynomial $E_L(\phi, [0, \lambda])$. The analysis of the lower bound on $E_L(\phi, [0, \lambda])$ can be found in Section 5.A. By using the lower bound (in Lemma 11), we prove Theorem 10 as follows.

Proof of Theorem 10. Set $L = \lfloor C_1 \ln n \rfloor$ and $\lambda = C_2 \frac{\ln n}{n}$ where C_1 and C_2 are universal constants such that $2eC_2 \leq C_1$. Assembling Lemmas 11 and 13 to 15, we have $M = C_2 \frac{\ln n}{k}$, $|\mathbf{E}[\phi(U) - \phi(U')]| = d \geq ck\left(\frac{\lambda}{L^2}\right)^{\alpha}$ where $c > 0$ is an universal constant. Also, we have

$$\begin{aligned}&\tilde{R}^*(n, k, \epsilon; \phi)\\ \geq&\frac{d^2}{16}\left(\frac{7}{8} - k\left(\frac{2eC_2 \ln n}{k\lfloor C_1 \ln n \rfloor}\right)^{\lfloor C_1 \ln n \rfloor} - \frac{64W^2\lfloor C_1 \ln n \rfloor^{4\alpha}}{c^2\alpha^2 k^{2\alpha+1}} - \frac{64c_1^2\lfloor C_1 \ln n \rfloor^{4\alpha}\lambda^{2-2\alpha}}{c^2k^3}\right).\end{aligned}$$

If $\alpha \in (1/2, 1)$, it is sufficient to prove Theorem 10 when $k \gtrsim n^{1-1/2\alpha} \ln n$ because of Theorem 9. Hence,

$$\frac{64W^2\lfloor C_1 \ln n \rfloor^{4\alpha}}{c^2\alpha^2 k^{2\alpha+1}} = o(1) \tag{5.13}$$

$$\frac{64c_1^2\lfloor C_1 \ln n \rfloor^{4\alpha}\lambda^{2-2\alpha}}{c^2k^3} = o(1). \tag{5.14}$$

If $\alpha \in (0, 1/2]$, we assume $k \gtrsim \ln^{\frac{4}{3}} n$. Then, we get Eqs (5.13) and (5.14). Moreover, for sufficiently large C_1, we get $k\left(\frac{2eC_2 \ln n}{k\lfloor C_1 \ln n \rfloor}\right)^{\lfloor C_1 \ln n \rfloor} = o(1)$.Thus, we have

$$\tilde{R}^*(n, k, \epsilon; \phi) \gtrsim d^2 \gtrsim \frac{k^2}{(n \ln n)^{2\alpha}}. \tag{5.15}$$

The second term in Lemma 12 is bounded above as

$$4\left(\frac{W}{\alpha}k^{1-\alpha} + |c_1|\right)^2 e^{-n/32} = o\left(\frac{k^2}{(n \ln n)^{2\alpha}}\right).$$

For $\alpha \in (0, 1)$, we get an upper bound on the fourth term in Lemma 12 as

$$\begin{aligned}c_1^2\epsilon^2 \leq&\frac{c_1^2\lambda^{2-2\alpha}L^{4\alpha}}{k^2} \cdot d^2\\ \leq&\frac{c_1^2\lambda^{2-2\alpha}\lfloor C_1 \ln n \rfloor^{4\alpha}}{k^2} \cdot d^2 = o(1) \cdot d^2.\end{aligned}$$

If $\alpha \in (1/2, 1)$, the third term in Lemma 12 is bounded above as

$$\begin{aligned}\frac{W^2}{\alpha^2}k^{2-2\alpha}\epsilon^{2\alpha} \leq&\frac{W^2L^{4\alpha}}{c^2\alpha^2k^{3\alpha}} \cdot d^2\\ \leq&\frac{W^2\lfloor C_1 \ln n \rfloor^{4\alpha}}{c^2\alpha^2k^{3\alpha}} \cdot d^2 = o(1) \cdot d^2.\end{aligned}$$

Then, Eq (5.15) and Lemma 12 gives

$$\tilde{R}^*(n,k;\phi) \gtrsim \frac{k^2}{(n\ln n)^{2\alpha}}.$$

If $\alpha \in (0,1/2]$, we assume $k \geq c' \ln^{\frac{4}{3}} n$ for an arbitrary constant $c' > 0$, and we get

$$\frac{W^2}{\alpha^2} k^{2-2\alpha}\epsilon^{2\alpha} \leq \frac{W^2 C_1^{4\alpha}}{c^2\alpha^2 c'^{3\alpha}} \cdot d^2.$$

Hence, for sufficiently small c', Eq (5.15) and Lemma 12 yields

$$\tilde{R}^*(n,k;\phi) \gtrsim \frac{k^2}{(n\ln n)^{2\alpha}}.$$

□

Appendix 5.C Proofs for Upper Bounds

We use the following helper lemma for proving Lemma 5.

Lemma 16 (Cai et al. (2011), Lemma 4)**.** *Suppose* $\mathbf{1}_{\mathcal{E}}$ *is an indicator random variable independent of* X *and* Y*, then*

$$\mathbf{Var}[X\mathbf{1}_{\mathcal{E}} + Y\mathbf{1}_{\mathcal{E}^c}] = \mathbf{Var}[X]\mathbb{P}\mathcal{E} + \mathbf{Var}[Y]\mathbb{P}\mathcal{E}^c + (\mathbf{E}[X] - \mathbf{E}[Y])^2\mathbb{P}\mathcal{E}\mathbb{P}\mathcal{E}^c.$$

Proof of Lemma 4. From the property of the absolute value, the bias is bounded above as

$$\begin{aligned}
&\mathbf{Bias}\Big[\hat{\theta}(\tilde{N}) - \theta(P)\Big] \\
\leq &\sum_{i=1}^{k}\Big(\mathbf{Bias}\Big[\mathbf{1}_{\tilde{N}_i' \geq 2\Delta_{n,k}}\Big(\phi_{\text{plugin}}(\tilde{N}_i) - \phi(p_i)\Big)\Big] + \mathbf{Bias}\Big[\mathbf{1}_{\tilde{N}_i' < 2\Delta_{n,k}}\Big(\phi_{\text{poly}}(\tilde{N}_i) - \phi(p_i)\Big)\Big]\Big).
\end{aligned}$$

Because of the independence between $\tilde{N}$ and $\tilde{N}'$, we have

$$\begin{aligned}
\mathbf{Bias}\Big[\mathbf{1}_{\tilde{N}_i' \geq 2\Delta_{n,k}}\Big(\phi_{\text{plugin}}(\tilde{N}_i) - \phi(p_i)\Big)\Big] =& \mathbf{Bias}\Big[\phi_{\text{plugin}}(\tilde{N}_i) - \phi(p_i)\Big]\mathbb{P}\Big\{\tilde{N}_i' \geq 2\Delta_{n,k}\Big\} \\
\mathbf{Bias}\Big[\mathbf{1}_{\tilde{N}_i' < 2\Delta_{n,k}}\Big(\phi_{\text{poly}}(\tilde{N}_i) - \phi(p_i)\Big)\Big] =& \mathbf{Bias}\Big[\phi_{\text{poly}}(\tilde{N}_i) - \phi(p_i)\Big]\mathbb{P}\Big\{\tilde{N}_i' < 2\Delta_{n,k}\Big\}
\end{aligned}$$

For $p \in [\frac{\Delta_{n,k}}{2n}, \frac{\Delta_{n,k}}{n}]$, from Lemmas 17 and 18, we have

$$\begin{aligned}
&\left|H_4\left(p;\phi,\frac{\Delta_{n,k}}{n},\frac{\Delta_{n,k}}{2n}\right)-\phi(p_i)\right| \\
\leq&\left|\sum_{m=1}^{4}\frac{\phi^{(m)}\left(\frac{\Delta_{n,k}}{n}\right)}{m!}\left(p-\frac{\Delta_{n,k}}{n}\right)^m\sum_{\ell=0}^{4-m}\frac{4+1}{4+\ell+1}\mathrm{B}_{\ell,4+\ell+1}\left(\frac{p-\frac{\Delta_{n,k}}{n}}{\frac{\Delta_{n,k}}{2n}-\frac{\Delta_{n,k}}{n}}\right)\right| \\
&\qquad+\left|\phi\left(\frac{\Delta_{n,k}}{n}\right)-\phi(p_i)\right| \\
\leq&\sum_{m=1}^{4}\frac{\left|\phi^{(m)}\left(\frac{\Delta_{n,k}}{n}\right)\right|}{m!}\left(\frac{\Delta_{n,k}}{2n}\right)^m\sum_{\ell=0}^{4-m}\binom{4+\ell}{\ell}\left(\frac{\ell}{4+\ell+1}\right)^{\ell}\left(\frac{4+1}{4+\ell+1}\right)^{4+1} \\
&\qquad+\frac{W}{\alpha}+|c_1| \\
\leq&\sum_{m=1}^{4}\frac{\left|\phi^{(m)}\left(\frac{\Delta_{n,k}}{n}\right)\right|}{m!}\left(\frac{\Delta_{n,k}}{2n}\right)^m(5-m)+\frac{W}{\alpha}+|c_1| \\
\leq&5\sum_{m=1}^{4}\left(\frac{\alpha_{m-1}W}{m!}\left(\frac{\Delta_{n,k}}{n}\right)^{\alpha}2^{-m}+c_m\left(\frac{\Delta_{n,k}}{2n}\right)^m\right)+\frac{W}{\alpha}+|c_1|,
\end{aligned}$$

where we use $0 \leq \mathrm{B}_{\nu,n}(x) \leq \mathrm{B}_{\nu,n}(\nu/n)$ to get the third line. From the assumption $\Delta_{n,k} \leq n$, we have

$$\left|H_4\left(p;\phi,\frac{\Delta_{n,k}}{n},\frac{\Delta_{n,k}}{2n}\right)-\phi(p)\right| \leq 5\sum_{m=1}^{4}\left(\frac{\alpha_{m-1}W}{m!2^m}+c_m\right)+\frac{W}{\alpha}+|c_1|.$$

Also, for $p \in [1,2]$, we have

$$\begin{aligned}
&|H_4(p;\phi,1,2)-\phi(p_i)| \\
\leq&\left|\sum_{m=1}^{4}\frac{\phi^{(m)}(1)}{m!}(p-1)^m\sum_{\ell=0}^{4-m}\frac{4+1}{4+\ell+1}\mathrm{B}_{\ell,4+\ell+1}\left(\frac{p-1}{2-1}\right)\right|+|\phi(1)-\phi(p_i)| \\
\leq&5\sum_{m=1}^{4}\frac{\left|\phi^{(m)}(1)\right|}{m!}+\frac{W}{\alpha}+|c_1| \\
\leq&5\sum_{m=1}^{4}(\alpha_{m-1}W+c_m)+\frac{W}{\alpha}+|c_1|.
\end{aligned}$$

For $p \in (\frac{\Delta_{n,k}}{n}, 1)$, we have by Lemma 18 that

$$|\phi(p)-\phi(p_i)| \leq \frac{W}{\alpha}+|c_1|.$$

Consequently, we have for $p \geq 0$

$$\left|\bar{\phi}_{\frac{\Delta_{n,k}}{n}}(p) - \phi(p_i)\right| \leq 5\sum_{m=1}^{4}(\alpha_{m-1}W + c_m) + \frac{W}{\alpha} + |c_1| \lesssim 1. \tag{5.16}$$

