Collaborative Learning of Discrete Distributions under Heterogeneity and Communication Constraints

Xinmeng Huang [*], **Donghwan Lee** [*]
Graduate Group in Applied Mathematics and Computational Science
University of Pennsylvania
Philadelphia, PA 19104
{xinmengh, dh7401}@sas.upenn.edu

Edgar Dobriban
Department of Statistics and Data Science
University of Pennsylvania
Philadelphia, PA 19104
dobriban@wharton.upenn.edu

Hamed Hassani
Department of Electrical and Systems Engineering
University of Pennsylvania
Philadelphia, PA 19104
hassani@seas.upenn.edu

Abstract

In modern machine learning, users often have to collaborate to learn distributions that generate the data. Communication can be a significant bottleneck. Prior work has studied homogeneous users—*i.e.*, whose data follow the same discrete distribution—and has provided optimal communication-efficient methods. However, these methods rely heavily on homogeneity, and are less applicable in the common case when users' discrete distributions are heterogeneous. Here we consider a natural and tractable model of heterogeneity, where users' discrete distributions only vary sparsely, on a small number of entries. We propose a novel two-stage method named SHIFT: First, the users collaborate by communicating with the server to learn a central distribution; relying on methods from robust statistics. Then, the learned central distribution is fine-tuned to estimate the individual distributions of users. We show that our method is minimax optimal in our model of heterogeneity and under communication constraints. Further, we provide experimental results using both synthetic data and n-gram frequency estimation in the text domain, which corroborate its efficiency.

1 Introduction

Research on learning from data distributed over multiple computational units (machines, users, devices) has grown in recent years, as data is commonly generated by multiple users, such as smart devices and wireless sensors. While many works focus on learning predictive models with distributed data, learning the data distribution itself is also increasingly popular [56, 3, 14, 15, 60]. In various applications, it is often required to reconstruct the data distribution from scattered measurements. Examples include sensor networks and P2P (Peer2Peer) systems, load balancing and query processing (see, *e.g.*, [41, 61, 46] and references therein). In these scenarios, communication costs and bandwidth are often bottlenecks on the performance of learning algorithms [22, 5, 29, 17]. The bottlenecks become even more severe in federated analytics [31], where many users coordinate with a server to learn central models, while communication via the wireless links is typically expensive and operates at low rates.

[*] Equal Contribution

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

Table 1: Estimation error $\mathbb{E}[\|\widehat{\mathbf{p}}^t - \mathbf{p}^t\|_2^2]$ of various methods when n is sufficiently large: $\mathbf{p}^t$ is the test distribution, $\widehat{\mathbf{p}}^t$ is the estimator, $\boldsymbol{\delta}^t = T^{-1}\sum_{t'\in[T]}\mathbf{p}^{t'} - \mathbf{p}^t$ is a non-vanishing measure of heterogeneity. See Section 1.1 for other notations. Constants and logarithmic factors are omitted for clarity. The "data usage" column indicates whether the estimate is obtained for each cluster separately or by pooling data.

Method	Estimation Error	Data Usage	Bound Type
Unif. Group./Hash. [26]	$O\left(\frac{d}{2^b n}\right)$	Separate	Upper
Unif. Group./Hash. [26]	$O\left(\|\boldsymbol{\delta}^t\|_2^2 + \frac{d}{2^b Tn}\right)$	Pool	Upper
Localize-then-Refine [15]	$O\left(\frac{\|\mathbf{p}^t\|_{1/2}}{2^b n}\right)^*$	Separate	Upper
Localize-then-Refine [15]	$O\left(\|\boldsymbol{\delta}^t\|_2^2 + \frac{\|\frac{1}{T}\sum_{t'\in[T]}\mathbf{p}^{t'}\|_{1/2}}{2^b Tn}\right)^*$	Pool	Upper
SHIFT (Theorem 1)	$\tilde{O}\left(\frac{\max\{2^b, s\}}{2^b n} + \frac{d}{2^b Tn}\right)$	–	Upper
SHIFT (Theorem 3)	$\Omega\left(\frac{\max\{2^b, s\}}{2^b n} + \frac{d}{2^b Tn}\right)$	–	Lower

* This method [15] requires interactive communication protocols, while other methods are non-interactive.

This paper considers learning high-dimensional discrete distributions from user data in the distributed setting. Many communication-efficient methods have been proposed, and their optimality under communication constraints has been established under various models [26, 6, 27, 2, 3, 14, 15]. However, the key challenge of *heterogeneity*, *i.e.*, that users' distributions can differ, is rarely considered. Heterogeneity is common, as users inevitably have unique characteristics [13]. Meanwhile, heterogeneity can cause a significant performance drop for learning algorithms designed only for i.i.d data [37, 34, 20, 59]. To use all the data, one needs to learn some central structure, transferable to all individual users. Then one may locally learn the unique components for each user [21, 53, 16, 57].

To study this paradigm, we first need to introduce a suitable model of heterogeneity. We consider, as an example, the heterogeneous frequencies of words across different texts, *e.g.*, news articles, books, plays (tragedies and comedies), viewed as users. Most words appear with almost the same probabilities in different texts, however, a few can have very different probabilities, such as "sorrowful" being common in tragedies and "convivial" being common in comedies. Motivated by this, we formulate a model of sparse heterogeneity. Specifically, suppose that the discrete distributions of all users differ from an underlying central distribution in at most s entries, where s is much smaller than the dimension d. Sparse heterogeneity is relevant to applications such as recommendation systems [28, 42, 35, 7] and medical risk scoring [50, 43, 40].

However, given data generated by multiple distributions with sparse heterogeneity, previous works [26, 2, 3, 15] either do not use all the data, or suffer from bias due to heterogeneity that does not vanish as the sample size increases. Here we propose a novel **s**parse **h**eterogeneity **i**nspired collaboration and **f**ine-**t**uning method (SHIFT) where we first collaboratively learn the central distribution, and then fine-tune the central estimate to individual distributions. Our method makes full use of heterogeneous data, leading to a significant improvement in error rates compared to prior methods. See Table 1 for an overview, explained in detail later.

1.1 Contributions

We consider the problem of learning d-dimensional distributions with s-sparse heterogeneity. We assume there are T clusters of user datapoints, and allow each datapoint to be transmitted with b bits of information to the server. Our setting embraces heterogeneous data and thus is a significant generalization of the models from [26, 6, 27, 2]. Our technical contributions are as follows:

- We propose the SHIFT method to learn heterogeneous distributions with collaboration and tuning, in a sample-efficient manner. Our method can in principle be used with an arbitrary robust estimate for the probability of each entry/coordinate. When entry-wise median and trimmed mean are used, we provide upper bounds on the estimation error of individual distributions in the ℓ_2 and ℓ_1 norms. We show a factor of $\min\{T, d/\max\{s, 2^b\}\}$ improvement in sample complexity compared to previous works; showing the benefit of collaboration (large T) and sparsity (small s), despite communication constraints and heterogeneity.

