Statistica Sinica **37** (2027), 1-23
doi:https://doi.org/10.5705/ss.202024.0310

AUXILIARY LEARNING AND ITS STATISTICAL UNDERSTANDING

Hanchao Yan[1], Feifei Wang*[2], Chuanxin Xia[3] and Hansheng Wang[3]

[1]*Peking University,* [2]*Renmin University of China*
and [3]*Ministry of Commerce People's Republic of China*

Abstract: Modern statistical analysis often encounters high-dimensional problems but with a limited sample size. It poses great challenges to traditional statistical estimation methods. In this work, we adopt auxiliary learning to solve the estimation problem in high-dimensional settings. We start with the linear regression setup. To improve the statistical efficiency of the parameter estimator for the primary task, we consider several auxiliary tasks, which share the same covariates with the primary task. Then a weighted estimator for the primary task is developed, which is a linear combination of the ordinary least squares estimators of both the primary task and auxiliary tasks. The optimal weight is analytically derived and the statistical properties of the corresponding weighted estimator are studied. We then extend the weighted estimator to generalized linear regression models. Extensive numerical experiments are conducted to verify our theoretical results. Last, a deep learning-related real-data example of smart vending machines is presented for illustration purposes.

Key words and phrases: Auxiliary learning, deep learning, high-dimensional data analysis, ordinary least squares estimator, smart vending machines.

1. Introduction

Modern statistical analysis frequently encounters datasets characterized by an extensive number of features, which even surpass the sample size. This is attributed to advancements in state-of-the-art technology, which facilitates more convenient collection of data and computing (Gao et al., 2022; Jiang, Guo and Wang, 2023; Li et al., 2024). For instance, in genetic studies, there are hundreds of thousands of single-nucleotide polymorphisms (SNPs) that serve as potential predictors. However, the sample size is limited due to the expensive cost of sequencing. In deep learning-related studies, millions or even billions of features are extracted from texts, images, and other forms of unstructured data. However, the sample size is also relatively limited due to the high cost of labeling. The fact that the feature dimension is larger than the sample size imposes significant challenges to classical statistical analysis. Consider the linear regression model as an example. To estimate the regression coefficient β, the popularly used ordinary

*Corresponding author. E-mail: feifei.wang@ruc.edu.cn