For $p \in (\frac{\Delta_{n,k}}{2n}, \frac{\Delta_{n,k}}{n})$,

$$\begin{aligned}
&\frac{p}{2n}\left|H_4^{(2)}\left(p;\phi,\frac{\Delta_{n,k}}{n},\frac{\Delta_{n,k}}{2n}\right)\right| \\
&= \frac{p}{2n}\left|\sum_{m=1}^{4}\phi^{(m)}\left(\frac{\Delta_{n,k}}{n}\right)\sum_{i=0}^{2}\binom{2}{i}\frac{1}{((m-i)\vee 0)!}\left(p-\frac{\Delta_{n,k}}{n}\right)^{(m-i)\vee 0}\right. \\
&\qquad\qquad \left.\sum_{\ell=0}^{4-m}\frac{4+1}{4+\ell+1}\,\mathrm{B}_{\ell,4+\ell+1}^{(2-i)}\left(\frac{p-\frac{\Delta_{n,k}}{n}}{-\frac{\Delta_{n,k}}{2n}}\right)\right| \\
&= \frac{p}{2n}\left|\sum_{m=1}^{4}\phi^{(m)}\left(\frac{\Delta_{n,k}}{n}\right)\sum_{i=0}^{2}\binom{2}{i}\frac{1}{((m-i)\vee 0)!}\left(p-\frac{\Delta_{n,k}}{n}\right)^{(m-i)\vee 0}\right. \\
&\quad \left.\sum_{\ell=0}^{4-m}\frac{(4+1)(4+\ell+1)!}{(4+\ell+1)(4+\ell-1+i)!}\sum_{j=0}^{(2-i)\wedge\ell}(-1)^j\binom{2-i}{j}\mathrm{B}_{\ell-j,4+\ell-1+i}\left(\frac{p-\frac{\Delta_{n,k}}{n}}{-\frac{\Delta_{n,k}}{2n}}\right)\right|,
\end{aligned}$$

where the last line is obtained by using the fact $\mathrm{B}_{\nu,n}^{(1)}(x) = n(\mathrm{B}_{\nu-1,n-1}(x) - \mathrm{B}_{\nu,n-1}(x))$. Again, the fact $0 \leq \mathrm{B}_{\nu,n}(x) \leq \mathrm{B}_{\nu,n}(\nu/n)$ gives

$$\begin{aligned}
&\frac{p}{2n}\left|H_4^{(2)}\left(p;\phi,\frac{\Delta_{n,k}}{n},\frac{\Delta_{n,k}}{2n}\right)\right| \\
&\leq \frac{p}{2n}\sum_{m=1}^{4}\left|\phi^{(m)}\left(\frac{\Delta_{n,k}}{n}\right)\right|\sum_{i=0}^{2}\binom{2}{i}\frac{1}{((m-i)\vee 0)!}\left(\frac{\Delta_{n,k}}{2n}\right)^{(m-i)\vee 0} \\
&\qquad \sum_{\ell=0}^{4-m}\sum_{j=0}^{(2-i)\wedge\ell}\binom{2-i}{j}\frac{(4+1)(4+\ell)!}{(\ell-j)!(4-1+i+j)!}\frac{(\ell-j)^{\ell-j}(4-1+i+j)^{4-1+i+j}}{(4+\ell-1+i)^{4+\ell-1+i}} \\
&\leq \frac{p}{2n}\sum_{m=1}^{4}\left|\phi^{(m)}\left(\frac{\Delta_{n,k}}{n}\right)\right|\left(\frac{5-m}{((m-2)\vee 0)!}\left(\frac{\Delta_{n,k}}{2n}\right)^{(m-2)\vee 0}\right. \\
&\qquad\qquad \left.+\frac{20(5-m)}{(m-1)!}\left(\frac{\Delta_{n,k}}{2n}\right)^{m-1}+\frac{20(4+(4-m)(5-m))}{2m!}\left(\frac{\Delta_{n,k}}{2n}\right)^{m}\right) \\
&\leq \frac{1}{n}\sum_{m=1}^{4}\left(\alpha_{m-1}W\left(\frac{\Delta_{n,k}}{n}\right)^{\alpha-m}+c_m\right)\left(\frac{5-m}{((m-2)\vee 0)!}\left(\frac{\Delta_{n,k}}{2n}\right)^{(m-1)\vee 1}\right. \\
&\qquad\qquad \left.+\frac{20(5-m)}{(m-1)!}\left(\frac{\Delta_{n,k}}{2n}\right)^{m}+\frac{20(4+(4-m)(5-m))}{2m!}\left(\frac{\Delta_{n,k}}{2n}\right)^{m+1}\right).
\end{aligned}$$

From the assumption $\Delta_{n,k} \leq n$, we have

$$\begin{aligned}
&\frac{p}{2n}\left|H_4^{(2)}\left(p;\phi,\frac{\Delta_{n,k}}{n},\frac{\Delta_{n,k}}{2n}\right)\right| \\
\leq &\frac{1}{n}\sum_{m=1}^{4}\left(\alpha_{m-1}W\left(\frac{\Delta_{n,k}}{n}\right)^{\alpha-1}+c_m\right) \\
&\qquad\left(\frac{(5-m)}{2^{m-1}((m-2)\vee 0)!}+\frac{20(5-m)}{2^m(m-1)!}+\frac{20(4+(4-m)(5-m))}{2^{m+2}m!}\right).
\end{aligned}$$

From the assumption, there is a universal constant $c > 0$ such that $\Delta_{n,k} \geq c$. Thus, we have

$$\begin{aligned}
&\frac{p}{2n}\left|H_4^{(2)}\left(p;\phi,\frac{\Delta_{n,k}}{n},\frac{\Delta_{n,k}}{2n}\right)\right| \\
\leq &\sum_{m=1}^{4}\left(\alpha_{m-1}W\frac{c^{\alpha-1}}{n^\alpha}+\frac{c_m}{n}\right) \\
&\qquad\left(\frac{(5-m)}{2^{m-1}((m-2)\vee 0)!}+\frac{20(5-m)}{2^m(m-1)!}+\frac{20(4+(4-m)(5-m))}{2^{m+2}m!}\right).
\end{aligned}$$

Also, for $p \in (1,2)$, we have

$$\begin{aligned}
&\frac{p}{2n}\left|H_4^{(2)}(p;\phi,1,2)\right| \\
\leq &\frac{1}{n}\sum_{m=1}^{4}(\alpha_{m-1}W+c_m)\left(\frac{5-m}{((m-2)\vee 0)!}+\frac{20(5-m)}{(m-1)!}+\frac{20(4+(4-m)(5-m))}{2m!}\right).
\end{aligned}$$

Thus, we have for $p \geq 0$

$$\begin{aligned}
&\left|\frac{p}{2n}\bar{\phi}^{(2)}_{\frac{\Delta_{n,k}}{n}}(p)\right| \\
\leq &\sum_{m=1}^{4}\left(\alpha_{m-1}W\frac{c^{\alpha-1}}{n^\alpha}+\frac{|c_m|}{n}\right) \\
&\qquad\left(1\vee\left(\frac{(5-m)}{((m-2)\vee 0)!}+\frac{20(5-m)}{(m-1)!}+\frac{20(4+(4-m)(5-m))}{2m!}\right)\right) \\
\lesssim &\frac{1}{n^\alpha}. \qquad (5.17)
\end{aligned}$$

Combining Eqs (5.16) and (5.17) yields for any $p_i \in [0,1]$

$$\mathbf{Bias}\left[\phi_{\text{plugin}}(\tilde{N}_i)-\phi(p_i)\right] \lesssim 1.$$

Then, we have

$$\begin{aligned}
&\mathbf{Bias}\Big[\phi_{\mathrm{plugin}}(\tilde{N}_i) - \phi(p_i)\Big]\mathbb{P}\Big\{\tilde{N}_i' \geq 2\Delta_{n,k}\Big\} \\
=&\mathbf{Bias}\Big[\phi_{\mathrm{plugin}}(\tilde{N}_i) - \phi(p_i)\Big]\mathbb{P}\Big\{\tilde{N}_i' \geq 2\Delta_{n,k}\Big\}\mathbf{1}_{np_i \leq \Delta_{n,k}} \\
&\qquad + \mathbf{Bias}\Big[\phi_{\mathrm{plugin}}(\tilde{N}_i) - \phi(p_i)\Big]\mathbb{P}\Big\{\tilde{N}_i' \geq 2\Delta_{n,k}\Big\}\mathbf{1}_{np_i > \Delta_{n,k}} \\
\lesssim&\mathbb{P}\Big\{\tilde{N}_i' \geq 2\Delta_{n,k}\Big\}\mathbf{1}_{np_i \leq \Delta_{n,k}} + \mathbf{Bias}\Big[\phi_{\mathrm{plugin}}(\tilde{N}_i) - \phi(p_i)\Big]\mathbf{1}_{np_i > \Delta_{n,k}}.
\end{aligned}$$

The Chernoff bound for the Poisson distribution gives $\mathbb{P}\Big\{\tilde{N}_i' \geq 2\Delta_{n,k}\Big\}\mathbf{1}_{np_i \leq \Delta_{n,k}} \leq (e/4)^{\Delta_{n,k}}$. Thus, we have

$$\begin{aligned}
&\mathbf{Bias}\Big[\phi_{\mathrm{plugin}}(\tilde{N}_i) - \phi(p_i)\Big]\mathbb{P}\Big\{\tilde{N}_i' \geq 2\Delta_{n,k}\Big\} \\
\lesssim&(e/4)^{\Delta_{n,k}} + \mathbf{Bias}\Big[\phi_{\mathrm{plugin}}(\tilde{N}_i) - \phi(p_i)\Big]\mathbf{1}_{np_i > \Delta_{n,k}}.
\end{aligned} \tag{5.18}$$

Similarly, we have by the final truncation of ϕ_{poly} and Lemma 18 that

$$\mathbf{Bias}\Big[\phi_{\mathrm{poly}}(\tilde{N}_i) - \phi(p_i)\Big] \leq \sup_{p\in[0,1]} |\phi(p) - \phi(p_i)| \leq \frac{W}{\alpha} + |c_1|.$$

The Chernoff bound yields $\mathbb{P}\Big\{\tilde{N}_i' < 2\Delta_{n,k}\Big\} \leq e^{-\Delta_{n,k}/8}$ for $p_i > 4\Delta_{n,k}$. Thus, we have

$$\begin{aligned}
&\mathbf{Bias}\Big[\phi_{\mathrm{poly}}(\tilde{N}_i) - \phi(p_i)\Big]\mathbb{P}\Big\{\tilde{N}_i' < 2\Delta_{n,k}\Big\} \\
\leq&\mathbf{Bias}\Big[\phi_{\mathrm{poly}}(\tilde{N}_i) - \phi(p_i)\Big]\mathbb{P}\Big\{\tilde{N}_i' < 2\Delta_{n,k}\Big\}\mathbf{1}_{np_i \leq 4\Delta_{n,k}} \\
&\quad + \mathbf{Bias}\Big[\phi_{\mathrm{poly}}(\tilde{N}_i) - \phi(p_i)\Big]\mathbb{P}\Big\{\tilde{N}_i' < 2\Delta_{n,k}\Big\}\mathbf{1}_{np_i > 4\Delta_{n,k}} \\
\leq&\mathbf{Bias}\Big[\phi_{\mathrm{poly}}(\tilde{N}_i) - \phi(p_i)\Big]\mathbf{1}_{np_i \leq 4\Delta_{n,k}} + \left(\frac{W}{\alpha} + |c_1|\right)e^{-\Delta_{n,k}/8}.
\end{aligned} \tag{5.19}$$