- To justify the optimality of our method, we prove minimax lower bounds on the estimation errors for individual distributions in the ℓ_2 and ℓ_1 norms, holding for all, possibly interactive, methods. These lower bounds, combined with our upper bounds, imply that our median-based method is minimax optimal.
- We support our method with experiments on both synthetic and empirical datasets, showing a significant improvement over previous methods.

1.2 Related Works

Learning with heterogeneity. Learning with heterogeneity is commonly found in the broader context of multi-task learning [12, 52, 8] and federated learning [4, 44, 24, 47], where a central model or representation is learned from multiple heterogeneous datasets. These central representations can be useful for few-shot learning tasks [51, 23] due to their ability to adapt to new tasks efficiently. In heterogeneous linear regression, [53, 21] show improved sample complexities by assuming a low dimensional central representation, compared to the i.i.d. setting [24, 36]. Related results are proved by [16] for personalized federated learning. [57] study a bandit problem where the unknown parameter in each dataset equals a global parameter plus a sparse instance-specific term. We study a different setting: learning distributions with sparse heterogeneity under communication constraints.

Estimating distributions under communication constraints. Estimating discrete distributions has a rich literature [10, 18, 45, 19]. Under communication constraints, [26, 6, 27, 2] consider the non-interactive scenario and establish the minimax optimal rates, in terms of data dimension and communication budget, via potentially shared randomness, when all users' data is homogeneous. The optimality for the general interactive (blackboard) methods is developed by [1]. A few works study the estimation of sparse distributions. In particular, [3] consider s-sparse distributions and establish minimax optimal rates under communication and privacy constraints, which are further improved by localization strategies in [14]. Complementary to minimax rates, [15] provides pointwise rates, governed by the half-norm of the distribution instead of its dimension. Our setting embraces heterogeneous data, and thus is a generalization of the one studied in above works.

Robust estimation & learning. Robust statistics and learning study algorithms resilient to unknown data corruption [30, 25]. The median-of-means method [32, 38, 39], partitions the data into subsets, computes an estimate from each, and takes their median. Similarly, some works study robustness from the optimization perspective, proposing to robustly aggregate gradients of the loss functions [48, 49, 9, 58]. We adapt some analysis techniques from [39, 58] to the significantly different setting of estimation with heterogeneity and communication constraints.

1.3 Notations

Throughout the paper, for an integer $d \geq 1$, we write $[d]$ for both $\{1, \ldots, d\}$ and $\{e_1, \ldots, e_d\} \subseteq \mathbb{R}^d$, where e_k is the k-th canonical basis vector of $\mathbb{R}^d$. For a vector $\mathbf{v} \in \mathbb{R}^d$, we refer to the entries of $\mathbf{v}$ by both $[\mathbf{v}]_1, \ldots, [\mathbf{v}]_d$ and $v_1, \ldots, v_d$. We denote $\|\mathbf{v}\|_p = (\sum_{k \in [d]} |v_k|^p)^{\frac{1}{p}}$ for all $p > 0$ with $\|\mathbf{v}\|_0$ defined additionally as the number of non-zero entries. We let $\mathcal{P}_d := \{\mathbf{p} = (p_1, \ldots, p_d) \in [0,1]^d : p_1 + \cdots + p_d = 1\}$ be the simplex of all d-dimensional discrete probability distributions. For $\mathbf{p} \in \mathcal{P}_d$, we denote by $\mathbb{B}_s(\mathbf{p})$ the s-distinct neighborhood $\{\mathbf{p}' \in \mathcal{P}_d : \|\mathbf{p}' - \mathbf{p}\|_0 \leq s\}$. For a random variable X, we denote n i.i.d. copies of X by $X^{[n]}$. Given any index set $\mathcal{I}$, we write $|\mathcal{I}|$ for its cardinality and denote by $[\mathbf{v}]_{\mathcal{I}}$ the sub-vector $([\mathbf{v}]_k)_{k \in \mathcal{I}}$ indexed by $\mathcal{I}$. We use the Bachmann-Landau asymptotic notations $\Omega(\cdot)$, $\Theta(\cdot)$, $O(\cdot)$ to hide constant factors, and use $\tilde{\Omega}(\cdot)$, $\tilde{O}(\cdot)$ to also hide logarithmic factors. We denote the categorical distribution with class probability vector $\mathbf{p} \in \mathcal{P}_d$ by $\mathrm{Cat}(\mathbf{p})$. We use $\check{\cdot}$ and $\hat{\cdot}$ to indicate the intermediate estimate and the final estimate, respectively.

2 Problem Setup

We consider the problem of collaboratively learning distributions defined according to the following model of heterogeneity (see Figure 1 for an illustration). There are $T \geq 1$ clusters $\{\mathcal{C}^t \triangleq (X^{t,j})_{j \in [n]}$:

$t \in [T]\}$ of user datapoints, each of which contains n i.i.d. local datapoints. Each datapoint $X^{t,j}$ is in a one-hot format, *i.e.*, $X^{t,j} \in \{e_1, e_2, \ldots, e_d\}$, and follows the categorical distribution $\mathrm{Cat}(\mathbf{p}^t)$ where $\mathbf{p}^t \in \mathcal{P}_d$ is unknown. Thus, user datapoints in the same cluster have an identical distribution $\mathbf{p}^t$, while the distribution $\mathbf{p}^t$ can vary, *i.e.*, be heterogeneous, across clusters $t \in [T]$. The datapoint $X^{t,j}$ is encoded by its user into a message $Y^{t,j}$, and then transmitted to a central server. We assume that the message sent by each datapoint is encoded into no more than b bits and b can be significantly smaller than $\log_2 d$ so that the communication is efficient. We also assume the server knows which cluster $t \in [T]$ each $Y^{t,j}$ belongs to, as well as the number of clusters T. This paper mainly addresses the communication bottleneck and does not involve privacy concerns.

The goal here is to collaboratively learn the distributions $\mathbf{p}^t$ from the collection of messages $\{Y^{t',[n]} \triangleq (Y^{t',j})_{j\in[n]} : t' \in [T]\}$ despite heterogeneity. More precisely, we aim to design per-cluster estimators $\widehat{\mathbf{p}}^t : \{Y^{t',[n]} : t' \in [T]\} \to \mathcal{P}_d$ to minimize the ℓ_2 errors

$$\mathbb{E}[\|\widehat{\mathbf{p}}^t - \mathbf{p}^t\|_2^2], \quad \text{for all } t \in [T].$$

We also study the widely-used ℓ_1 error metric (in addition to the ℓ_2 metric). When $T = 1$, *i.e.*, all user datapoints are homogeneous and there is a single distribution to learn, the problem reduces to the one studied by [26, 6, 27, 2, 1].

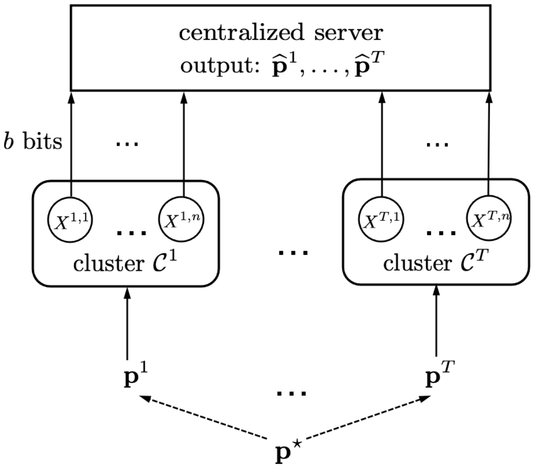

Figure 1: Learning distributions with heterogeneity and communication constraints.

