Optimal Distributed Nonparametric Regression with a Single Index: Sufficient Dimension Reduction and Cross-Fitted Adaptive Kernel Smoothing

Abstract

Distributed estimation of single-index models is a fundamental challenge when massive datasets are partitioned across many machines. We propose a communication-efficient, non-iterative framework that integrates sufficient dimension reduction with adaptive kernel ridge regression. The index vector is estimated by spectrally aggregating local direction estimators on the Grassmann manifold, which naturally resolves sign ambiguity and avoids alignment preprocessing. A leave-one-out cross-fitting design decouples the parametric and nonparametric stages, guaranteeing that the projected direction is statistically independent of the local regression data. For the link function, each machine computes a local kernel ridge estimator and transmits only a small set of non-sensitive basis coefficients; a central server synthesizes a global approximation used as the validation target for adaptive tuning. This procedure requires only two communication rounds of basis coefficients and prevents the overfitting inherent in purely local validation. The theoretical analysis accommodates general heterogeneous partitions without balanced sample allocation, establishes $\sqrt{N}$-consistency and asymptotic normality for the direction estimator, and derives the minimax optimal rate $N^{-\alpha/(2\alpha+1)}$ for the link function. Extensive simulations and an airline on-time performance study confirm the strong predictive accuracy, robustness, and computational speed-up in unbalanced distributed environments.

Keywords: Distributed single-index model, Sufficient dimension reduction, Grassmann manifold, Adaptive kernel ridge regression, Cross-fitting, Communication efficiency

Mathematics Subject Classification (2020): 62G08, 62R07

1 Introduction

The contemporary era of big data has witnessed an unprecedented explosion in data volume across scientific domains, and massive datasets are frequently distributed across multiple machines due to storage constraints, privacy considerations, or computational limitations. Let $\{(\boldsymbol{X}_i, Y_i)\}_{i=1}^N$ be independent and identically distributed (i.i.d.) copies of $(\boldsymbol{X}, Y)$, where $\boldsymbol{X} \in \mathbb{R}^p$ and $Y \in \mathbb{R}$; in a typical distributed setting, the data are partitioned across m local machines, with the k-th machine holding a subset S_k of size n_k, where $\sum_{k=1}^m n_k = N$. Within this framework, the single-index model (SIM), a prototypical and widely adopted class of nonlinear regression, serves as a fundamental tool for capturing complex relationships beyond linearity in

nonparametric and semiparametric regression. The SIM is specified as

$$Y = f(b^\top \boldsymbol{X}) + \varepsilon, \tag{1}$$

where $Y \in \mathbb{R}$ is the response variable, $\boldsymbol{X} \in \mathbb{R}^p$ denotes a p-dimensional predictor vector, $b \in \mathbb{R}^p$ with $\|b\| = 1$ is the index vector identifying the effective direction, $f(\cdot)$ is an unknown link function, and ε is a random error independent of $\boldsymbol{X}$ with $\mathbb{E}[\varepsilon] = 0$ and $\mathrm{Var}(\varepsilon) = \sigma^2$. The SIM significantly alleviates the curse of dimensionality inherent in fully nonparametric regression while retaining flexibility in modeling nonlinear relationships; consequently, developing distributed estimation techniques for such nonlinear models is critical for extracting nonlinear insights from massive datasets without compromising privacy or overwhelming single-machine memory.

Despite its importance, extending nonlinear regression techniques to distributed environments presents unique challenges. Recently, to handle SIM estimation with massive data, scholars have proposed various distributed strategies. For example, Jiang et al. (2022) developed a divide-and-conquer (DC) approach based on composite quasi-likelihood, while Li and Yang (2025) combined refined outer product gradients (rOPG) to propose the DC-lsrOPG and DC-qrOPG algorithms. These methods broadly fall into two categories: one-shot averaging approaches that suffer efficiency loss as the number of subsets grows, and iterative algorithms that incur repeated synchronization costs. The iterative divide-and-conquer (IDC) method proposed by Jiang and Peng (2023) removes the restriction on block numbers via multiple rounds of refinement; however, this comes at the expense of repeated communication between local and central machines.