Combining Eqs (5.18) and (5.19) gives the desired result. □

Proof of Lemma 5. Because of the independence of $\tilde{N}_1, .., \tilde{N}_k, \tilde{N}_1', ..., \tilde{N}_k'$, ap-

plying Lemma 16 gives

$$
\begin{aligned}
&\mathbf{Var}\Big[\hat{\theta}(\tilde{N}) - \theta(P)\Big] \\
\leq &\mathbf{Var}\left[\sum_{i=1}^{k} \mathbf{1}_{\tilde{N}_i' \geq 2\Delta_{n,k}} \Big(\phi_{\text{plugin}}(\tilde{N}_i) - \phi(p_i)\Big) + \mathbf{1}_{\tilde{N}_i' < 2\Delta_{n,k}} \Big(\phi_{\text{poly}}(\tilde{N}_i) - \phi(p_i)\Big)\right] \\
\leq &\sum_{i=1}^{k} \mathbf{Var}\Big[\mathbf{1}_{\tilde{N}_i' \geq 2\Delta_{n,k}} \Big(\phi_{\text{plugin}}(\tilde{N}_i) - \phi(p_i)\Big) + \mathbf{1}_{\tilde{N}_i' < 2\Delta_{n,k}} \Big(\phi_{\text{poly}}(\tilde{N}_i) - \phi(p_i)\Big)\Big] \\
\leq &\sum_{i=1}^{k} \Bigg(\mathbf{Var}\Big[\phi_{\text{plugin}}(\tilde{N}_i) - \phi(p_i)\Big] \mathbb{P}\Big\{\tilde{N}_i' \geq 2\Delta_{n,k}\Big\} \\
&\qquad + \mathbf{Var}\Big[\phi_{\text{poly}}(\tilde{N}_i) - \phi(p_i)\Big] \mathbb{P}\Big\{\tilde{N}_i' < 2\Delta_{n,k}\Big\} \\
&\qquad + \Big(\mathbf{E}\Big[\phi_{\text{plugin}}(\tilde{N}_i) - \phi(p_i)\Big] - \mathbf{E}\Big[\phi_{\text{poly}}(\tilde{N}_i) - \phi(p_i)\Big]\Big)^2 \\
&\qquad\qquad \mathbb{P}\Big\{\tilde{N}_i' \geq 2\Delta_{n,k}\Big\} \mathbb{P}\Big\{\tilde{N}_i' < 2\Delta_{n,k}\Big\}\Bigg).
\end{aligned}
\tag{5.20}
$$

We can derive upper bounds on the first two terms of Eq (5.20) in the same manner of Eqs (5.18) and (5.19) as

$$
\begin{aligned}
&\mathbf{Var}\Big[\phi_{\text{plugin}}(\tilde{N}_i) - \phi(p_i)\Big] \mathbb{P}\Big\{\tilde{N}_i' \geq 2\Delta_{n,k}\Big\} \\
&\qquad \lesssim (e/4)^{\Delta_{n,k}} + \mathbf{Var}\Big[\phi_{\text{plugin}}(\tilde{N}_i) - \phi(p_i)\Big] \mathbf{1}_{np_i > \Delta_{n,k}},
\end{aligned}
$$

and

$$
\begin{aligned}
&\mathbf{Var}\Big[\phi_{\text{poly}}(\tilde{N}_i) - \phi(p_i)\Big] \mathbb{P}\Big\{\tilde{N}_i' < 2\Delta_{n,k}\Big\} \\
&\qquad \lesssim \mathbf{Var}\Big[\phi_{\text{poly}}(\tilde{N}_i) - \phi(p_i)\Big] \mathbf{1}_{np_i \leq 4\Delta_{n,k}} + e^{-\Delta_{n,k}/8}.
\end{aligned}
$$

By the Chernoff bound, we have

$$
\begin{aligned}
&\mathbb{P}\Big\{\tilde{N}_i' \geq 2\Delta_{n,k}\Big\} \mathbb{P}\Big\{\tilde{N}_i' < 2\Delta_{n,k}\Big\} \\
=&(\mathbf{1}_{p_i < \Delta_{n,k}} + \mathbf{1}_{p_i > 4\Delta_{n,k}} + \mathbf{1}_{\Delta_{n,k} \leq p_i \leq 4\Delta_{n,k}}) \mathbb{P}\Big\{\tilde{N}_i' \geq 2\Delta_{n,k}\Big\} \mathbb{P}\Big\{\tilde{N}_i' < 2\Delta_{n,k}\Big\} \\
\leq &(e/4)^{\Delta_{n,k}} + e^{-\Delta_{n,k}/8} + \mathbf{1}_{\Delta_{n,k} \leq p_i \leq 4\Delta_{n,k}}.
\end{aligned}
$$

Thus, we have the upper bound of the last term of Eq (5.20) as

$$\begin{aligned}
&\left(\mathbf{E}\left[\phi_{\text{plugin}}(\tilde{N}_i) - \phi(p_i)\right] - \mathbf{E}\left[\phi_{\text{poly}}(\tilde{N}_i) - \phi(p_i)\right]\right)^2 \mathbb{P}\left\{\tilde{N}_i' \geq 2\Delta_{n,k}\right\}\mathbb{P}\left\{\tilde{N}_i' < 2\Delta_{n,k}\right\} \\
&\leq \left(\mathbf{Bias}\left[\phi_{\text{plugin}}(\tilde{N}_i) - \phi(p_i)\right] + \mathbf{Bias}\left[\phi_{\text{poly}}(\tilde{N}_i) - \phi(p_i)\right]\right)^2 \\
&\qquad\qquad \left((e/4)^{\Delta_{n,k}} + e^{-\Delta_{n,k}/8} + \mathbf{1}_{\Delta_{n,k} \leq p_i \leq 4\Delta_{n,k}}\right) \\
&\lesssim (e/4)^{\Delta_{n,k}} + e^{-\Delta_{n,k}/8} \\
&\quad + \left(\mathbf{Bias}\left[\phi_{\text{plugin}}(\tilde{N}_i) - \phi(p_i)\right] + \mathbf{Bias}\left[\phi_{\text{poly}}(\tilde{N}_i) - \phi(p_i)\right]\right)^2 \mathbf{1}_{\Delta_{n,k} \leq p_i \leq 4\Delta_{n,k}}.
\end{aligned}$$

□

Next, we prove the upper bounds on the bias and the variance of the best polynomial estimator as follows:

Proof of Lemma 6. Let $\phi'_{\text{sup},\Delta} = \phi_{\text{sup},\Delta} \vee \sup_{p\in[0,\Delta]} \phi_L(p)$ and $\phi'_{\text{inf},\Delta} = \phi_{\text{inf},\Delta} \wedge \inf_{p\in[0,\Delta]} \phi_L(p)$. By the triangle inequality and the fact that g_L is an unbiased estimator of ϕ_L, we have

$$\begin{aligned}
&\mathbf{Bias}\left[(g_L(\tilde{N}) \wedge \phi_{\text{sup},\Delta}) \vee \phi_{\text{inf},\Delta} - \phi(p)\right] \\
&\quad \leq \mathbf{Bias}\left[(g_L(\tilde{N}) \wedge \phi_{\text{sup},\Delta}) \vee \phi_{\text{inf},\Delta} - (g_L(\tilde{N}) \wedge \phi'_{\text{sup},\Delta}) \vee \phi'_{\text{inf},\Delta}\right] \\
&\quad + \mathbf{Bias}\left[(g_L(\tilde{N}) \wedge \phi'_{\text{sup},\Delta}) \vee \phi'_{\text{inf},\Delta} - \phi_L(p)\right] + \mathbf{Bias}\left[g_L(\tilde{N}) - \phi(p)\right].
\end{aligned}$$

By Chebyshev alternating theorem (Petrushev et al. 2011), the first term is bounded above as

$$\begin{aligned}
&\mathbf{Bias}\left[(g_L(\tilde{N}) \wedge \phi_{\text{sup},\Delta}) \vee \phi_{\text{inf},\Delta} - (g_L(\tilde{N}) \wedge \phi'_{\text{sup},\Delta}) \vee \phi'_{\text{inf},\Delta}\right] \\
&\leq (\phi'_{\text{sup},\Delta} - \phi_{\text{sup},\Delta}) \vee (\phi_{\text{inf},\Delta} - \phi'_{\text{inf},\Delta}) \leq E_L(\phi, [0,\Delta]).
\end{aligned}$$

Also, the third term is bounded above as

$$\mathbf{Bias}\left[g_L(\tilde{N}) - \phi(p)\right] = |\phi_L(p) - \phi(p)| \leq E_L(\phi, [0,\Delta]).$$

The error bound of $E_L(\phi, [0,\Delta])$ is derived in Section 5.A. From Lemma 10, we have $E_L(\phi, [0,\Delta]) \lesssim \left(\frac{\Delta}{L^2}\right)^{\alpha}$. The second term has upper bound as

$$\begin{aligned}
\mathbf{Bias}\left[(g_L(\tilde{N}) \wedge \phi'_{\text{sup},\Delta}) \vee \phi'_{\text{inf},\Delta} - \phi_L(p)\right] &= \sqrt{\left(\mathbf{E}\left[(g_L(\tilde{N}) \wedge \phi'_{\text{sup},\Delta}) \vee \phi'_{\text{inf},\Delta} - \phi_L(p)\right]\right)^2} \\
&\leq \sqrt{\mathbf{E}\left[\left((g_L(\tilde{N}) \wedge \phi'_{\text{sup},\Delta}) \vee \phi'_{\text{inf},\Delta} - \phi_L(p)\right)^2\right]}.
\end{aligned}$$

Since $\phi_L(p) \in [\phi'_{\inf,\Delta}, \phi'_{\sup,\Delta}]$ for $p \in [0,\Delta]$, we have $\Big((g_L(\tilde{N}) \wedge \phi'_{\sup,\Delta}) \vee \phi'_{\inf,\Delta} - \phi_L(p)\Big)^2 \leq \Big(g_L(\tilde{N}) - \phi_L(p)\Big)^2$. Thus, we have

$$\mathbf{Bias}\Big[(g_L(\tilde{N}) \wedge \phi'_{\sup,\Delta}) \vee \phi'_{\inf,\Delta} - \phi_L(p)\Big] \leq \sqrt{\mathbf{Var}\Big[g_L(\tilde{N}) - \phi_L(p)\Big]}.$$

□

Proof of Lemma 7. It is obviously that truncation does not increase the variance, i.e.,

$$\mathbf{Var}\Big[(g_L(\tilde{N}) \wedge \phi_{\sup,\Delta}) \vee \phi_{\inf,\Delta} - \phi(p)\Big] \leq \mathbf{Var}\Big[g_L(\tilde{N}) - \phi(p)\Big].$$

Letting $\phi_\Delta(p) = \phi(\Delta x)$ and $a_0, ..., a_L$ be coefficients of the optimal uniform approximation of ϕ_Δ by degree-L polynomials on $[0,1]$, we have $\sum_{m=0}^{L} \frac{\Delta^m a_m}{n^m}(\tilde{N})_m = g_L(\tilde{N})$. Then, since the standard deviation of sum of random variables is at most the sum of individual standard deviation, we have

$$\mathbf{Var}\Big[g_L(\tilde{N}) - \phi(p)\Big] \leq \left(\sum_{m=1}^{L} \frac{\Delta^m |a_m|}{n^m} \sqrt{\mathbf{Var}(\tilde{N})_m}\right)^2.$$

From (Petrushev et al. 2011) and the fact from Lemma 18 that ϕ is bounded, there is a positive constant C such that $|a_m| \leq C 2^{3L}$. From (Yihong Wu et al. 2016), $\mathbf{Var}(\tilde{N})_m$ is decreasing monotonously as m increases, and for $X \sim \mathrm{Poi}(\lambda)$

$$\mathbf{Var}(X)_m \leq (\lambda m)^m \left(\frac{(2e)^{2\sqrt{\lambda m}}}{\pi\sqrt{\lambda m}} \vee 1\right).$$