Model of heterogeneity. In heterogeneous settings, collaboration among the users is most beneficial if the local distributions are related. We model this by assuming that the local distributions are sparse perturbations of an unknown *central distribution* $\mathbf{p}^\star \in \mathcal{P}_d$. The distribution $\mathbf{p}^t$ of each cluster t differs from $\mathbf{p}^\star$ in at most $s \geq 0$ entries:

$$\|\mathbf{p}^t - \mathbf{p}^\star\|_0 \leq s, \quad \forall\, t \in [T]. \tag{1}$$

The central distribution $\mathbf{p}^\star$ can be viewed as the central structure across heterogeneous clusters of datapoints. The level of heterogeneity is controlled by the parameter s. When s is much smaller than d, the local distributions differ from the center in a small number of entries.

While motivated by word frequencies of different texts, our model of sparse heterogeneity is also relevant for recommendation systems, where the high-dimensional item-preference vectors of users can vary sparsely [28, 42, 35, 7]; and medical risk scoring, where hospitals enjoy similar predicted results but with a few systematic differences in diagnosis behavior, healthcare utilization, etc [50, 43, 40].

3 Algorithm

We now introduce our method for leveraging heterogeneous data to improve per-cluster sample efficiency. We first discuss a hashing-based method to handle the communication constraint. Since communication between each user datapoint and the server is restricted to at most $b > 0$ bits (where we may have $2^b \ll d$), the datapoint $X^{t,j}$ needs to be encoded by an encoding function $W^{t,j} : \mathcal{X} \triangleq [d] \to \mathcal{Y}$. The b-bit constraint enforces $|\mathcal{Y}| \leq 2^b$. Then the encoded message $Y^{t,j} := W^{t,j}(X^{t,j})$ is sent to the server, where it is decoded and used.

Under relatively sophisticated protocols, the design of encoding functions can be *interactive* [15, 14, 1], *i.e.*, depend on previously sent messages. Here we adopt a *non-interactive* encoding-decoding scheme, based on uniform hashing [3, 14] where $W^{t,j}$ depends only on $X^{t,j}$ and is independent of other messages. Specifically, each datapoint $X^{t,j}$ is encoded via an independent random hash function $h^{t,j} : [d] \to [2^b]$. Upon receiving all messages, the server counts the empirical frequencies of all symbols, leading to hashed estimates $\check{\mathbf{b}}^t$. The communication scheme based on uniform hashing

Algorithm 1 SHIFT: Sparse Heterogeneity Inspired collaboration and Fine-Tuning

```
input: individual hashed estimators $\check{\mathbf{b}}^1, \dots, \check{\mathbf{b}}^T$, threshold parameter $\alpha$
▷ Stage I: Collaborative Learning
Estimate $\mathbf{b}^\star$ via robust statistical methods: $\check{\mathbf{b}}^\star \leftarrow \mathrm{robust\_estimate}(\{\check{\mathbf{b}}^t : t \in [T]\})$
▷ Stage II: Fine-Tuning
for $k = 1, \dots, d$ do
    for $t = 1, \dots, T$ do
        $[\widehat{\mathbf{b}}^t]_k \leftarrow [\check{\mathbf{b}}^\star]_k$ if $|[\check{\mathbf{b}}^\star]_k - [\check{\mathbf{b}}^t]_k| \leq \sqrt{\alpha[\check{\mathbf{b}}^t]_k/n}$, else $[\check{\mathbf{b}}^t]_k$
        $[\widehat{\mathbf{p}}^t]_k \leftarrow \mathrm{Proj}_{[0,1]}(\frac{2^b[\widehat{\mathbf{b}}^t]_k - 1}{2^b - 1})$
    end for
end for
output: estimates $\widehat{\mathbf{p}}^1, \dots, \widehat{\mathbf{p}}^T$
```

is summarized below.

$$(\text{Encoding}): \text{Send the message } Y^{t,j} = h^{t,j}(X^{t,j}) \text{ encoded by a hash function } h^{t,j} : [d] \to [2^b];$$
$$(\text{Decoding}): \text{Count } N_k^t(Y^{t,[n]}) = |\{j \in [n] : h^{t,j}(k) = Y^{t,j}\}| \text{ and return } [\check{\mathbf{b}}^t]_k = N_k^t/n.$$

It is easy to verify that $\mathbb{E}[\check{\mathbf{b}}^t] = [(2^b - 1)\mathbf{p}^t + 1]/2^b \triangleq \mathbf{b}^t$; and thus the hashed estimate $\check{\mathbf{b}}^t$ is biased for $\mathbf{p}^t$. We also write $\mathbf{b}^\star = [(2^b - 1)\mathbf{p}^\star + 1]/2^b$ for the mean of a hashed datapoint sampled from the central distribution. More details on the hashed estimator $\check{\mathbf{b}}^t$ are given in Appendix A.

3.1 The SHIFT Method

We now introduce the SHIFT method, which consists of two stages: *collaborative learning* and *fine-tuning*. The first stage estimates the central hashed distribution $\mathbf{b}^\star$ using all hashed estimates $\{\check{\mathbf{b}}^t : t \in [T]\}$. This is achieved via tools from robust statistics such as the median or trimmed mean. The key insight here is that, since the heterogeneity is sparse, for each entry at which the individual distributions less mismatch with the central one, most datapoints (used to estimate that entry) are sampled from the probability of the central distribution. Hence, to estimate those entries of the central distribution, we can treat the datapoints generated by heterogeneous users as corrupted, and leverage robust statistical methods to mitigate their influences.

In the second stage—fine-tuning—we detect mismatched entries between individual hashed estimates $\check{\mathbf{b}}^t$ and the central estimate $\check{\mathbf{b}}^\star$. Recall that the central and individual distributions differ in only a few entries. For entries $k \in [d]$ such that $|[\check{\mathbf{b}}^\star]_k - [\check{\mathbf{b}}^t]_k|$ is below $(\alpha[\check{\mathbf{b}}^\star]_k/n)^{1/2}$ for some threshold parameter α, we may expect that $p_k^t = p_k^\star$. As a result, the estimate $[\check{b}^\star]_k$ of $b_k^\star$ can be more accurate as it is learned collaboratively. Thus, we assign $[\check{\mathbf{b}}^\star]_k$ as the final estimate $[\widehat{\mathbf{b}}^t]_k$ of b_k^t.

On the other hand, for the entries where the central and individual distributions differ, *i.e.*, $p_k^t \neq p_k^\star$, the threshold is more likely to be exceeded. In this case, we keep the individual estimate $[\check{\mathbf{b}}^t]_k$ as $[\widehat{\mathbf{b}}^t]_k$. Finally, since the hashed distributions $\mathbf{b}^t$ are biased, we debias them in the final estimates of $\mathbf{p}^t$ where $\mathrm{Proj}_{[0,1]}(\cdot)$ indicates truncating the input with upper/lower bound 1 and 0. Our method does not require sample splitting, despite using two stages, leading to increased sample-efficiency. We remark that the SHIFT does not require the knowledge of the sparse heterogeneity s.