Recent distributed strategies for single-index models have focused primarily on the parametric index vector. Jiang and Peng (2023) proposed an iterative divide-and-conquer method that refines the direction estimator through multiple rounds of aggregation, while Li and Yang (2025) developed one-shot DC-rOPG estimators via Euclidean averaging of local directions. Both lines of work establish asymptotic normality for the index coefficient; however, *the distributed convergence properties of the nonparametric link function remain entirely untreated.* In particular, neither provides a theoretical rate for estimating $f(\cdot)$ under data partitioning, nor addresses how the parametric estimation error propagates into the subsequent nonparametric smoothing stage. Moreover, both methods rely on the restrictive balanced-sample assumption $n_k = N/m$, leaving their validity under general heterogeneous partitions unexamined.

We close this gap by establishing *two-stage optimality.* Theorem 4.1 establishes parametric consistency and asymptotic normality for the aggregated direction estimator under general heterogeneous partitions; the convergence guarantees hold under both fixed and diverging numbers of machines (with $m = o(\sqrt{N})$ in the latter case, which is automatically satisfied under the stronger condition of Theorem 4.2). Theorem 4.2 further derives the minimax optimal rate for the link function estimator under adaptive parameter selection, accommodating realistic unbalanced workloads without requiring balanced sample allocation. Having obtained the direction via spectral aggregation on the Grassmann manifold, we employ adaptive distributed kernel ridge regression on the projected covariates $\widehat{b}^\top \boldsymbol{X}$. Theorem 4.2 derives the minimax optimal rate $N^{-\alpha/(2\alpha+1)}$ for the aggregated link function estimator under an adaptive parameter selection

strategy that requires no centralization of raw data; this rate depends solely on the univariate smoothness α rather than the ambient dimension p.

A critical obstacle in realizing this two-stage pipeline is the statistical dependence between the direction estimator $\widehat{b}$ and the local kernel ridge regression estimators, which are conventionally computed from the *same* data. This dependence, commonly referred to as "circularity" in semiparametric inference, complicates the theoretical analysis because the projection error contaminates the oracle nonparametric error through intricate conditional dependencies. In high-dimensional causal estimation, Chernozhukov et al. (2018) demonstrated that *cross-fitting*—constructing nuisance estimators on independent subsamples—systematically eliminates such dependence. Their double machine learning (DML) framework showed that this decoupling removes the need for uniform convergence conditions over nuisance realizations and yields asymptotically normal estimators at the $\sqrt{N}$ rate even when nuisance estimators converge slowly. Early antecedents of this idea trace back to sample-splitting techniques in classical semiparametrics (Schick, 1986; Bickel, 1982). We adapt this principle to the distributed setting by proposing a *leave-one-out cross-fitting* variant: the direction estimator sent to machine k is constructed exclusively from the remaining $m-1$ machines, making it statistically independent of the data on machine k. Consequently, the projection error no longer contaminates the oracle nonparametric error, and the resulting convergence rate is freed from the extraneous $\sqrt{\log N}$ factors that typically plague coupled two-stage analyses.

To circumvent the heavy communication burden of iterative methods and the structural failures of simple averaging, we propose a *communication-efficient distributed learning framework* based on sufficient dimension reduction (SDR) and adaptive kernel ridge regression. Rather than iterative gradient aggregation, we employ distributed SDR techniques that transmit low-dimensional projection directions from local machines to the central server. Specifically, SDR seeks a $p \times d$ matrix B such that

$$Y \perp \boldsymbol{X} \mid B^\top \boldsymbol{X},$$

see (Cook, 2009). Here, the symbol $\perp$ denotes statistical independence. For the single-index model defined in (1), we have $d = 1$ and the matrix B reduces to the index vector b.

A key advantage of many SDR estimators is that they admit closed-form solutions via eigenvalue decomposition or quadratic programming. This enables genuine one-shot aggregation where local estimates of the central mean subspace are combined at the central server through spectral methods on the Grassmann manifold, bypassing the sign ambiguity and alignment issues inherent in simple Euclidean averaging. Unlike the iterative approach of Jiang and Peng (2023)—which requires repeated synchronization with communication costs scaling linearly in the number of iterations—our framework proceeds through a small, fixed number of non-iterative communication rounds. Specifically, the SDR stage involves one round of transmitting local directions and one round of broadcasting leave-one-out aggregated directions; the adaptive distributed kernel ridge regression tuning stage involves one round of transmitting basis coefficients and one round of broadcasting globally synthesized coefficients. Each round transmits only low-dimensional, non-sensitive summary statistics (direction vectors or basis coefficients), and there is no iterative refinement or repeated optimization.