By the assumption of $p \leq \Delta$ and monotonous, $\mathbf{Var}(\tilde{N})_m \leq \mathbf{Var}(X)_m$ where $X \sim \mathrm{Poi}(\Delta n)$. Thus, we have

$$\begin{aligned}\mathbf{Var}\Big[g_L(\tilde{N})\Big] \lesssim& \left(\sum_{m=1}^{L} \frac{\Delta^m 2^{3L}}{n^m} \sqrt{(\Delta n L)^m (2e)^{2\sqrt{\Delta n L}}}\right)^2 \\ \leq& \left(\sum_{m=1}^{L} \sqrt{\frac{\Delta^{3m} L^m}{n^m}} 2^{3L} (2e)^{\sqrt{\Delta n L}}\right)^2.\end{aligned}$$

From the assumption $\frac{\Delta^3 L}{n} \leq \frac{1}{2}$, we have

$$
\begin{aligned}
&\left(\sum_{m=1}^{L} c^m \sqrt{\frac{\Delta^{3m} L^m}{n^m}} 2^{3L} (2e)^{\sqrt{\Delta n L}}\right)^2 \\
\leq &\left(2^{3L}(2e)^{\sqrt{\Delta n L}} \sum_{m=1}^{L} \left(\sqrt{\frac{\Delta^3 L}{n}}\right)^m\right)^2 \\
\leq &\left(2^{3L}(2e)^{\sqrt{\Delta n L}} \left(\sqrt{\frac{\Delta^3 L}{n}} + \int_1^L \left(\sqrt{\frac{\Delta^3 L}{n}}\right)^x dx\right)\right)^2 \\
\leq &\left(2^{3L}(2e)^{\sqrt{\Delta n L}} \left(\sqrt{\frac{\Delta^3 L}{n}} + \frac{2}{\ln\left(\frac{\Delta^3 L}{n}\right)} \left(\left(\sqrt{\frac{\Delta^3 L}{n}}\right)^L - \sqrt{\frac{\Delta^3 L}{n}}\right)\right)\right)^2 \\
= &\left(\sqrt{\frac{\Delta^3 L}{n}} 2^{3L}(2e)^{\sqrt{\Delta n L}} \left(1 + \frac{2}{\ln 2}\left(1 - \left(\sqrt{\frac{\Delta^3 L}{n}}\right)^{L-1}\right)\right)\right)^2 \\
\leq &\frac{16\Delta^3 L 64^L (2e)^{2\sqrt{\Delta n L}}}{n} \\
\lesssim &\frac{\Delta^3 L 64^L (2e)^{2\sqrt{\Delta n L}}}{n}.
\end{aligned}
$$

□

The proofs of the upper bounds on the bias and the variance of the bias-corrected plugin estimator are obtained as follows.

Proof of Lemma 8. Applying Taylor theorem yields

$$
\begin{aligned}
&\mathbf{Bias}\left[\bar{\phi}_\Delta\left(\frac{\tilde{N}}{n}\right) - \phi(p)\right] \\
= &\left|\mathbf{E}\left[\phi^{(1)}(p)\frac{\tilde{N} - np}{n} + \frac{\phi^{(2)}(p)}{2}\left(\frac{\tilde{N}}{n} - p\right)^2 - \frac{\tilde{N}}{2n}\bar{\phi}_\Delta^{(2)}\left(\frac{\tilde{N}}{n}\right)\right.\right. \\
&\qquad\left.\left. + \frac{\phi^{(3)}(p)}{6}\left(\frac{\tilde{N}}{n} - p\right)^3 + R_3\left(\frac{\tilde{N}}{n}; \bar{\phi}_\Delta, p\right)\right]\right| \\
\leq &\frac{1}{2n}\left|\mathbf{E}\left[p\phi^{(2)}(p) - \frac{\tilde{N}}{n}\bar{\phi}_\Delta^{(2)}\left(\frac{\tilde{N}}{n}\right)\right]\right| + \frac{p|\phi^{(3)}(p)|}{6n^2} + \left|\mathbf{E}\left[R_3\left(\frac{\tilde{N}}{n}; \bar{\phi}_\Delta, p\right)\right]\right|,
\end{aligned}
\tag{5.21}
$$

where we use the fact that for $X \sim \mathrm{Poi}(\lambda)$, $\mathbf{E}[(X-\lambda)^2] = \lambda$, $\mathbf{E}[(X-\lambda)^3] = \lambda$, and $R_3(x; \bar{\phi}_\Delta, p)$ denotes the reminder term of the Taylor theorem. The first term of Eq (5.21) is bounded above as

$$\begin{aligned}
&\frac{1}{2n}\left|\mathbf{E}\left[p\phi^{(2)}(p) - \frac{\tilde{N}}{n}\bar{\phi}_\Delta^{(2)}\left(\frac{\tilde{N}}{n}\right)\right]\right| \\
=&\frac{1}{2n}\left|\mathbf{E}\left[\phi^{(2)}(p)\left(p - \frac{\tilde{N}}{n}\right) + \frac{\tilde{N}}{n}\left(\phi^{(2)}(p) - \bar{\phi}_\Delta^{(2)}\left(\frac{\tilde{N}}{n}\right)\right)\right]\right| \\
=&\frac{1}{2n}\left|\mathbf{E}\left[\frac{\tilde{N}\phi^{(3)}(p)}{n}\left(\frac{\tilde{N}}{n} - p\right) + \frac{\tilde{N}}{n}R_1\left(\frac{\tilde{N}}{n}; \bar{\phi}_\Delta^{(2)}, p\right)\right]\right| \\
\leq&\frac{p\left|\phi^{(3)}(p)\right|}{2n^2} + \left|\mathbf{E}\left[\frac{\tilde{N}}{2n^2}R_1\left(\frac{\tilde{N}}{n}; \bar{\phi}_\Delta^{(2)}, p\right)\right]\right|, \qquad (5.22)
\end{aligned}$$

where the last line is obtained by using the fact that for $X \sim \mathrm{Poi}(\lambda)$, $\mathbf{E}[X(X-\lambda)] = \lambda$, and $R_1(x; \bar{\phi}_\Delta^{(2)}, p)$ denotes the reminder term of the Taylor theorem. From Lemma 17, the second term of Eq (5.21) and the first term of Eq (5.22) are bounded above as

$$\frac{p\left|\phi^{(3)}(p)\right|}{6n^2} \leq \frac{\alpha_2 W p^{\alpha-2} + c_3 p}{6n^2} \lesssim \frac{1}{n^2\Delta^{2-\alpha}} + \frac{p}{n^2} \qquad (5.23)$$

$$\frac{p\left|\phi^{(3)}(p)\right|}{2n^2} \leq \frac{\alpha_2 W p^{\alpha-2} + c_3 p}{2n^2} \lesssim \frac{1}{n^2\Delta^{2-\alpha}} + \frac{p}{n^2}. \qquad (5.24)$$

The rest is to derive the upper bound on $\left|\mathbf{E}\left[R_3\left(\frac{\tilde{N}}{n}; \bar{\phi}_\Delta, p\right)\right]\right|$ and $\left|\mathbf{E}\left[\frac{\tilde{N}}{2n^2}R_1\left(\frac{\tilde{N}}{n}; \bar{\phi}_\Delta^{(2)}, p\right)\right]\right|$. Let $\hat{p} = \frac{\tilde{N}}{n}$. From the mean value theorem, letting a function $G(x)$ be continuous on the closed interval and differentiable with non-vanishing derivative on the open interval between p and $\hat{p}$, there exists ξ between p and $\hat{p}$ such that

$$R_3\big(\hat{p}; \bar{\phi}_\Delta, p\big) = \frac{\bar{\phi}_\Delta^{(4)}(\xi)}{6}(\hat{p} - \xi)^3\frac{G(\hat{p}) - G(p)}{G^{(1)}(\xi)}.$$

Define $G(x) = \frac{1}{x^2}(\hat{p} - x)^4$. Then, there exists ξ such that

$$\begin{aligned}
R_3\big(\hat{p}; \bar{\phi}_\Delta, p\big) &= -\frac{\bar{\phi}_\Delta^{(4)}(\xi)}{12}(\hat{p} - \xi)^3\frac{\xi^3(\hat{p} - p)^4}{p^2(\xi + \hat{p})(\hat{p} - \xi)^3} \\
&= -\frac{\xi^3\bar{\phi}_\Delta^{(4)}(\xi)}{12p^2(\xi + \hat{p})}(\hat{p} - p)^4 \qquad (5.25)
\end{aligned}$$

Thus, we have

$$\begin{aligned}\left|\mathbf{E}\left[R_3\left(\hat{p};\bar{\phi}_\Delta,p\right)\right]\right| \leq & \mathbf{E}\left[\frac{\xi^3\left|\bar{\phi}_\Delta^{(4)}(\xi)\right|}{12p^2(\xi+\hat{p})}(\hat{p}-p)^4\right] \\ \leq & \frac{1}{12p^2}\mathbf{E}\left[\xi^2\left|\bar{\phi}_\Delta^{(4)}(\xi)\right|(\hat{p}-p)^4\right] \\ \leq & \frac{\sup_{\xi\in\mathbb{R}_+}\xi^2\left|\bar{\phi}_\Delta^{(4)}(\xi)\right|}{12p^2}\mathbf{E}\left[(\hat{p}-p)^4\right] \\ \leq & \left(\frac{1}{4n^2}+\frac{1}{12pn^3}\right)\sup_{\xi\in\mathbb{R}_+}\xi^2\left|\bar{\phi}_\Delta^{(4)}(\xi)\right|,\end{aligned}$$

where we use the fact that for $X\sim\mathrm{Poi}(\lambda)$, $\mathbf{E}[X^4]=3\lambda^2+\lambda$. For $\xi\in(\frac{\Delta}{2},\Delta)$, we have

$$\begin{aligned}&\left|H^{(4)}\left(\xi;\phi,\frac{\Delta}{2},\Delta\right)\right| \\ =&\left|\sum_{m=1}^4\phi^{(m)}(\Delta)\sum_{i=0}^4\binom{4}{i}\frac{1}{((m-i)\vee 0)!}(\xi-\Delta)^{(m-i)\vee 0}\sum_{\ell=0}^{4-m}\frac{5}{5+\ell}\,\mathrm{B}_{\ell,5+\ell}^{(4-i)}\left(\frac{\xi-\Delta}{\Delta/2}\right)\right| \\ =&\left|\sum_{m=1}^4\phi^{(m)}(\Delta)\sum_{i=0}^4\binom{4}{i}\frac{1}{((m-i)\vee 0)!}(\xi-\Delta)^{(m-i)\vee 0}\right. \\ &\left.\sum_{\ell=0}^{4-m}\frac{5(5+\ell)!}{(5+\ell)(1+\ell+i)}\sum_{j=0}^{(4-i)\wedge\ell}(-1)^j\binom{4-i}{j}\,\mathrm{B}_{\ell-j,1+\ell+i}\left(\frac{\xi-\Delta}{\Delta/2}\right)\right|,\end{aligned}$$

where we use $\mathrm{B}_{\nu,n}^{(1)}(x)=n(\mathrm{B}_{\nu-1,n-1}(x)-\mathrm{B}_{\nu,n-1}(x))$. Since $0\leq\mathrm{B}_{\nu,n}(x)\leq\mathrm{B}_{\nu,n}(\nu/n)\leq 1$, there is a universal constant $c>0$ such that for any $i=0,...,4$

$$\left|\sum_{\ell=0}^{4-m}\frac{5(5+\ell)!}{(5+\ell)(1+\ell+i)}\sum_{j=0}^{(4-i)\wedge\ell}(-1)^j\binom{4-i}{j}\,\mathrm{B}_{\ell-j,1+\ell+i}\left(\frac{\xi-\Delta}{\Delta/2}\right)\right|\leq c.$$