Knowledge transfer to new clusters. The collaboratively learned central distribution from Algorithm 1 is adaptable to new clusters, with possibly few datapoints. This can be particularly beneficial for sample efficiency because most entries of the target distribution are well-estimated through collaborative learning. One can transfer those entries, and it suffices to estimate the few remaining entries, instead of the whole distribution. See Theorem 2 for the details. The knowledge transfer utility further motivates the importance of collaborative learning.

3.2 The Median-Based SHIFT

In this section, we provide statistical guarantees, in the form of upper bounds on the error, for the median-based SHIFT method, where $\mathrm{robust\_estimate}(\{\check{\mathbf{b}}^t : t \in [T]\})$ in Algorithm 1 is the entry-wise median. Specifically, we let

$$[\check{\mathbf{b}}^\star]_k = \mathrm{median}\big(\{[\check{\mathbf{b}}^t]_k : t \in [T]\}\big), \quad \text{for each } k \in [d].$$

When there is no ambiguity, we write $\check{\mathbf{b}}^\star = \mathrm{median}\big(\{\check{\mathbf{b}}^t\}_{t\in[T]}\big)$. We also provide results for the trimmed-mean-based SHIFT method, see Appendix D.

By setting the threshold parameter α in Algorithm 1 as $\alpha = \Theta(\ln(n))$, we prove the upper bounds on the final individual ℓ_2 estimation errors as follows. The results for the ℓ_1 error are in Appendix C.

Theorem 1. *Suppose $n \geq 2^{b+6}\ln(n)$ and $\alpha = \Theta(\ln(n))$*[2]*. Then, for the median-based SHIFT method, for any $t \in [T]$,*

$$\mathbb{E}\left[\|\widehat{\mathbf{p}}^t - \mathbf{p}^t\|_2^2\right] = \tilde{O}\left(\frac{\max\{2^b, s\}}{2^b n} + \frac{d}{2^b T n} + \frac{d}{n^2}\right).$$

When $n \geq 2^{b+6}\ln(n)$ and $n = \Omega(2^b \min\{T, d/\max\{2^b, s\}\})$, the rate further becomes

$$\mathbb{E}\left[\|\widehat{\mathbf{p}}^t - \mathbf{p}^t\|_2^2\right] = \tilde{O}\left(\frac{\max\{2^b, s\}}{2^b n} + \frac{d}{2^b T n}\right). \tag{2}$$

The upper bound in (2) consists of two terms. The first term $\max\{2^b, s\}/(2^b n)$ is independent of the dimension d, and is a factor $d/\max\{2^b, s\}$ smaller compared to the rate $d/(2^b n)$ obtained by the minimax optimal method using only homogeneous datapoints [26]. Thus, it brings a significant benefit under sparse heterogeneity, *i.e.*, when $s \ll d$. Meanwhile, the second, dimension-dependent, term $d/(2^b T n)$ is T times smaller than $d/(2^b n)$, since it depends on the *total sample-size Tn* used collaboratively, despite heterogeneity. Therefore, our method shows a factor of $\min\{T, d/\max\{2^b, s\}\}$ improvement in sample efficiency, compared to previous work designed for homogeneous datapoints.

For completeness, we also consider a heuristic version of estimators of prior works [26, 15] in which all datapoints are pooled to learn a global distribution $T^{-1}\sum_{t\in[T]}\mathbf{p}^t$, which is then used by each cluster. While this uses all datapoints, it inevitably introduces a non-vanishing bias $\boldsymbol{\delta}^t = \mathbf{p}^t - T^{-1}\sum_{t'\in[T]}\mathbf{p}^{t'}$ in estimating individual distributions, and can behave poorly when the bias is large. See Table 1 for more details.

Finally, we discuss our results on knowledge transfer. The central estimator $\check{\mathbf{b}}^\star$ is adaptable to a new cluster $\mathcal{C}^{T+1}$ in the following way. We adjust the fine-tuning procedure in Algorithm 1 to $[\widehat{\mathbf{b}}^{T+1}]_k \leftarrow [\check{\mathbf{b}}^\star]_k$ if $|[\check{\mathbf{b}}^\star]_k - [\check{\mathbf{b}}^{T+1}]_k| \leq \sqrt{\alpha[\check{\mathbf{b}}^{T+1}]_k/\tilde{n}}$, and $[\widehat{\mathbf{b}}^{T+1}]_k \leftarrow [\check{\mathbf{b}}^{T+1}]_k$ otherwise, where $\tilde{n}$ is the size of $\mathcal{C}^{T+1}$. We then show the following result.

Theorem 2. *Let $\check{\mathbf{b}}^{T+1}$ be the hashed estimate of any new cluster $\mathcal{C}^{T+1}$ with $\tilde{n}$ datapoints such that $n \geq \tilde{n} \geq 2^{b+6}\ln(\tilde{n})$ and $\tilde{n} = \Omega(2^b \min\{T, d/\max\{2^b, s\}\})$. Let the threshold parameter be $\alpha = \Theta(\ln(\tilde{n}))$. Then, the median-based SHIFT method has error bounded by*

$$\mathbb{E}\left[\|\widehat{\mathbf{p}}^{T+1} - \mathbf{p}^{T+1}\|_2^2\right] = \tilde{O}\left(\frac{\max\{2^b, s\}}{2^b \tilde{n}} + \frac{d}{2^b T n}\right).$$

Similarly, one can see that adaptation to new clusters with the median-based SHIFT method achieves a factor of $\min\{Tn/\tilde{n}, d/\max\{2^b, s\}\}$ improvement in sample-efficiency compared to the case where the distribution of the new cluster is estimated from scratch (*i.e.*without any knowledge transfer).

3.2.1 Highlights of Theoretical Analysis

In this section, we introduce the key analytical ideas behind the proof of Theorems 1 and 2. Our analysis is novel compared to previous analyses for methods with homogeneous datapoints. The

[2] To be precise, we require $\alpha = O(\ln(n))$ and $\alpha \geq c\ln(n)$ for some absolute constant c. The analogous statement applies in Theorem 2.

final individual estimation errors relate to the error of estimating the central hashed distribution $\mathbf{b}^\star$. However, we only expect high accuracy at the center for entries at which the heterogeneity is mild. To quantify the influence of heterogeneity, for any $0 < \eta \leq 1$, we define the set of *η-well-aligned entries* as

$$\mathcal{I}_\eta := \{k \in [d] : |\mathcal{B}_k| < \eta T\}, \quad \text{where} \quad \mathcal{B}_k \triangleq \{t \in [T] : b_k^t \neq b_k^\star, \textit{ i.e.}, \; p_k^t \neq p_k^\star\}$$

is the set of clusters whose distribution differs from $\mathbf{p}^\star$ in the k-th entry. We aim to estimate the η-well-aligned entries accurately by using robust statistical methods.

