Learning from Biased and Costly Data Sources: Minimax-optimal Data Collection under a Budget

Michael O. Harding MOHARDING@WISC.EDU
Department of Statistics, University of Wisconsin-Madison
Vikas Singh VSINGH@BIOSTAT.WISC.EDU
Department of Biostatistics, University of Wisconsin-Madison
Kirthevasan Kandasamy KANDASAMY@CS.WISC.EDU
Department of Computer Sciences, University of Wisconsin-Madison

Editors: Steve Hanneke and Tor Lattimore

Abstract

Data collection is a critical component of modern statistical and machine learning pipelines, particularly when data must be gathered from multiple heterogeneous sources to study a target population of interest. In many use cases, such as medical studies or political polling, different sources incur different sampling costs. Observations often have associated group identities—for example, health markers, demographics, or political affiliations—and the relative composition of these groups may differ substantially, both among the source populations and between sources and target population. Moreover, while group proportions are often known at the source and population levels, individual group membership may only be revealed after data collection.

In this work, we study multi-source data collection under a fixed budget, focusing on the estimation of population means and group-conditional means. We show that naive data collection strategies (e.g. attempting to "match" the target distribution) or relying on standard estimators (e.g. sample mean) can be highly suboptimal. Instead, we develop a sampling plan which maximizes the *effective sample size*—the total sample size divided by $D_{\chi^2}(q \,||\, \overline{p}) + 1$, where q is the target distribution, $\overline{p}$ is the aggregated source distribution, and D_{χ^2} is the χ^2-divergence. We pair this sampling plan with a classical post-stratified estimator, which is able to leverage large but systematically biased datasets, and upper bound its risk. We provide lower bounds with exactly matching leading terms, proving that our approach achieves the budgeted minimax-optimal risk up to additive lower order terms.

Our techniques also extend to prediction problems when minimizing the excess risk. In this setting, we pair the effective-sample-size-maximizing sampling plan with a weighted empirical risk minimzer and upper bound its risk. A key contribution to this end is the development of a general information-theoretic lower bound framework for prediction problems under possibly differing source and target distributions, which may be of independent interest outside of this work. We apply this framework to the case of binary classification to establish a lower bound which matches the upper bound of our proposed approach up to a $\sqrt{K/q_{\min}}$-factor, where K is the number of groups and $q_{\min}$ is the smallest group identity probability under the target distribution q. Our framework enables us to respect the information geometry of the problem via dependence on $D_{\chi^2}(q \,||\, \overline{p})$ that matches the upper bound, which was not possible with existing techniques.[1]

Keywords: Data collection, Multi-source learning, Minimax optimality, Covariate shift

Acknowledgments

The authors would like to thank Michael A. Newton for his role as a co-advisor to Michael O. Harding through the NSF's TRIPODS Program for the Institute for Foundations of Data Science. This reseach was supported by NSF Awards IIS-2441796 and DMS-2023239.

1. Extended abstract. Full version appears as [arXiv:2602.17894,v2]