Thus, we have from Lemma 17 that

$$\begin{aligned}
&\xi^2\left|H^{(4)}\left(\xi;\phi,\frac{\Delta}{2},\Delta\right)\right| \\
\leq&\sum_{m=1}^{4}\left|\phi^{(m)}(\Delta)\right|\sum_{i=0}^{4}\binom{4}{i}\frac{c}{((m-i)\vee 0)!}\left|\xi^2(\xi-\Delta)^{(m-i)\vee 0}\right| \\
\leq&\sum_{m=1}^{4}\left(\alpha_{m-1}W\Delta^{\alpha-m}+c_m\right)\sum_{i=0}^{4}\binom{4}{i}\frac{c}{((m-i)\vee 0)!}\Delta^{(2+m-i)\vee 2} \\
=&\sum_{m=1}^{4}\sum_{i=0}^{4}\binom{4}{i}\frac{c}{((m-i)\vee 0)!}\left(\alpha_{m-2}W\Delta^{(2+\alpha-i)\vee(2+\alpha-m)}+c_m\Delta^{(2+m-i)\vee 2}\right) \\
\lesssim&\Delta^{\alpha-2}.
\end{aligned}$$

Similarly, for $\xi\in(1,2)$

$$\begin{aligned}
&\xi^2\left|H^{(4)}(\xi;\phi,2,1)\right| \\
\leq&\sum_{m=1}^{4}\left|\phi^{(m)}(1)\right|\sum_{i=0}^{4}\binom{4}{i}\frac{c}{((m-i)\vee 0)!}\left|\xi^2(\xi-1)^{(m-i)\vee 0}\right| \\
\leq&\sum_{m=1}^{4}(\alpha_{m-1}W+c_m)\sum_{i=0}^{4}\binom{4}{i}\frac{4c}{((m-i)\vee 0)!} \\
\lesssim&1.
\end{aligned}$$

For $\xi\in[\Delta,1]$, we have from Lemma 17 that

$$\left|\xi^2\phi^{(4)}(\xi)\right|\leq\alpha_1W\xi^{\alpha-2}+c_4\xi^2\lesssim\Delta^{\alpha-2}.$$

Since $\bar{\phi}_\Delta(\xi)=0$ for $\xi\in[0,\Delta/2]$ and $\xi\geq 2$ by the construction, we have

$$\sup_{\xi\in\mathbb{R}_+}\xi^2\left|\bar{\phi}_\Delta^{(4)}(\xi)\right|\lesssim\Delta^{\alpha-2}. \tag{5.26}$$

Thus, we have

$$\left|\mathbf{E}\left[R_3\left(\hat{p};\bar{\phi}_\Delta,p\right)\right]\right|\lesssim\frac{1}{n^2\Delta^{2-\alpha}}+\frac{1}{n^3\Delta^{3-\alpha}}. \tag{5.27}$$

Define $G(x)=\frac{1}{2}(\frac{\hat{p}}{x}-1)^2$. Then, the mean value theorem stats that there exists ξ such that

$$\begin{aligned}
R_1(\hat{p};\bar{\phi}_\Delta^{(2)},p)=&\frac{\bar{\phi}_\Delta^{(4)}(\xi)}{2}(\hat{p}-\xi)\frac{\xi^2(\frac{\hat{p}}{p}-1)^2}{\hat{p}(\frac{\hat{p}}{\xi}-1)} \\
=&\frac{\bar{\phi}_\Delta^{(4)}(\xi)}{2}\frac{\xi^3(\hat{p}-p)^2}{p^2\hat{p}}.
\end{aligned}$$

Thus, we have

$$\begin{aligned}\left|\mathbf{E}\left[\frac{\hat{p}}{2n}R_1(\hat{p};\bar{\phi}_\Delta^{(2)},p)\right]\right| \leq & \mathbf{E}\left[\frac{\left|\bar{\phi}_\Delta^{(4)}(\xi)\right|}{4n}\frac{\xi^3(\hat{p}-p)^2}{p^2}\right] \\ \leq & \frac{\sup_{\xi\in\mathbb{R}_+}\xi^3\left|\bar{\phi}_\Delta^{(4)}(\xi)\right|}{4np^2}\mathbf{E}\left[(\hat{p}-p)^2\right] \\ = & \frac{1}{4n^2p}\sup_{\xi\in\mathbb{R}_+}\xi^3\left|\bar{\phi}_\Delta^{(4)}(\xi)\right|.\end{aligned}$$

In the similar manner of Eq (5.26), we have

$$\sup_{\xi\in\mathbb{R}_+}\xi^3\left|\bar{\phi}_\Delta^{(4)}(\xi)\right| \lesssim \Delta^{\alpha-1}.$$

Thus, we have

$$\left|\mathbf{E}\left[\frac{\hat{p}}{2n}R_1(\hat{p};\bar{\phi}_\Delta^{(2)},p)\right]\right| \lesssim \frac{1}{n^2p\Delta^{1-\alpha}} \leq \frac{1}{n^2\Delta^{2-\alpha}}. \tag{5.28}$$

By the assumption $\Delta \gtrsim \frac{1}{n}$, we have $\frac{1}{n^3\Delta^{3-\alpha}} \lesssim \frac{1}{n^2\Delta^{2-\alpha}}$. Assembling Eqs (5.23), (5.24), (5.27) and (5.28) gives the desired result. □

Proof of Lemma 9. From the property of the variance and the triangle inequality, we have

$$\begin{aligned}&\mathbf{Var}\left[\bar{\phi}_\Delta\left(\frac{\tilde{N}}{n}\right)-\frac{\tilde{N}}{2n^2}\bar{\phi}_\Delta^{(2)}\left(\frac{\tilde{N}}{n}\right)-\phi(p)+\frac{p\phi^{(2)}(p)}{2n}\right] \\ \leq & \mathbf{E}\left[\left(\bar{\phi}_\Delta\left(\frac{\tilde{N}}{n}\right)-\frac{\tilde{N}}{2n^2}\bar{\phi}_\Delta^{(2)}\left(\frac{\tilde{N}}{n}\right)-\phi(p)+\frac{p\phi^{(2)}(p)}{2n}\right)^2\right] \\ \leq & 2\mathbf{E}\left[\left(\bar{\phi}_\Delta\left(\frac{\tilde{N}}{n}\right)-\phi(p)\right)^2\right]+2\mathbf{E}\left[\left(\frac{\tilde{N}}{2n^2}\bar{\phi}_\Delta^{(2)}\left(\frac{\tilde{N}}{n}\right)-\frac{p\phi^{(2)}(p)}{2n}\right)^2\right].\end{aligned} \tag{5.29}$$

Applying Taylor theorem to the first term of Eq (5.29) gives

$$\begin{aligned}&\left|\bar{\phi}_\Delta\left(\frac{\tilde{N}}{n}\right)-\phi(p)\right| \\ = & \left|\phi^{(1)}(p)\left(\frac{\tilde{N}}{n}-p\right)+\frac{\phi^{(2)}(p)}{2}\left(\frac{\tilde{N}}{n}-p\right)^2+\frac{\phi^{(3)}(p)}{6}\left(\frac{\tilde{N}}{n}-p\right)^3+R_3\left(\frac{\tilde{N}}{n};\bar{\phi}_\Delta,p\right)\right|,\end{aligned}$$

where $R_3\left(\frac{\tilde{N}}{n}; \bar{\phi}_\Delta, p\right)$ denotes the reminder term of the Taylor theorem. From the triangle inequality, we have

$$\left(\bar{\phi}_\Delta\left(\frac{\tilde{N}}{n}\right) - \phi(p)\right)^2 = 4\left(\phi^{(1)}(p)\right)^2\left(\frac{\tilde{N}}{n} - p\right)^2 + \left(\phi^{(2)}(p)\right)^2\left(\frac{\tilde{N}}{n} - p\right)^4 + \frac{\left(\phi^{(3)}(p)\right)^2}{9}\left(\frac{\tilde{N}}{n} - p\right)^6 + 4\left(R_3\left(\frac{\tilde{N}}{n}; \bar{\phi}_\Delta, p\right)\right)^2. \quad (5.30)$$

The central moments for $X \sim \mathrm{Poi}(\lambda)$ are given as $\mathbf{E}[(X-\lambda)^2] = \lambda$,$\mathbf{E}[(X-\lambda)^4] = 3\lambda^2 + \lambda$, and $\mathbf{E}[(X-\lambda)^6] = 15\lambda^3 + 25\lambda^2 + \lambda$. Lemma 17, the triangle inequality and the assumption $\frac{1}{n} \gtrsim \Delta$, the expectation of the first three terms in Eq (5.30) have upper bounds as

$$\mathbf{E}\left[4\left(\phi^{(1)}(p)\right)^2\left(\frac{\tilde{N}}{n} - p\right)^2\right] \leq \frac{8W^2p^{2\alpha-1} + 8c_1^2p}{n} \lesssim \frac{p^{2\alpha-1}}{n} + \frac{p}{n},$$

$$\begin{aligned}\mathbf{E}\left[\left(\phi^{(2)}(p)\right)^2\left(\frac{\tilde{N}}{n} - p\right)^4\right] \leq& \left(2\alpha_1^2W^2p^{2\alpha-4} + c_2^2\right)\left(\frac{3p^2}{n^2} + \frac{p}{n^3}\right)\\ \lesssim& \frac{p^{2\alpha-1}}{n^2\Delta} + \frac{p^{2\alpha-1}}{n^3\Delta^2} + \frac{p}{n^2}\\ \lesssim& \frac{p^{2\alpha-1}}{n} + \frac{p}{n^2},\end{aligned}$$

and

$$\begin{aligned}\mathbf{E}\left[\frac{\left(\phi^{(3)}(p)\right)^2}{9}\left(\frac{\tilde{N}}{n} - p\right)^6\right] \leq& \left(2\alpha_2^2W^2p^{2\alpha-6} + c_3^2\right)\left(\frac{15p^3}{n^3} + \frac{25p^2}{n^4} + \frac{p}{n^5}\right)\\ \lesssim& \frac{p^{2\alpha-1}}{n^3\Delta^2} + \frac{p^{2\alpha-1}}{n^4\Delta^3} + \frac{p^{2\alpha-1}}{n^5\Delta^4} + \frac{p}{n^3}\\ \lesssim& \frac{p^{2\alpha-1}}{n} + \frac{p}{n^3}.\end{aligned}$$

From Eq (5.25), there exists ξ between p and $\hat{p}$ such that

$$\begin{aligned}
&4\mathbf{E}\left[\left(R_3\left(\frac{\tilde{N}}{n};\bar{\phi}_\Delta,p\right)\right)^2\right]\\
=&4\mathbf{E}\left[\left(\frac{\xi^3\bar{\phi}_\Delta^{(4)}(\xi)}{12p^2(\xi+\hat{p})}(\hat{p}-p)^4\right)^2\right]\\
\leq&\frac{\sup_{\xi\in\mathbb{R}_+}\left|\xi^2\bar{\phi}_\Delta^{(4)}(\xi)\right|^2}{36p^4}\mathbf{E}\left[(\hat{p}-p)^8\right]\\
\leq&\left(\frac{105}{36n^4}+\frac{490}{36n^5p}+\frac{119}{36n^6p^2}+\frac{1}{36n^7p}\right)\sup_{\xi\in\mathbb{R}_+}\left|\xi^2\bar{\phi}_\Delta^{(4)}(\xi)\right|^2,
\end{aligned}$$

where we use $\mathbf{E}[(X-\lambda)^8]=105\lambda^4+490\lambda^3+119\lambda^2+\lambda$ for $X\sim\mathrm{Poi}(\lambda)$. Since $\sup_{\xi\in\mathbb{R}_+}\left|\xi^2\bar{\phi}_\Delta^{(4)}(\xi)\right|^2\lesssim\Delta^{2\alpha-4}$ from Eq (5.26) and $\Delta\gtrsim\frac{1}{n}$ by the assumption, we have

$$\begin{aligned}
&4\mathbf{E}\left[\left(R_3\left(\frac{\tilde{N}}{n};\bar{\phi}_\Delta,p\right)\right)^2\right]\\
\lesssim&\frac{1}{n^4\Delta^{4-2\alpha}}+\frac{1}{n^5\Delta^{5-2\alpha}}+\frac{1}{n^6\Delta^{6-2\alpha}}+\frac{1}{n^7\Delta^{7-2\alpha}}\\
\lesssim&\frac{1}{n^4\Delta^{4-2\alpha}}.
\end{aligned}$$