Further, we argue that there are few poorly-aligned entries, and they affect the final per-cluster error only mildly. By the pigeonhole principle, the number of entries that are not η-well-aligned is upper bounded by $|\mathcal{I}_\eta^c| \triangleq |[d] \backslash \mathcal{I}_\eta| \leq \frac{sT}{\eta T} = \frac{s}{\eta}$. Therefore, given an estimator $\check{\mathbf{b}}^\star$ that is accurate for the η-well-aligned entries, the entries of $\mathbf{b}^\star$ can be estimated accurately except for at most s/η entries. The following technical lemma bounds the error for each entry $k \in \mathcal{I}_\eta$.

Lemma 1. *Suppose* $\check{\mathbf{b}}^\star = \mathrm{median}\big(\{\check{\mathbf{b}}^t\}_{t\in[T]}\big)$. *Then for any* $0 < \eta \leq 1/5$ *and* $k \in \mathcal{I}_\eta$, *it holds that*

$$\mathbb{E}[([\check{\mathbf{b}}^\star]_k - [\mathbf{b}^\star]_k)^2] = \tilde{O}\left(\frac{|\mathcal{B}_k|^2 b_k^\star(1-b_k^\star)}{T^2 n} + \frac{b_k^\star(1-b_k^\star)}{Tn} + \frac{1}{n^2}\right).$$

Lemma 1 provides an upper bound relating to the frequency $|\mathcal{B}_k|/T$ of misalignment (smaller than η), and a variance term $b_k^\star(1-b_k^\star)$. This result cannot be obtained by directly applying the standard Chernoff or Hoeffding bounds to random variables distributed in $[0,1]$ as in previous works [15] for two reasons: 1) the datapoints are heterogeneous, 2) the variance $b_k^\star(1-b_k^\star)$ here can be small, compared to general random variables in $[0,1]$, implying more concentration than follows from Hoeffding's inequality. To address these issues, we analyze the concentration of the empirical $(1/2 \pm |\mathcal{B}_k|/T)$-quantiles to mitigate the influence of heterogeneity, and we also use Bernstein's inequality, which is variance-dependent [54], to obtain bounds relying on both the sample size Tn and the variance $b_k^\star(1-b_k^\star)$.

Also, the constant $1/5$, controlling the heterogeneity, is not essential (we choose $1/5$ for clarity). It can be replaced with any number below $1/2$ so that estimating the central probability distribution becomes possible, as the information conveyed by homogeneous datapoints dominates.

Lemma 1 reveals that well-aligned entries of the central distribution are accurately estimated. Thus one can use the central estimate for the entries where the central distribution $\mathbf{p}^t$ aligns with the target distribution $\mathbf{p}^t$. The remaining entries, that are neither well-aligned nor satisfy $p_k^\star = p_k^t$, can be estimated by the individual estimator. We argue that a properly chosen threshold parameter α filters out the desired entries to be estimated individually with high probability, leading to Theorems 1 and 2.

While estimating $\mathbf{p}^\star$ is not our main goal, one can readily obtain from Lemma 1 the following bound on estimating $\mathbf{p}^\star$ by summing up the errors for all entries $k \in [d] = \mathcal{I}_\eta$ with $\eta = \max_{k\in[d]} |\mathcal{B}_k|$. Corollary 1 reveals that the central distribution can be accurately estimated if the mismatching of distributions happens uniformly across all entries, *i.e.*, each entry differs in $O(sT/d)$ of clusters.

Corollary 1. *Let* $\widehat{\mathbf{p}}^\star = \mathrm{Proj}_{[0,1]}(\frac{2^b\check{\mathbf{b}}^\star-1}{2^b-1})$ *be obtained by the debiasing operation from Algorithm 1. Suppose* $|\mathcal{B}_k| = O(sT/d)$ *for any* $k \in [d]$, *with* $\eta = \max_{k\in[d]} |\mathcal{B}_k|/T$, *the median-based SHIFT method enjoys*

$$\mathbb{E}[\|\widehat{\mathbf{p}}^\star - \mathbf{p}^\star\|_2^2] = \tilde{O}\left(\frac{s^2}{d2^b n} + \frac{d}{2^b Tn} + \frac{d}{n^2}\right).$$

4 Lower Bounds

To complement our upper bounds, we now provide minimax lower bounds for estimating distributions under heterogeneity. Since our setting contains T heterogeneous clusters of datapoints, our minimax error metric is slightly different from the one studied in [26, 6, 27, 2]. Using the ℓ_2 error as the loss, the lower bound metric is defined as

$$\inf_{\substack{(W^{t',[n]})_{t'\in[T]} \\ \widehat{\mathbf{p}}^t}} \; \sup_{\substack{\mathbf{p}^\star\in\mathcal{P}_d \\ \{\mathbf{p}^{t'}:t'\in[T]\}\subseteq\mathbb{B}_s(\mathbf{p}^\star)}} \; \mathbb{E}\left[\left\|\widehat{\mathbf{p}}^t - \mathbf{p}^t\right\|_2^2\right], \tag{3}$$

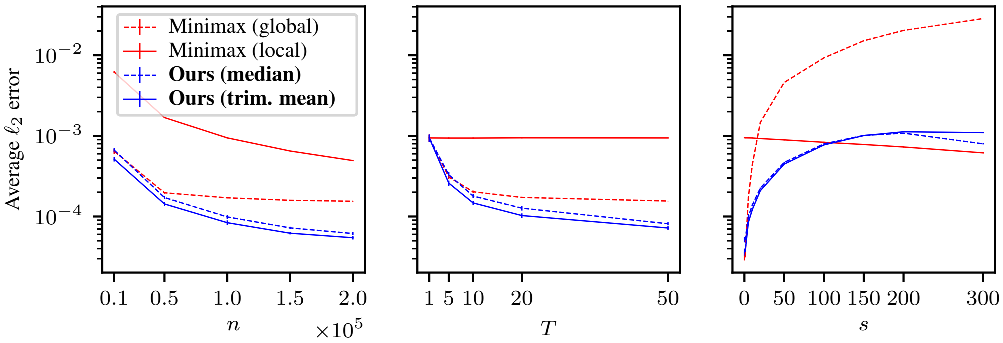

Figure 2: Average ℓ_2 estimation error in synthetic experiment. (Left): Fixing $s = 5$, $T = 30$ and varying n. (Middle): Fixing $s = 5$, $n = 100,000$ and varying T. (Right): Fixing $T = 30$, $n = 100,000$ and varying s. The standard error bars are obtained from 10 independent runs.

where the supremum is taken over all possible central distributions $\mathbf{p}^\star \in \mathcal{P}_d$ and individual distributions $\{\mathbf{p}^t : t \in [T]\}$ in $\mathbb{B}_s(\mathbf{p}^\star) \triangleq \{\mathbf{p} \in \mathcal{P}_d : \|\mathbf{p} - \mathbf{p}^\star\|_0 \leq s\}$, and the infimum is taken over all estimation methods $\hat{\mathbf{p}}^t$ that use all heterogeneous messages $\{Y^{t',j} \triangleq W^{t',j}(X^{t',j}) : j \in [n], t' \in [T]\}$ encoded (possibly interactively) by any encoding functions $\{W^{t',j} : j \in [n], t' \in [T]\}$ with output in $[2^b]$, *e.g.*, the random hashing maps. The measure (3) characterizes the best possible worst-case performance of estimating distributions under our model of heterogeneity.