Our methodological contribution is twofold. First, we establish a unified distributed SDR

framework that accommodates three representative dimension reduction paradigms: moment-based, machine-learning-enhanced, and model-free. By transmitting p-dimensional index vectors from m local machines to construct the aggregate matrix $\widehat{\boldsymbol{B}} = (\widehat{b}_1, \ldots, \widehat{b}_m)$, followed by spectral aggregation on the Grassmann manifold, the proposed method yields $\sqrt{N}$-consistent estimates while requiring only $O(mp)$ communication cost. Second, we propose a decoupled two-stage estimation procedure that separates high-dimensional parametric estimation from univariate nonparametric function approximation. Having obtained the index vector estimate via any distributed SDR method, we employ adaptive distributed kernel ridge regression to estimate the link function $f(\cdot)$ using the projected covariates $\widehat{b}^\top \boldsymbol{X}$.

Crucially, the adaptive parameter selection mechanism follows the adaptive distributed kernel ridge regression (AdaDKRR) principle of Lin et al. (2025): each local machine computes an approximation of the global kernel ridge regression rule via a fixed basis expansion, transmits the basis coefficients (which are non-sensitive) to the central server, and receives back a globally synthesized approximation. Using this global approximation as the validation target, each machine then selects its regularization parameter by standard hold-out or cross-validation. This design avoids the over-fitting problem inherent in purely local tuning—where the locally optimal regularization parameter is typically larger than the globally optimal one—and guarantees that the distributed estimator performs similarly to the oracle estimator based on the full centralized sample.

The remainder of this article is organized as follows. Section 2 reviews three representative SDR methods within a unified framework tailored for the single-index model. Section 3 details the distributed estimation procedure for the index vector, the adaptive distributed kernel ridge regression methodology for link function estimation, and an analysis of the overall communication cost. Section 4 establishes the asymptotic properties of the proposed estimators. Section 5 presents extensive numerical studies. Section 6 concludes the paper. Proofs of the main theorems are deferred to the Appendix.

2 Preliminaries and Review of SDR Methods

The literature on sufficient dimension reduction has evolved into a rich methodological landscape, encompassing moment-based approaches (Li, 1991; Cook and Weisberg, 1991; Li and Wang, 2007; Zhu et al., 2010), likelihood-based methods, and nonparametric smoothing techniques (Xia et al., 2002; Xia, 2007). For the distributed estimation framework developed in this paper, we focus on three representative SDR paradigms that offer distinct theoretical foundations while sharing computational tractability: *sliced inverse regression* (SIR, Li 1991) as the canonical moment-based method, *principal support vector machines* (PSVM, Li et al. 2011) as a prominent example of machine-learning-enhanced dimension reduction, and *distance covariance* (DCOV, Sheng and Yin 2016) as a model-free dependence measure. These methods are selected not only for their complementary strengths—SIR for its simplicity and computational efficiency under linearity conditions, PSVM for its robustness via large-margin classification, and DCOV for its distribution-free property that circumvents restrictive design conditions—but also because they all admit closed-form or efficiently computable solutions via eigen-decomposition or convex optimization. This common algorithmic structure renders them particularly amenable to one-

shot distributed aggregation.

In this section, we review three prominent SDR methods within a unified framework tailored for the single-index model (1). Throughout, we assume $\{(\boldsymbol{X}_i, Y_i)\}_{i=1}^n$ is a random sample from $(\boldsymbol{X}, Y)$, where $\boldsymbol{X} \in \mathbb{R}^p$ and $Y \in \mathbb{R}$. The central subspace $\mathcal{S}_{Y|\boldsymbol{X}}$ is one-dimensional, spanned by the unit index vector $b \in \mathbb{R}^p$ satisfying $b^\top b = 1$, where $\mu = \mathbb{E}(\boldsymbol{X})$ and $\Sigma = \mathrm{Cov}(\boldsymbol{X})$.