Letting $g(p)=p\bar{\phi}_\Delta^{(2)}(p)$, application of the Taylor theorem to the second term of Eq (5.29) yields

$$\left|\frac{\tilde{N}}{2n^2}\bar{\phi}_\Delta^{(2)}\left(\frac{\tilde{N}}{n}\right)-\frac{p\phi^{(2)}(p)}{2n}\right|\leq\frac{1}{2n}\left|(\phi(2)(p)+p\phi(3)(p))\left(\frac{\tilde{N}}{n}-p\right)+R_1\left(\frac{\tilde{N}}{n};g,p\right)\right|.$$

The triangle inequality and $\mathbf{E}[(X-\lambda)^2]=\lambda$ for $X\sim\mathrm{Poi}(\lambda)$ give

$$\begin{aligned}
&\mathbf{E}\left[\left(\frac{\tilde{N}}{2n^2}\bar{\phi}_\Delta^{(2)}\left(\frac{\tilde{N}}{n}\right)-\frac{p\phi^{(2)}(p)}{2n}\right)^2\right]\\
\leq&\frac{(\phi(2)(p))^2+(p\phi(3)(p))^2}{n^2}\mathbf{E}\left[\left(\frac{\tilde{N}}{n}-p\right)^2\right]+\frac{1}{2n^2}\mathbf{E}\left[\left(R_1\left(\frac{\tilde{N}}{n};g,p\right)\right)^2\right]\\
=&\frac{p(\phi(2)(p))^2+p(p\phi(3)(p))^2}{n^3}+\frac{1}{2n^2}\mathbf{E}\left[\left(R_1\left(\frac{\tilde{N}}{n};g,p\right)\right)^2\right].
\end{aligned}$$

Applying Lemma 17 gives

$$\begin{aligned}
&\frac{p(\phi(2)(p))^2 + p(p\phi(3)(p))^2}{n^3} \\
\leq &\frac{1}{n^3}\left(2\alpha_1^2 W^2 p^{2\alpha-3} + 2pc_2^2 + 2\alpha_1^2 W^2 p^{2\alpha-3} + 2p^3 c_3^2\right) \\
\lesssim &\frac{p^{2\alpha-1}}{n^3\Delta^2} + \frac{p}{n^3} \\
\lesssim &\frac{p^{2\alpha-1}}{n} + \frac{p}{n^3}.
\end{aligned}$$

Let $\hat{p} = \frac{\tilde{N}}{n}$ and $G(x) = \frac{1}{x}(\hat{p} - x)^2$. Then, the mean value theorem gives that there exists ξ between p and $\hat{p}$ such that

$$\begin{aligned}
\mathbf{E}\left[(R_1(\hat{p}; g, p))^2\right] =&\mathbf{E}\left[\left(g^{(1)}(\xi)(\hat{p} - \xi)\frac{G(\hat{p}) - G(p)}{G^{(1)}(\xi)}\right)^2\right] \\
=&\mathbf{E}\left[\left(g^{(1)}(\xi)\frac{\xi^2(\hat{p} - p)^2}{p(\hat{p} + \xi)}\right)^2\right] \\
\leq&\left(\frac{3}{n^2} + \frac{1}{n^3 p}\right)\sup_{\xi\in\mathbb{R}_+}\left|\xi g^{(1)}(\xi)\right|^2 \\
\leq&\left(\frac{3}{n^2} + \frac{1}{n^3 p}\right)\sup_{\xi\in\mathbb{R}_+}\left|2\xi\bar{\phi}_\Delta^{(3)}(\xi) + \xi^2\bar{\phi}_\Delta^{(4)}(\xi)\right|^2 \\
\leq&\left(\frac{3}{n^2} + \frac{1}{n^3 p}\right)\left(2\sup_{\xi\in\mathbb{R}_+}\left|\xi\bar{\phi}_\Delta^{(3)}(\xi)\right|^2 + 2\sup_{\xi\in\mathbb{R}_+}\left|\xi^2\bar{\phi}_\Delta^{(4)}(\xi)\right|^2\right).
\end{aligned}$$

In the similar manner of Eq (5.26), we have

$$\sup_{\xi\in\mathbb{R}_+}\left|\xi\bar{\phi}_\Delta^{(3)}(\xi)\right|^2 \lesssim \Delta^{2\alpha-4}, \text{ and } \sup_{\xi\in\mathbb{R}_+}\left|\xi^2\bar{\phi}_\Delta^{(4)}(\xi)\right|^2 \lesssim \Delta^{2\alpha-4}.$$

Thus, we have

$$\frac{1}{2n^2}\mathbf{E}\left[\left(R_1\left(\frac{\tilde{N}}{n}; g, p\right)\right)^2\right] \lesssim \frac{1}{n^4\Delta^{4-2\alpha}} + \frac{1}{n^5\Delta^{5-2\alpha}} \lesssim \frac{1}{n^4\Delta^{4-2\alpha}}.$$

Consequently, we get the bound of the variance as

$$\frac{p^{2\alpha-1}}{n} + \frac{1}{n^4\Delta^{4-2\alpha}} + \frac{p}{n}.$$

□

Appendix 5.D Proof of Proposition 3

Proof of Proposition 3. It is obviously that if the output domain of ϕ is unbounded, i.e., there is a point $p_0 \in [0,1]$ such that $|\phi(p)| \to \infty$ as $p \to p_0$, there is no consistent estimator. Letting $p_0 = \left(\frac{W}{W \vee -c_1'}\right)$, $\phi^{(1)}(p)$ has same sign in $(0, p_0]$. Thus, for any $p \in (0, p_0]$, we have

$$\begin{aligned}
|\phi(p) - \phi(p_0)| =& \left| \int_{p_0}^{p} \phi^{(1)}(x) dx \right| \\
=& \int_{p}^{p_0} \left| \phi^{(1)}(x) \right| dx \\
\geq& W \int_{p}^{p_0} p^{-1} dx + c_1'(p_0 - p) \\
\geq& W \ln(p_0/p) + c_1'(p_0 - p).
\end{aligned}$$

Since $|\phi(p) - \phi(p_0)| \to \infty$ as $p \to 0$, ϕ is unbounded and we gets the claim. □

Appendix 5.E Additional Lemmas

Here, we introduce some additional lemmas and their proofs.

Lemma 17. *For a non-integer α, let ϕ be a m times continuously differentiable function on $(0,1]$ where $m \geq 1 + \alpha$. Suppose that there exist finite constants $W > 0$, c_m and c_m' such that*

$$\left|\phi^{(m)}(p)\right| \leq \alpha_{m-1} W p^{\alpha - m} + c_m, \text{ and } \left|\phi^{(m)}(p)\right| \geq \alpha_{m-1} W p^{\alpha - m} + c_m'.$$

Then, there exists finite constants c_{m-1} and c_{m-1}' such that

$$\left|\phi^{(m-1)}(p)\right| \leq \alpha_{m-2} W p^{\alpha - m + 1} + c_{m-1}, \text{ and } \left|\phi^{(m-1)}(p)\right| \geq \alpha_{m-2} W p^{\alpha - m + 1} + c_{m-1}',$$

where $\alpha_0 = 1$ and $\alpha_i = \prod_{j=1}^{i} (j - \alpha)$ for $i = 1, ..., m$.

Proof of Lemma 17. Let $p_m = \left(\frac{\alpha_{m-1} W}{\alpha_{m-1} W \vee -c_m'}\right)^{1/(m-\alpha)}$. Then, $\left|\phi^{(m)}(p)\right| > 0$ for $p \in (0, p_m)$. From continuousness of $\phi^{(m)}$, $\phi^{(m)}(p)$ has same sign in $p \in (0, p_m]$, and thus we have either $\phi^{(m)}(p) \geq \alpha_{m-1} W p^{\alpha - m} + c_m'$ or $\phi^{(m)}(p) \leq -\alpha_{m-1} W p^{\alpha - m} - c_m'$ in $p \in (0, p_m]$. Since $\phi^{(m-1)}$ is absolutely continuous on $(0,1]$, we have for any $p \in (0,1]$

$$\phi^{(m-1)}(p) = \phi^{(m-1)}(p_m) + \int_{p_m}^{p} \phi^{(m)}(x) dx.$$

The absolute value of the second term has an upper bound as

$$\begin{aligned}\left|\int_{p_m}^{p}\phi^{(m)}(x)dx\right| \leq& \left|\int_{p_m}^{p}\alpha_{m-1}Wx^{\alpha-m}+c_m dx\right| \\ \leq& \left|\alpha_{m-2}W\left(p_m^{\alpha-m+1}-p^{\alpha-m+1}\right)+c_m(p-p_m)\right| \\ \leq& \alpha_{m-2}Wp^{\alpha-m+1}+\left|\alpha_{m-2}Wp_m^{\alpha-m+1}+c_m(p_m-p)\right| \\ \leq& \alpha_{m-2}Wp^{\alpha-m+1}+\alpha_{m-2}Wp_m^{\alpha-m+1}+|c_m|.\end{aligned}$$

Also, we have a lower bound of the second term as

$$\begin{aligned}\left|\int_{p_m}^{p}\phi^{(m)}(x)dx\right| =& \left|\int_{p_m}^{p\wedge p_m}\phi^{(m)}(x)dx+\int_{p\wedge p_m}^{p}\phi^{(m)}(x)dx\right| \\ \geq& \left|\int_{p_m}^{p\wedge p_m}\alpha_{m-1}Wx^{\alpha-m}+c_m' dx\right|-\left|\int_{p\wedge p_m}^{p}\alpha_{m-1}Wp_m^{\alpha-m}+c_m dx\right| \\ \geq& \left|\alpha_{m-2}W\left(p_m^{\alpha-m+1}-(p\wedge p_m)^{\alpha-m+1}\right)+c_m'((p\wedge p_m)-p_m)\right| \\ & -\left|\left(\alpha_{m-1}Wp_m^{\alpha-m}+c_m\right)(p-(p\wedge p_m))\right| \\ \geq& \alpha_{m-2}W(p\wedge p_m)^{\alpha-m+1}-\alpha_{m-2}Wp_m^{\alpha-m+1}-|c_m'(p_m-(p\wedge p_m))| \\ & -\left(\alpha_{m-1}Wp_m^{\alpha-m}+c_m\right)(p-(p\wedge p_m)) \\ \geq& \alpha_{m-2}Wp^{\alpha-m+1}-\alpha_{m-2}Wp_m^{\alpha-m+1}-|c_m'|p_m \\ & -\left(\alpha_{m-1}Wp_m^{\alpha-m}+c_m\right)(1-p_m)\end{aligned}.$$

Applying the triangle inequality and the reverse triangle inequality gives

$$\left|\int_{p_m}^{p}\phi^{(m)}(x)dx\right|-\left|\phi^{(m-1)}(p_m)\right| \leq \left|\phi^{(m-1)}(p)\right| \leq \left|\int_{p_m}^{p}\phi^{(m)}(x)dx\right|+\left|\phi^{(m-1)}(p_m)\right|.$$