Since the supremum is taken over all distributions $\mathbf{p}^\star, \mathbf{p}^1, \ldots, \mathbf{p}^T$ in $\mathcal{P}_d$ such that $\|\mathbf{p}^t - \mathbf{p}^\star\|_0 \leq s$ for all $t \in [T]$, we consider two representative cases therein: 1) The *homogeneous* case where $\mathbf{p}^1 = \cdots = \mathbf{p}^T = \mathbf{p}^\star \in \mathcal{P}_d$. Then the setting essentially reduces to the single-cluster problem but with nT datapoints, and the goal here is to estimate $\mathbf{p}^\star$, leading to the lower bound $\Omega(d/(2^b Tn))$. 2) The $s/2$-*sparse* case where $\|\mathbf{p}^\star\|_0 \leq s/2$ and $\|\mathbf{p}^t\|_0 \leq s/2$ for all $t \in [T]$. Then it naturally holds that $\{\mathbf{p}^t : t \in [T]\} \subseteq \mathbb{B}_s(\mathbf{p}^\star)$. By constructing independent priors for $\{\mathbf{p}^t : t \in [T]\}$ and $\mathbf{p}^\star$, one can show that only datapoints generated by $\mathbf{p}^t$ itself are informative for estimating $\mathbf{p}^t$. In this case, we show the lower bound $\Omega(\max\{2^b, s\}/(2^b n))$. Combining the two cases, we find the following lower bound. The formal argument is provided in Appendix E.

Theorem 3. *For any—possibly interactive—estimation method, and for any $t \in [T]$, we have*

$$\inf_{\substack{(W^{t',[n]})_{t' \in [T]} \\ \hat{\mathbf{p}}^t}} \sup_{\substack{\mathbf{p}^\star \in \mathcal{P}_d \\ \{\mathbf{p}^{t'} : t' \in [T]\} \subseteq \mathbb{B}_s(\mathbf{p}^\star)}} \mathbb{E}[\|\widehat{\mathbf{p}}^t - \mathbf{p}^t\|_2^2] = \Omega\left(\frac{\max\{2^b, s\}}{2^b n} + \frac{d}{2^b T n}\right). \tag{4}$$

By a similar argument but with an additional $(T+1)$-st cluster of $\tilde{n}$ users, we obtain a lower bound for adapting to a new cluster.

Theorem 4. *For any—possibly interactive—estimation method, and a new cluster $\mathcal{C}^{T+1}$, we have*

$$\inf_{\substack{(W^{t',[n]})_{t' \in [T]} \\ W^{T+1,[\tilde{n}]}, \widehat{\mathbf{p}}^{T+1}}} \sup_{\substack{\mathbf{p}^\star \in \mathcal{P}_d \\ \{\mathbf{p}^{t'} : t' \in [T+1]\} \subseteq \mathbb{B}_s(\mathbf{p}^\star)}} \mathbb{E}[\|\widehat{\mathbf{p}}^{T+1} - \mathbf{p}^{T+1}\|_2^2] = \Omega\left(\frac{\max\{2^b, s\}}{2^b \tilde{n}} + \frac{d}{2^b T n}\right). \tag{5}$$

Theorem 3 and 4, combined with the upper bounds in Section 3, imply that our method is minimax optimal up to logarithmic terms. We provide similar lower bounds for the ℓ_1 error in Appendix E.

5 Experiments

We test SHIFT on synthetic data as well as the Shakespeare dataset [11]. As a baseline method, we use the estimator based on uniform grouping in [26] that is minimax optimal under homogeneity, *i.e.*, in the single-task regime. We apply the baseline method both locally and globally. In the local case, the estimator $\widehat{\mathbf{p}}^t$ for each cluster is computed without datapoints from other clusters. In the global case, we pool data from all clusters, and compute estimators $\widehat{\mathbf{p}} = \widehat{\mathbf{p}}^1 = \cdots = \widehat{\mathbf{p}}^T$. The performance measure for estimating $\widehat{\mathbf{p}}^t$, $t \in [T]$ is taken as $T^{-1} \sum_{t=1}^T \|\mathbf{p}^t - \widehat{\mathbf{p}}^t\|_2^2$.

$k = 2$	$b = 2$	$b = 4$	$b = 6$	$b = 8$
Unif. Group. (local)	640 ± 6.0	142 ± 1.2	40 ± 0.40	14 ± 0.13
Unif. Group. (global)	33 ± 1.8	17 ± 0.37	14 ± 0.081	13 ± 0.037
SHIFT (median)	47 ± 2.4	21 ± 0.66	14 ± 0.17	11 ± 0.10
SHIFT (trimmed mean)	36 ± 2.2	19 ± 0.51	13 ± 0.24	10 ± 0.062
$k = 3$	$b = 2$	$b = 4$	$b = 6$	$b = 8$
Unif. Group. (local)	15000 ± 21	3000 ± 5.9	720 ± 2.1	180 ± 0.39
Unif. Group. (global)	4400 ± 5.7	100 ± 1.4	38 ± 0.35	23 ± 0.090
SHIFT (median)	7300 ± 9.6	180 ± 2.1	53 ± 1.0	20 ± 0.18
SHIFT (trimmed mean)	5100 ± 6.3	140 ± 2.3	43 ± 0.66	18 ± 0.18

Table 2: Average ℓ_2 error for estimating distributions of k-grams in the Shakespeare dataset. Numbers are scaled by 10^{-5}.

5.1 Synthetic Data

We set the uniform distribution, $\mathbf{p}^\star = (1/d, \dots, 1/d)$ as the central distribution. In Appendix F, we also experiment on the truncated geometric distribution and compare our method with the localization-refinement method [15]. Among the d entries of $\mathbf{p}^\star$, we draw s entries uniformly at random and assign new values for them uniformly at random over $[0, 1]$, with re-normalization to preserve their sum. We repeat this procedure T times to obtain sparsely perturbed distributions $\mathbf{p}^1, \dots, \mathbf{p}^T \in \mathcal{P}_d$. Then, n i.i.d. datapoints $X^{t,1}, \dots, X^{t,n} \sim \mathrm{Cat}(\mathbf{p}^t)$ are generated for each cluster $t \in [T]$. We set the dimension to $d = 300$ and run the simulation by varying n, T, s. As we see from (2), the error of our method depends on s only when $2^b < s$. For this reason, we let $b = 2$ in our experiments.