All three methods proceed by standardizing the predictors. Let $\boldsymbol{Z} = \Sigma^{-1/2}(\boldsymbol{X} - \mu)$ denote the standardized predictor with $\mathbb{E}(\boldsymbol{Z}) = 0$ and $\mathrm{Cov}(\boldsymbol{Z}) = \boldsymbol{I}_p$. In practice, we compute sample moments $\widehat{\mu} = \bar{\boldsymbol{X}}$ and $\widehat{\Sigma} = n^{-1}\sum_{i=1}^n (\boldsymbol{X}_i - \widehat{\mu})(\boldsymbol{X}_i - \widehat{\mu})^\top$, yielding standardized data $\widehat{\boldsymbol{Z}}_i = \widehat{\Sigma}^{-1/2}(\boldsymbol{X}_i - \widehat{\mu})$. Estimation is performed in the Z-space to obtain a direction vector η, which is then transformed back to the original scale via $b = \Sigma^{-1/2}\eta$. Thus, the final estimator takes the form $\widehat{b} = \widehat{\Sigma}^{-1/2}\widehat{\eta}$, normalized to unit length.

2.1 Sliced Inverse Regression

SIR (Li, 1991) exploits the inverse regression curve $\mathbb{E}(\boldsymbol{X} \mid Y)$. Partition the range of Y into H slices $\{I_h\}_{h=1}^H$ with cut-off points $-\infty = q_0 < q_1 < \cdots < q_H = \infty$, where $I_h = (q_{h-1}, q_h]$. Let $p_h = \Pr(Y \in I_h)$ and $\delta_h(Y) = \mathbf{1}(Y \in I_h)$.

Within the standardized space, define the slice-specific conditional mean:

$$\zeta_h = \mathbb{E}(\boldsymbol{Z} \mid Y \in I_h).$$

Under the linearity condition (Li, 1991), the centered inverse regression curve $\mathbb{E}(\boldsymbol{Z} \mid Y)$ lies within the standardized central subspace. The *candidate matrix* is formed as the weighted covariance of these slice means:

$$\boldsymbol{M}_0^{\mathrm{SIR}} = \sum_{h=1}^H p_h \zeta_h \zeta_h^\top.$$

Let η be the leading eigenvector of $\boldsymbol{M}_0^{\mathrm{SIR}}$. The population index vector is $b = \Sigma^{-1/2}\eta$.

Empirically, we estimate slice probabilities $\widehat{p}_h = n^{-1}\sum_{i=1}^n \delta_h(Y_i)$ and slice means $\widehat{\zeta}_h = (n\widehat{p}_h)^{-1}\sum_{i=1}^n \widehat{\boldsymbol{Z}}_i \delta_h(Y_i)$. The sample candidate matrix is $\widehat{\boldsymbol{M}}_n^{\mathrm{SIR}} = \sum_{h=1}^H \widehat{p}_h \widehat{\zeta}_h \widehat{\zeta}_h^\top$. The SIR estimator $\widehat{b}_{\mathrm{SIR}}$ is obtained as the leading eigenvector of $\widehat{\Sigma}^{-1/2}\widehat{\boldsymbol{M}}_n^{\mathrm{SIR}}\widehat{\Sigma}^{-1/2}$, normalized to satisfy $\widehat{b}_{\mathrm{SIR}}^\top \widehat{b}_{\mathrm{SIR}} = 1$.

2.2 Principal Support Vector Machines

PSVM (Li et al., 2011) extends support vector machines to SDR by dichotomizing Y. Using cut-off points $\{q_\ell\}_{\ell=1}^{R-1}$, define pseudo-responses:

$$\widetilde{Y}^{(\ell)} = \mathbf{1}(Y > q_\ell) - \mathbf{1}(Y \le q_\ell), \quad \ell = 1, \ldots, R-1.$$

For each ℓ, consider the population SVM-type objective:

$$L(\psi_\ell, t_\ell) = \psi_\ell^\top \psi_\ell + \lambda^* \mathbb{E}\{1 - \widetilde{Y}^{(\ell)}(\psi_\ell^\top \boldsymbol{Z} - t_\ell)\}^+,$$

where $\lambda^* > 0$ is a regularization parameter and $[\cdot]^+$ denotes the hinge loss. Let $(\psi_{0,\ell}, t_{0,\ell})$ minimize L. Under the linearity condition (Li, 1991), $\psi_{0,\ell}$ resides in the standardized central

subspace.

Aggregating across all $R-1$ binary partitions yields the candidate matrix:

$$\boldsymbol{M}_0^{\mathrm{PSVM}} = \sum_{\ell=1}^{R-1} \psi_{0,\ell}\psi_{0,\ell}^\top.$$

Let η be the leading eigenvector of $\boldsymbol{M}_0^{\mathrm{PSVM}}$. The population direction satisfies $b = \Sigma^{-1/2}\eta$.