Thus, setting $c_{m-1} = \alpha_{m-2}Wp_m^{\alpha-m+1}+|c_m|+\left|\phi^{(m-1)}(p_m)\right|$ and $c_{m-1}' = -\alpha_{m-2}Wp_m^{\alpha-m+1}-|c_m'|p_m-(\alpha_{m-1}Wp_m^{\alpha-m}+c_m)(1-p_m)-\left|\phi^{(m-1)}(p_m)\right|$ yields the claim. □

Lemma 18. *Under Assumption 1 or Assumption 2, for any $p, p' \in [0,1]$*

$$|\phi(p)-\phi(p')| \leq \frac{W}{\alpha}|p-p'|^{\alpha}+|c_1(p-p')|.$$

Proof of Lemma 18. We can assume $p' \leq p$ without loss of generality. The absolute continuously of ϕ gives

$$|\phi(p)-\phi(p')| = \left|\int_{p'}^{p}\phi^{(1)}(x)dx\right| \leq \left|\int_{p'}^{p}\left|\phi^{(1)}(x)\right|dx\right|.$$

From Lemma 17, we have

$$\begin{aligned}|\phi(p)-\phi(p')| \leq& \left|\int_{p'}^{p}(Wx^{\alpha-1}+c_1)dx\right| \\ =& \left|\frac{W}{\alpha}(p^\alpha - p'^\alpha)+c_1(p-p')\right| \\ \leq& \frac{W}{\alpha}|p-p'|^\alpha + |c_1(p-p')|,\end{aligned}$$

where the last line is obtained since a function x^α for $\alpha \in (0,1)$ is α-Holder continuous. This is valid for the case $p'=0$. Indeed,

$$\begin{aligned}|\phi(p)-\phi(0)| =& \lim_{p'\to 0}|\phi(p)-\phi(p')| \\ \leq& \lim_{p'\to 0}\left(\frac{W}{\alpha}|p-p'|^\alpha + |c_1(p-p')|\right) \\ =& \frac{W}{\alpha}|p-0|^\alpha + |c_1(p-0)|.\end{aligned}$$

□

Lemma 19. *Given* $\alpha \in [0,1]$, $\sup_{P\in\mathcal{M}_k}\sum_{i=1}^k p_i^\alpha = k^{1-\alpha}$.

Proof of Lemma 19. If $\alpha = 1$, the claim is obviously true. Thus, we assume $\alpha < 1$. We introduce the Lagrange multiplier λ for a constraint $\sum_{i=1}^n p_i = 1$, and let the partial derivative of $\sum_{i=1}^k p_i^\alpha + \lambda(1-\sum_{i=1}^k p_i)$ with respect to p_i be zero. Then, we have

$$\alpha p_i^{\alpha-1} - \lambda = 0. \tag{5.31}$$

Since $p^{\alpha-1}$ is a monotone function, the solution of Eq (5.31) is given as $p_i = (\lambda/\alpha)^{1/(\alpha-1)}$, i.e., the values of $p_1,...,p_k$ are same. Thus, the function $\sum_{i=1}^k p_i^\alpha$ is maximized at $p_i = 1/k$ for $i = 1,...,k$. Substituting $p_i = 1/k$ into $\sum_{i=1}^k p_i^\alpha$ gives the claim. □

Lemma 20. *Given* $\alpha < 0$ *and* $\Delta \leq \frac{1}{k}$, $\sup_{P\in\mathcal{M}_k:\forall i,p_i\geq\Delta}\sum_{i=1}^k p_i^\alpha = ((1-(k-1)\Delta)^\alpha + (k-1)\Delta^\alpha) \leq k\Delta^\alpha$.

Proof. From the Karush–Kuhn–Tucker conditions, letting $P^* = (p_1^*,...,p_k^*)$ be a probability vector that attains the supremum, there exist real values λ and $\delta_i \geq 0$ such that

$$(p_i^*)^{\alpha-1} - \lambda - \delta_i = 0,$$

and $p_i^* = \Delta$ only if $\delta_i > 0$. Thus, we have

$$p_i^* = \lambda^{1/(\alpha-1)} \text{ or } p_i^* = \Delta.$$

Hence,

$$\sup_{P \in \mathcal{M}_k : \forall i, p_i \geq \Delta} \sum_{i=1}^{k} p_i^\alpha = \max_{m=1,\ldots,k-1} (m\Delta^\alpha + (k-m)(1-m\Delta)^\alpha).$$

Since $\Delta^\alpha \geq (1-m\Delta)^\alpha$ for $m = 1, ..., k-1$, the maximum is attained at $m = k-1$. Moreover, we have $(1-(k-1)\Delta)^\alpha \leq \Delta^\alpha$, and thus we get the claim. □

Chapter 6

Conclusion

In this thesis, we tackle the problems of supervised learning under fairness. Our contributions in this thesis are summarized as follows.

Model-based fairness to remove the hidden disparate impact. In Chapter 3, we deal with a problem of removing the disparate impact coming from the hidden sensitive attribute. To overcome this difficulty, we develop maximum likelihood estimation for supervised learning which ensures fairness against a given prediction model of the sensitive attribute. The experimental results demonstrate that our proposed method successes to control the trade-off between prediction accuracy and fairness even if the sensitive attribute is not contained in the training samples.

Fair empirical risk minimization with populational fairness guarantee. In Chapter 4, we develop a variant framework of ERM, in which we prove that the learned predictor ensures populational fairness. The framework achieves a fair predictor by adding the empirically evaluated fairness measure as one of regularizers, the empirical neutrality risk, into the objective function of the ERM. To guarantee populational fairness, we prove a bound on the error between the empirically evaluated fairness measure and the populationally evaluated fairness measure. As a result, we show that the error is bounded above by the Rademacher complexity of the class of the predictors and $O(\sqrt{1/n})$ term. Furthermore, we conduct the comparison experiments with an instance of our NERM framework, neutral SVM. The result shows that the neutral SVM achieves the most efficient trade-off between prediction accuracy and fairness in many datasets.

Optimal populational fairness evaluation. In Chapter 5, we investigate a minimax (rate-)optimal estimator for the additive functional, which is a scalar value parameterized by a function ϕ. The additive functional covers many variants of entropy, and thus the estimator of them is useful when we want to evaluate the populational fairness. As a result, we show that the minimax rate is characterized by the divergence speed of the fourth derivative of ϕ. We reveal that if the divergence speed is faster than a certain speed, there is no consistent estimator. Moreover, we show that if the divergence speed is in a range indexed by $\alpha \in (0, 1)$, the minimax rate is $\frac{k^2}{(n \ln n)^{2\alpha}} + \frac{k^{2-2\alpha}}{n}$.

As mentioned in the introduction, social communities intensively demand fairness for machine learning algorithms. To meet such requirements, our contributions solve the largest issues in developing and analyzing the machine learning algorithms with a guarantee on fairness.

Bibliography

Acharya, Jayadev, Alon Orlitsky, Ananda Theertha Suresh, and Himanshu Tyagi (2015). "The Complexity of Estimating Rényi Entropy". In: *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2015, San Diego, CA, USA, January 4-6, 2015*. SIAM, pp. 1855–1869. DOI: 10.1137/1.9781611973730.124.

Achieser, Naum I (2013). *Theory of approximation*. Courier Corporation.

Akaike, Hirotugu (1998). "Information theory and an extension of the maximum likelihood principle". In: *Selected Papers of Hirotugu Akaike*. Springer, pp. 199–213.

Antos, András and Ioannis Kontoyiannis (2001). "Convergence properties of functional estimates for discrete distributions". In: *Random Structures & Algorithms* 19.3-4, pp. 163–193. ISSN: 1098-2418. DOI: 10.1002/rsa.10019.

Barocas, Solon and Andrew D. Selbst (2016). *Big Data's Disparate Impact*. Tech. rep. California Law Review, Inc.

Bartlett, Peter L. and Shahar Mendelson (2002). "Rademacher and Gaussian Complexities: Risk Bounds and Structural Results". In: *Journal of Machine Learning Research* 3, pp. 463–482.

Bolukbasi, Tolga, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Tauman Kalai (2016). "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings". In: *Advances in Neural Information Processing Systems 29*. Ed. by Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett. Barcelona, Spain, pp. 4349–4357.

Bu, Y., S. Zou, Y. Liang, and V. V. Veeravalli (2016). "Estimation of KL divergence between large-alphabet distributions". In: *2016 IEEE International Symposium on Information Theory (ISIT)*, pp. 1118–1122. DOI: 10.1109/ISIT.2016.7541473.

Cai, T Tony, Mark G Low, et al. (2011). "Testing composite hypotheses, Hermite polynomials and optimal estimation of a nonsmooth functional". In: *The Annals of Statistics* 39.2, pp. 1012–1041.

Calders, Toon and Sicco Verwer (2010). "Three Naive Bayes Approaches for Discrimination-Free Classification". In: *Data Mining and Knowledge Discovery* 21.2, pp. 277–292.

Dhillon, Inderjit S, Subramanyam Mallela, and Dharmendra S Modha (2003). "Information-theoretic co-clustering". In: *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 89–98.

Ditzian, Zeev and Vilmos Totik (2012). *Moduli of smoothness*. Vol. 9. Springer Science & Business Media.

Dutch Central Bureau for Statistics (2001). *Volkstelling*.

Dwork, Cynthia, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel (2012). "Fairness through awareness". In: *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. Ed. by Shafi Goldwasser. Cambridge, MA, USA: ACM, pp. 214–226.

Feldman, Michael, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian (2015). "Certifying and Removing Disparate Impact". In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Ed. by Longbing Cao, Chengqi Zhang, Thorsten Joachims, Geoffrey I. Webb, Dragos D. Margineantu, and Graham Williams. Sydney, NSW, Australia: ACM, pp. 259–268. DOI: 10.1145/2783258.2783311.

Frank, A. and A. Asuncion (2010). *UCI Machine Learning Repository*. URL: http://archive.ics.uci.edu/ml.

Fukuchi, Kazuto, Jun Sakuma, and Toshihiro Kamishima (2013). "Prediction with Model-Based Neutrality". In: *Machine Learning and Knowledge Discovery in Databases - European Conference*. Ed. by Hendrik Blockeel, Kristian Kersting, Siegfried Nijssen, and Filip Zelezný. Vol. 8189. Lecture Notes in Computer Science Part II. Prague, Czech Republic: Springer, pp. 499–514. ISBN: 978-3-642-40990-5. DOI: 10.1007/978-3-642-40991-2_32.

Grassberger, Peter (1988). "Finite sample corrections to entropy and dimension estimates". In: *Physics Letters A* 128.6, pp. 369–373.

Gu, Yu, Andrew McCallum, and Don Towsley (2005). "Detecting anomalies in network traffic using maximum entropy estimation". In: *Proceedings of the 5th ACM SIGCOMM conference on Internet Measurement*. USENIX Association, pp. 32–32.

Han, Y., J. Jiao, and T. Weissman (2016). "Minimax rate-optimal estimation of KL divergence between discrete distributions". In: *2016 International*

Symposium on Information Theory and Its Applications (ISITA), pp. 256–260.

Han, Yanjun, Jiantao Jiao, and Tsachy Weissman (2015). "Does Dirichlet Prior Smoothing Solve the Shannon Entropy Estimation Problem?" In: *CoRR* abs/1502.00327. URL: http://arxiv.org/abs/1502.00327.

Hardt, Moritz, Eric Price, and Nati Srebro (2016). "Equality of Opportunity in Supervised Learning". In: *Advances in Neural Information Processing Systems 29*. Ed. by Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett. Barcelona, Spain, pp. 3315–3323.

Holste, D, I Grosse, and H Herzel (1998). "Bayes' estimators of generalized entropies". In: *Journal of Physics A: Mathematical and General* 31.11, p. 2551.