We run SHIFT with the entry-wise median and entry-wise trimmed mean as the robust estimate. We set the threshold parameter $\alpha = \ln(n)$ and the trimming proportion $\omega = 0.1$. In Appendix F, we provide results for other choices of the hyperparameters α, ω and discuss a heuristic for choosing α. Figure 2 illustrates that our method outperforms the baselines for most choices of n, T, s. Specifically, as Theorem 1 predicts, the ℓ_2 error of our method decreases as the number of clusters T increases. On the other hand, when the baseline methods are applied globally, without considering heterogeneity, they show a bias that does not disappear as the sample size n or the number of clusters T increases. This shows that the fine-tuning step in SHIFT is crucial for the estimation of heterogeneous distributions. Finally, the right panel of Figure 2 shows that our method is effective only when s is small compared to the dimension d; which highlights the crucial role of sparse heterogeneity. When s is close to d, the distributions $\mathbf{p}^1, \dots \mathbf{p}^T$ could be considerably different without any meaningful central structure, making collaboration less useful than local estimation.

5.2 Shakespeare Dataset

The Shakespeare dataset was proposed as a benchmark for federated learning in [11]. The dataset consists of dialogues of 1,129 speaking roles in Shakespeare's 35 different plays. In our experiment, we study the distribution of k-grams (k-tuples of consecutive letters from the 26-letter English alphabet, see Chapter 3 of [33]) appearing in the dialogues. We consider each play as a cluster $\mathcal{C}^t$ and estimate the distribution $\mathbf{p}^t \in \mathcal{P}_d$, $d = 26^k$ of k-grams. Since the ground-truth distribution $\mathbf{p}^t$ is unknown, we regard the empirical frequency as $\mathbf{p}^t$.

To verify the heterogeneity, we run the chi-squared goodness-of-fit test for each pair of distributions from distinct clusters $\mathbf{p}^u$ and $\mathbf{p}^v$. Resulting p-values were essentially zero within machine precision, which suggests that the distributions of k-grams are strongly heterogeneous. We also perform entry-wise tests comparing $[\mathbf{p}^u]_i$ and $[\mathbf{p}^v]_i$ for all $u \neq v \in [T], i \in [d]$. In total, 25.8% of the tests were rejected at the significance level of 5%. This again supports the heterogeneity.

We draw $n = 20,000$ datapoints with replacement from each cluster and test SHIFT with communication budgets $b \in \{2, 4, 6, 8\}$. We set the fine-tuning threshold $\alpha = \ln(n)$ and the trimming proportion $\omega = 0.1$, which we choose following the heuristic discussed in Appendix F. We repeat the experiment ten times by taking different datapoints and report the average ℓ_2 error of estimation in Table 2. The standard deviations are small even over ten repetitions. SHIFT shows competitive performance on the empirical dataset, even though we do not rigorously know if the sparse heterogeneity model (1) applies.

6 Conclusion and Future Directions

We formulate the problem of learning distributions under sparse heterogeneity and communication constraints. We propose the SHIFT method, which first learns a central distribution, and then fine-tunes the estimate to adapt to individual distributions. We provide both theoretical and experimental results to show its sample-efficiency improvement compared to classical methods that target only homogeneous data. Many interesting directions remain to be explored, including investigating if there is a point-wise optimal method with rate depending on $\{\mathbf{p}^t : t \in [T]\}$ and $\mathbf{p}^\star$; and designing methods for other information constraints, such as local differential privacy constraints.

Acknowledgements

During this work, Xinmeng Huang was supported in part by the NSF TRIPODS 1934960, NSF DMS 2046874 (CAREER), NSF CAREER award CIF-1943064; Donghwan Lee was supported in part by ARO W911NF-20-1-0080, DCIST, Air Force Office of Scientific Research Young Investigator Program (AFOSR-YIP) award #FA9550-20-1-0111.

References

[1] J. Acharya, C. L. Canonne, and H. Tyagi. General lower bounds for interactive high-dimensional estimation under information constraints. *arXiv: Data Structures and Algorithms*, 2020.

[2] J. Acharya, C. L. Canonne, and H. Tyagi. Inference under information constraints ii: Communication constraints and shared randomness. *IEEE Transactions on Information Theory*, 66:7856–7877, 2020.

[3] J. Acharya, P. Kairouz, Y. Liu, and Z. Sun. Estimating sparse discrete distributions under privacy and communication constraints. In *International Conference on Algorithmic Learning Theory*, 2021.

[4] I. Achituve, A. Shamsian, A. Navon, G. Chechik, and E. Fetaya. Personalized federated learning with gaussian processes. *ArXiv*, abs/2106.15482, 2021.

[5] M. F. Balcan, A. Blum, S. Fine, and Y. Mansour. Distributed learning, communication complexity and privacy. In *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23, pages 26.1–26.22, 2012.

[6] L. P. Barnes, Y. Han, and A. Ozgur. Lower bounds for learning distributions under communication constraints via fisher information. *arXiv: Information Theory*, 2019.

[7] H. Bastani. Predicting with proxies: Transfer learning in high dimension. *Management Science*, 67:2964–2984, 2021.

[8] J. Baxter. A model of inductive bias learning. *Journal of Artificial Intelligence Research*, 12:149–198, 2000.

[9] P. Blanchard, E. M. E. Mhamdi, R. Guerraoui, and J. Stainer. Byzantine-tolerant machine learning. *ArXiv*, abs/1703.02757, 2017.

[10] H. D. Brunk, R. E. Barlow, D. J. Bartholomew, and J. M. Bremner. Statistical inference under order restrictions : the theory and application of isotonic regression. *International Statistical Review*, 41:395, 1972.

[11] S. Caldas, S. M. K. Duddu, P. Wu, T. Li, J. Konečnỳ, H. B. McMahan, V. Smith, and A. Talwalkar. Leaf: A benchmark for federated settings. *arXiv preprint arXiv:1812.01097*, 2018.

[12] R. Caruana. Multitask learning. *Machine Learning*, 28:41–75, 1997.

[13] W.-N. Chen, P. Kairouz, and A. Ozgur. Breaking the communication-privacy-accuracy trilemma. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 3312–3324. Curran Associates, Inc., 2020.

[14] W.-N. Chen, P. Kairouz, and A. Özgür. Breaking the dimension dependence in sparse distribution estimation under communication constraints. *arXiv preprint arXiv:2106.08597*, 2021.

[15] W.-N. Chen, P. Kairouz, and A. Özgür. Pointwise bounds for distribution estimation under communication constraints. *ArXiv*, abs/2110.03189, 2021.

[16] L. Collins, H. Hassani, A. Mokhtari, and S. Shakkottai. Exploiting shared representations for personalized federated learning. *ArXiv*, abs/2102.07078, 2021.

[17] H. Daumé, J. M. Phillips, A. Saha, and S. Venkatasubramanian. Efficient protocols for distributed classification and optimization. *ArXiv*, abs/1204.3523, 2012.

[18] L. Devroye and L. Györfi. Nonparametric density estimation : the l1 view. *Journal of the American Statistical Association*, 82:344, 1987.

[19] L. Devroye and G. Lugosi. Combinatorial methods in density estimation. In *Springer series in statistics*, 2001.

[20] E. Dobriban and Y. Sheng. Distributed linear regression by averaging. *The Annals of Statistics*, 2018.

[21] S. S. Du, W. Hu, S. M. Kakade, J. Lee, and Q. Lei. Few-shot learning via learning the representation, provably. *ArXiv*, abs/2002.09434, 2021.