The sample estimate $\widehat{\psi}_{n,\ell}$ minimizes the empirical objective $\widehat{L}(\psi_\ell, t_\ell) = \psi_\ell^\top\psi_\ell + \lambda\mathbb{E}_n\{1 - \widetilde{Y}_i^{(\ell)}(\psi_\ell^\top\widehat{\boldsymbol{Z}}_i - t_\ell)\}^+$. Forming $\widehat{\boldsymbol{M}}_n^{\mathrm{PSVM}} = \sum_{\ell=1}^{R-1}\widehat{\psi}_{n,\ell}\widehat{\psi}_{n,\ell}^\top$, the PSVM estimator $\widehat{b}_{\mathrm{PSVM}}$ is the leading eigenvector of $\widehat{\Sigma}^{-1/2}\widehat{\boldsymbol{M}}_n^{\mathrm{PSVM}}\widehat{\Sigma}^{-1/2}$, normalized to unit length.

2.3 Distance Covariance Based SDR

DCOV (Székely et al., 2007; Sheng and Yin, 2013) provides a model-free dependence measure that circumvents linearity conditions. In the standardized space, we seek the unit vector η maximizing the distance covariance between $\eta^\top\boldsymbol{Z}$ and Y:

$$\max_{\|\eta\|=1} \mathcal{V}^2(\eta^\top\boldsymbol{Z}, Y), \tag{2}$$

where $\mathcal{V}^2(U,V)$ denotes the squared distance covariance. For random variables U and V with independent copies (U', V') and (U'', V''), this is defined as

$$\mathcal{V}^2(U,V) = \mathbb{E}|U-U'||V-V'| + \mathbb{E}|U-U'|\mathbb{E}|V-V'| - \mathbb{E}|U-U'||V-V''| - \mathbb{E}|U-U''||V-V'|.$$

Under mild conditions, the maximizer η of (2) spans the standardized central subspace, yielding $b = \Sigma^{-1/2}\eta$.

Given the sample, we compute the U-statistic estimator:

$$\mathcal{V}_n^2(\eta^\top\widehat{\boldsymbol{Z}}, Y) = \frac{1}{n^2}\sum_{k,l=1}^{n} A_{kl}(\eta)B_{kl},$$

where $A_{kl}(\eta)$ and B_{kl} are centered pairwise distance matrices for $\eta^\top\widehat{\boldsymbol{Z}}$ and Y, respectively (Székely et al., 2007). The DCOV estimator $\widehat{\eta}_{\mathrm{DCOV}}$ solves $\max_{\|\eta\|=1}\mathcal{V}_n^2(\eta^\top\widehat{\boldsymbol{Z}}, Y)$ via sequential quadratic programming, and the estimator is $\widehat{b}_{\mathrm{DCOV}} = \widehat{\Sigma}^{-1/2}\widehat{\eta}_{\mathrm{DCOV}}$ with $\widehat{b}_{\mathrm{DCOV}}^\top\widehat{b}_{\mathrm{DCOV}} = 1$.

2.4 Unified Asymptotic Properties

Under standard regularity conditions (linearity condition for SIR and PSVM; compact support and finite first moments for DCOV), all three estimators achieve parametric convergence. Specifically, for $\widehat{b} \in \{\widehat{b}_{\mathrm{SIR}}, \widehat{b}_{\mathrm{PSVM}}, \widehat{b}_{\mathrm{DCOV}}\}$ normalized such that $\widehat{b}^\top\widehat{b} = 1$:

$$\|\widehat{b} - b\| = O_p(n^{-1/2}). \tag{3}$$

Furthermore, under mild regularity conditions specific to each method, all three estimators satisfy asymptotic normality:

$$\sqrt{n}(\widehat{b} - b) \xrightarrow{d} \mathcal{N}(0, \boldsymbol{\Sigma}_b),$$

where the exact form of the asymptotic covariance matrix $\boldsymbol{\Sigma}_b$ depends on the specific SDR method employed; see Li (1991) for SIR, Li et al. (2011) for PSVM, and Sheng and Yin (2016) for DCOV.

This $\sqrt{n}$-consistency and asymptotic normality establish that each method attains statistical efficiency comparable to the oracle estimator, providing rigorous guarantees for their distributed implementation.