Jabbari, Shahin, Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth (2017). "Fairness in Reinforcement Learning". In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by Doina Precup and Yee Whye Teh. Vol. 70. Proceedings of Machine Learning Research. Sydney, NSW, Australia: PMLR, pp. 1617–1626.

Jiao, J., Y. Han, and T. Weissman (2016). "Minimax estimation of the L1 distance". In: *2016 IEEE International Symposium on Information Theory (ISIT)*, pp. 750–754. DOI: 10.1109/ISIT.2016.7541399.

Jiao, Jiantao, Kartik Venkat, Yanjun Han, and Tsachy Weissman (2015). "Minimax estimation of functionals of discrete distributions". In: *Information Theory, IEEE Transactions on* 61.5, pp. 2835–2885.

Joseph, Matthew, Michael Kearns, Jamie H. Morgenstern, and Aaron Roth (2016). "Fairness in Learning: Classic and Contextual Bandits". In: *Advances in Neural Information Processing Systems 29*. Ed. by Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett. Barcelona, Spain, pp. 325–333.

Kamiran, Faisal, Toon Calders, and Mykola Pechenizkiy (2010). "Discrimination Aware Decision Tree Learning". In: *The 10th IEEE International Conference on Data Mining*. IEEE Computer Society, pp. 869–874.

Kamishima, Toshihiro, Shotaro Akaho, Hideki Asoh, and Jun Sakuma (2012a). "Enhancement of the Neutrality in Recommendation". In: *Proceedings of the 2nd Workshop on Human Decision Making in Recommender Systems*. Vol. 893. CEUR Workshop Proceedings. CEUR-WS.org, pp. 8–14.

Kamishima, Toshihiro, Shotaro Akaho, Hideki Asoh, and Jun Sakuma (2012b). "Fairness-Aware Classifier with Prejudice Remover Regularizer". In: *Machine Learning and Knowledge Discovery in Databases - European Conference*. Ed. by Peter A. Flach, Tijl De Bie, and Nello Cristianini. Vol. 7524. Lecture Notes in Computer Science Part II. Bristol, UK: Springer, pp. 35–50.

Kamishima, Toshihiro, Shotaro Akaho, Hideki Asoh, and Jun Sakuma (2013). "Efficiency Improvement of Neutrality-Enhanced Recommendation". In: *Proceedings of the 3rd Workshop on Human Decision Making in Recommender Systems in conjunction with the 7th ACM Conference on Recommender Systems (RecSys 2013)*. Vol. 1050. CEUR Workshop Proceedings. CEUR-WS.org, pp. 1–8.

Lake, Douglas E and J Randall Moorman (2011). "Accurate estimation of entropy in very short". In: *Am J Physiol Heart Circ Physiol* 300, H319–H325.

Le Cam, Lucien M (1986). *Asymptotic Methods in Statistical Theory*. New York, NY, USA: Springer-Verlag New York, Inc. ISBN: 0-387-96307-3.

Lepski, Oleg, Arkady Nemirovski, and Vladimir Spokoiny (1999). "On estimation of the L_r norm of a regression function". In: *Probability theory and related fields* 113.2, pp. 221–253.

Miller, G. A. (1955). *Note on the bias of information estimates*.

Munoz, Cecilia, Megan Smith, and D J Patil (2016). *Big data: A report on algorithmic systems, opportunity, and civil rights*. Tech. rep. The White House: Executive Office of the President.

Nemenman, Ilya, William Bialek, and Rob de Ruyter van Steveninck (2004). "Entropy and information in neural spike trains: Progress on the sampling problem". In: *Physical Review E* 69.5, p. 056111.

Pan, Sinno Jialin and Qiang Yang (2010). "A Survey on Transfer Learning". In: *IEEE Trans. Knowl. Data Eng.* 22.10, pp. 1345–1359.

Paninski, Liam (2004). "Estimating entropy on m bins given fewer than m samples". In: *IEEE Transactions on Information Theory* 50.9, pp. 2200–2203.

Pariser, Eli (2011). *The Filter Bubble: What The Internet Is Hiding From You*. London: Viking.

Pedreschi, Dino, Salvatore Ruggieri, and Franco Turini (2008). "Discrimination-aware data mining". In: *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, pp. 560–568.

Peng, Hanchuan, Fuhui Long, and Chris Ding (2005). "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy". In: *IEEE Transactions on pattern analysis and machine intelligence* 27.8, pp. 1226–1238.

Petrushev, Penco Petrov and Vasil Atanasov Popov (2011). *Rational approximation of real functions*. Vol. 28. Cambridge University Press.

Podesta, John, Penny Pritzker, Ernest J Moniz, John Holdern, and Jeffrey Zients (2014). *Big Data - Seizing Opportunities, Preserving Values*. Tech. rep. The White House: Executive Office of the President.

Quinlan, J. Ross (1986). "Induction of decision trees". In: *Machine learning* 1.1, pp. 81–106.

Remez, Eugene Y (1934). "Sur la détermination des polynômes d'approximation de degré donnée". In: *Comm. Soc. Math. Kharkov* 10, pp. 41–63.

Romei, Andrea and Salvatore Ruggieri (2013). "A multidisciplinary survey on discrimination analysis". In: *The Knowledge Engineering Review*, pp. 1–57. ISSN: 1469-8005.

Ruggieri, Salvatore, Dino Pedreschi, and Franco Turini (2010). "DCUBE: discrimination discovery in databases". In: *Proceedings of the ACM SIGMOD International Conference on Management of Data*. ACM, pp. 1127–1130.

Schober, S. (2013). "Some worst-case bounds for Bayesian estimators of discrete distributions". In: *2013 IEEE International Symposium on Information Theory*, pp. 2194–2198. DOI: 10.1109/ISIT.2013.6620615.

Schürmann, Thomas and Peter Grassberger (1996). "Entropy estimation of symbol sequences". In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 6.3, pp. 414–427.

Shor, N. Z., Krzysztof C. Kiwiel, and Andrzej Ruszcayǹski (1985). *Minimization Methods for Non-differentiable Functions*. New York, NY, USA: Springer-Verlag New York, Inc. ISBN: 0-387-12763-1.

Spitzbart, A (1960). "A generalization of Hermite's interpolation formula". In: *The American Mathematical Monthly* 67.1, pp. 42–46.

Suzuki, Taiji, Masashi Sugiyama, Takafumi Kanamori, and Jun Sese (2009). "Mutual information estimation reveals global associations between stimuli and biological processes". In: *BMC Bioinformatics* 10.S-1.

Sweeney, Latanya (2013). "Discrimination in online ad delivery". In: *Communications of the ACM* 56.5, pp. 44–54. DOI: 10.1145/2447976.2447990.

Thanh, Binh Luong, Salvatore Ruggieri, and Franco Turini (2011). "k-NN as an implementation of situation testing for discrimination discovery and prevention". In: *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, pp. 502–510.

Timan, A-F, J Berry, and J Cossar (1965). "Theory of approximation of functions of a real variable." In:

Torkkola, Kari (2003). "Feature Extraction by Non-Parametric Mutual Information Maximization". In: *Journal of Machine Learning Research* 3, pp. 1415–1438.

Tsybakov, Alexandre B. (2009). *Introduction to Nonparametric Estimation*. Springer series in statistics. Springer. ISBN: 978-0-387-79051-0. DOI: 10.1007/b13794.

Valiant, Gregory and Paul Valiant (2011a). "Estimating the unseen: an n/log(n)-sample estimator for entropy and support size, shown optimal

via new CLTs". In: *Proceedings of the 43rd ACM Symposium on Theory of Computing, STOC 2011, San Jose, CA, USA, 6-8 June 2011*. Ed. by Lance Fortnow and Salil P. Vadhan. ACM, pp. 685–694. DOI: 10.1145/1993636.1993727. URL: http://doi.acm.org/10.1145/1993636.1993727.

Valiant, Gregory and Paul Valiant (2011b). "The Power of Linear Estimators". In: *IEEE 52nd Annual Symposium on Foundations of Computer Science, FOCS 2011, Palm Springs, CA, USA, October 22-25, 2011*. Ed. by Rafail Ostrovsky. IEEE Computer Society, pp. 403–412. DOI: 10.1109/FOCS.2011.81.

Vapnik, Vladimir N (1998). "Statistical learning theory". In:

Wächter, Andreas and Lorenz T. Biegler (2006). "On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming". In: *Mathematical Programming* 106.1, pp. 25–57.

Woodworth, Blake E., Suriya Gunasekar, Mesrob I. Ohannessian, and Nathan Srebro (2017). "Learning Non-Discriminatory Predictors". In: *Proceedings of the 30th Conference on Learning Theory, COLT 2017, Amsterdam, The Netherlands, 7-10 July 2017*. Ed. by Satyen Kale and Ohad Shamir. Vol. 65. Proceedings of Machine Learning Research. PMLR, pp. 1920–1953.

Wu, Y. and P. Yang (2015). "Chebyshev polynomials, moment matching, and optimal estimation of the unseen". In: *ArXiv e-prints*. eprint: 1504.01227.

Wu, Yihong and Pengkun Yang (2016). "Minimax rates of entropy estimation on large alphabets via best polynomial approximation". In: *IEEE Transactions on Information Theory* 62.6, pp. 3702–3720.

Zafar, Muhammad Bilal, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi (2017). "Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment". In: *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, pp. 1171–1180.

Zahl, Samuel (1977). "Jackknifing An Index of Diversity". In: *Ecology* 58.4, pp. 907–913. ISSN: 00129658, 19399170.

Zemel, Richard S., Yu Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork (2013). "Learning Fair Representations". In: *Proceedings of The 30th International Conference on Machine Learning*. Vol. 28. JMLR Workshop and Conference Proceedings. Atlanta, GA, USA: JMLR.org, pp. 325–333.

Zliobaite, Indre, Faisal Kamiran, and Toon Calders (2011a). "Handling Conditional Discrimination". In: *The 11th IEEE International Conference on Data Mining*. IEEE Computer Society, pp. 992–1001.

Zliobaite, Indre, Faisal Kamiran, and Toon Calders (2011b). "Handling conditional discrimination". In: *Data Mining (ICDM), 2011 IEEE 11th International Conference on*. IEEE, pp. 992–1001.

Publications

Fukuchi, Kazuto, Toshihiro Kamishima, and Jun Sakuma (2015). "Prediction with Model-Based Neutrality". In: *IEICE Transactions* 98-D.8, pp. 1503–1516.

Fukuchi, Kazuto and Jun Sakuma (2014). "Neutralized Empirical Risk Minimization with Generalization Neutrality Bound". In: *Machine Learning and Knowledge Discovery in Databases - European Conference*. Ed. by Toon Calders, Floriana Esposito, Eyke Hüllermeier, and Rosa Meo. Vol. 8724. Lecture Notes in Computer Science Part I. Nancy, France: Springer, pp. 418–433. ISBN: 9783662448472. DOI: 10.1007/978-3-662-44848-9_27. eprint: 1511.01987.

Fukuchi, Kazuto and Jun Sakuma (2017). "Minimax optimal estimators for additive scalar functionals of discrete distributions". In: *2017 IEEE International Symposium on Information Theory, ISIT 2017, Aachen, Germany, June 25-30, 2017*. IEEE, pp. 2103–2107. DOI: 10.1109/ISIT.2017.8006900.

Fukuchi, Kazuto, Jun Sakuma, and Toshihiro Kamishima (2013). "Prediction with Model-Based Neutrality". In: *Machine Learning and Knowledge Discovery in Databases - European Conference*. Ed. by Hendrik Blockeel, Kristian Kersting, Siegfried Nijssen, and Filip Zelezný. Vol. 8189. Lecture Notes in Computer Science Part II. Prague, Czech Republic: Springer, pp. 499–514. ISBN: 978-3-642-40990-5. DOI: 10.1007/978-3-642-40991-2_32.