[22] A. Garg, T. Ma, and H. L. Nguyen. On communication cost of distributed statistical estimation and dimensionality. In *Advances in Neural Information Systems*, 2014.

[23] P. Goyal, D. K. Mahajan, A. K. Gupta, and I. Misra. Scaling and benchmarking self-supervised visual representation learning. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6390–6399, 2019.

[24] F. Grimberg, M.-A. Hartley, S. P. Karimireddy, and M. Jaggi. Optimal model averaging: Towards personalized collaborative learning. *ArXiv*, abs/2110.12946, 2021.

[25] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. *Robust statistics: the approach based on influence functions*, volume 196. John Wiley & Sons, 2011.

[26] Y. Han, P. Mukherjee, A. Özgür, and T. Weissman. Distributed statistical estimation of high-dimensional and nonparametric distributions. *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 506–510, 2018.

[27] Y. Han, A. Özgür, and T. Weissman. Geometric lower bounds for distributed parameter estimation under communication constraints. *IEEE Transactions on Information Theory*, 67:8248–8263, 2021.

[28] L. Hu, S. Jian, L. Cao, Z. Gu, Q. Chen, and A. Amirbekyan. Hers: Modeling influential contexts with heterogeneous relations for sparse and cold-start recommendation. In *The Thirty-Third AAAI Conference on Artificial Intelligence*, 2019.

[29] X. Huang, Y. Chen, W. Yin, and K. Yuan. Lower bounds and nearly optimal algorithms in distributed learning with communication compression. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[30] P. J. Huber. Robust statistics. *Wiley Series in Probability and Mathematical Statistics*, 1981.

[31] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. B. Charles, G. Cormode, R. Cummings, R. G. L. D'Oliveira, S. Y. E. Rouayheb, D. Evans, J. Gardner, Z. Garrett, A. Gascón, B. Ghazi, P. B. Gibbons, M. Gruteser, Z. Harchaoui, C. He, L. He, Z. Huo, B. Hutchinson, J. Hsu, M. Jaggi, T. Javidi, G. Joshi, M. Khodak, J. Konecný, A. Korolova, F. Koushanfar, O. Koyejo, T. Lepoint, Y. Liu, P. Mittal, M. Mohri, R. Nock, A. Özgür, R. Pagh, M. Raykova, H. Qi, D. Ramage, R. Raskar, D. X. Song, W. Song, S. U. Stich, Z. Sun, A. T. Suresh, F. Tramèr, P. Vepakomma, J. Wang, L. Xiong, Z. Xu, Q. Yang, F. X. Yu, H. Yu, and S. Zhao. Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 2021.

[32] M. Lerasle and R. I. Oliveira. Robust empirical mean estimators. *arXiv: Statistics Theory*, 2011.

[33] J. Leskovec, A. Rajaraman, and J. D. Ullman. *Mining of massive data sets*. Cambridge university press, 2020.

[34] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang. On the convergence of fedavg on non-iid data. *ArXiv*, abs/1907.02189, 2020.

[35] T. Liu, Z. Wang, J. Tang, S. Yang, G. Y. Huang, and Z. Liu. Recommender systems with heterogeneous side information. *The World Wide Web Conference*, 2019.

[36] A. Maurer, M. Pontil, and B. Romera-Paredes. The benefit of multitask representation learning. *Journal of Machine Learning Research*, 17:81:1–81:32, 2016.

[37] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, 2017.

[38] S. Minsker. Geometric median and robust estimation in banach spaces. *Bernoulli*, 2015.

[39] S. Minsker and N. Strawn. Distributed statistical estimation and rates of convergence in normal approximation. *ArXiv*, 2019.

[40] S. Mullainathan and Z. Obermeyer. Does machine learning automate moral hazard and error? *The American Economic Review*, 107(5):476–480, 2017.

[41] R. D. Nowak. Distributed em algorithms for density estimation in sensor networks. *International Conference on Acoustics, Speech, and Signal Processing*, 2003.

[42] M. Qian, L. Hong, Y. Shi, and S. Rajan. Structured sparse regression for recommender systems. *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, 2015.

[43] J. Quionero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence. *Dataset Shift in Machine Learning*. MIT Press, 2009.

[44] A. Shamsian, A. Navon, E. Fetaya, and G. Chechik. Personalized federated learning using hypernetworks. In *International Conference on Machine Learning*, 2021.

[45] B. W. Silverman. *Density estimation for statistics and data analysis*. London: Chapman & Hall, 1986.

[46] V. Slavov and P. R. Rao. A gossip-based approach for internet-scale cardinality estimation of xpath queries over distributed semistructured data. *The International Journal on Very Large Data Bases*, 2013.

[47] V. Smith, C.-K. Chiang, M. Sanjabi, and A. S. Talwalkar. Federated multi-task learning. In *Advances in neural information processing systems*, 2017.

[48] L. Su and N. H. Vaidya. Fault-tolerant multi-agent optimization: Optimal iterative distributed algorithms. In *Proceedings of the 2016 ACM Symposium on Principles of Distributed Computing*, 2016.

[49] L. Su and N. H. Vaidya. Non-bayesian learning in the presence of byzantine agents. In *International Symposium on Distributed Computing*, pages 414–427, 2016.

[50] A. Subbaswamy and S. Saria. From development to deployment: dataset shift, causality, and shift-stable models in health ai. *Biostatistics*, 2019.

[51] C. Sun, A. Shrivastava, S. Singh, and A. K. Gupta. Revisiting unreasonable effectiveness of data in deep learning era. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 843–852, 2017.

[52] S. Thrun and L. Y. Pratt. Learning to learn: Introduction and overview. In *Learning to Learn*. Springer, 1998.

[53] N. Tripuraneni, C. Jin, and M. I. Jordan. Provable meta-learning of linear representations. In *International Conference on Machine Learning*, 2021.

[54] J. V. Uspensky. *Introduction to mathematical probability*. McGraw-Hill Book Company, 1937.

[55] R. Vershynin. *High-Dimensional Probability*. 2018.

[56] D. Wang, S. Shi, Y. Zhu, and Z. Han. Federated analytics: Opportunities and challenges. *IEEE Network*, 2022.

[57] K. Xu and H. Bastani. Learning across bandits in high dimension via robust statistics. *ArXiv*, abs/2112.14233, 2021.

[58] D. Yin, Y. Chen, K. Ramchandran, and P. L. Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. *ArXiv*, abs/1803.01498, 2018.

[59] K. Yuan, S. A. Alghunaim, and X. Huang. Removing data heterogeneity influence enhances network topology dependence of decentralized sgd. *arXiv preprint arXiv:2105.08023*, 2021.

[60] K. Yuan, Y. Chen, X. Huang, Y. Zhang, P. Pan, Y. Xu, and W. Yin. Decentlam: Decentralized momentum sgd for large-batch deep training. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[61] M. Zhou, H. T. Shen, X. Zhou, W. Qian, and A. Zhou. Effective data density estimation in ring-based p2p networks. *International Conference on Data Engineering*, 2012.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [N/A]
   (c) Did you discuss any potential negative societal impacts of your work? [No]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (*e.g.*, data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (*e.g.*, with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (*e.g.*, type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (*e.g.*, code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]