3 Distributed Estimation

3.1 Distributed SDR for Index Vector Estimation

We now extend the SDR methods reviewed in Section 2 to the distributed setting. Suppose the full dataset $\{(\boldsymbol{X}_i, Y_i)\}_{i=1}^N$ is partitioned across m local machines, with the k-th machine holding a subset S_k of size n_k, where $\sum_{k=1}^m n_k = N$. Our framework accommodates general heterogeneous partitions without requiring balanced allocation; the convergence guarantees for the aggregated estimators are established under suitable conditions on the total sample size N and the number of machines m, thereby covering a broad range of realistic distributed computing environments including unbalanced workload distributions.

On each local machine k, we apply any of the three SDR methods (SIR, PSVM, or DCOV) described in Section 2 to obtain the local estimate $\widehat{b}_k$. Specifically, machine k computes its local sample moments $\widehat{\mu}_k$ and $\widehat{\Sigma}_k$, standardizes the local data to $\widehat{\boldsymbol{Z}}_i^{(k)} = \widehat{\Sigma}_k^{-1/2}(\boldsymbol{X}_i - \widehat{\mu}_k)$ for $i \in S_k$, and solves for the local direction $\widehat{\eta}_k$ in the standardized space. The local index vector is then $\widehat{b}_k = \widehat{\Sigma}_k^{-1/2}\widehat{\eta}_k$, normalized to satisfy $\widehat{b}_k^\top \widehat{b}_k = 1$.

Since each $\widehat{b}_k$ estimates the global direction b up to sign and lies on the ellipsoid, simple arithmetic averaging is inappropriate due to potential sign cancellations and the non-Euclidean geometry of the parameter space. We propose a spectral aggregation method: the central machine collects the local estimates $\widehat{\boldsymbol{B}} = (\widehat{b}_1, \ldots, \widehat{b}_m) \in \mathbb{R}^{p\times m}$ and computes the final estimator $\widehat{b}$ as the leading eigenvector of the aggregate matrix $\widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^\top$. This approach respects the Grassmann manifold structure of the projection space and achieves consensus through the dominant mode of variation across local estimates.

Formally, the central server constructs the symmetric aggregate matrix $\widehat{\boldsymbol{M}} = \sum_{k=1}^m \frac{n_k}{N}\widehat{b}_k\widehat{b}_k^\top \in \mathbb{R}^{p\times p}$ from the collected local estimates. The aggregated direction $\widehat{b}$ is then identified as the eigenvector corresponding to the largest eigenvalue of $\widehat{\boldsymbol{M}}$, and is subsequently normalized to unit length so that $\|\widehat{b}\| = 1$.

Leave-one-out cross-fitted directions. To decouple the direction estimation from the link-function estimation on each machine, we construct *leave-one-out* direction estimators. For each machine k, the central server forms the leave-one-out aggregate matrix using all local estimates except the k-th:

$$\widehat{\boldsymbol{M}}_{[-k]} = \sum_{j\neq k} \frac{n_j}{N_{-k}} \widehat{b}_j\widehat{b}_j^\top \in \mathbb{R}^{p\times p}, \qquad N_{-k} := \sum_{j\neq k} n_j, \tag{4}$$

and computes the leave-one-out direction estimator $\widehat{b}_{[-k]}$ as the leading eigenvector of $\widehat{\boldsymbol{M}}_{[-k]}$, normalized to unit length. This construction mirrors the *cross-fitting* (or K-fold sample splitting)

strategy widely employed in double machine learning (Chernozhukov et al., 2018; Newey and Robins, 2018): the nuisance estimator ($\widehat{b}_{[-k]}$) is trained on one subset of the data (the $m-1$ other machines) and evaluated on a disjoint subset (machine k), thereby guaranteeing statistical independence between the nuisance estimate and the downstream regression data. A crucial consequence is that $\widehat{b}_{[-k]}$ is statistically independent of the data residing on machine k, because the latter is excluded from the spectral aggregation. This independence completely removes the circularity issue in the two-stage estimation and allows for a substantially cleaner theoretical analysis of the link function estimator. This stage requires two rounds of communication: (i) local machines transmit their p-dimensional direction estimates $\widehat{b}_k$ to the central server; (ii) the server computes and broadcasts the personalized leave-one-out directions $\widehat{b}_{[-k]}$ back to each machine.

3.2 Adaptive Distributed Kernel Ridge Regression with Cross-Fitting

Upon receiving its dedicated direction $\widehat{b}_{[-k]}$, machine k projects its local covariates as $Z_i = \widehat{b}_{[-k]}^\top \boldsymbol{X}_i$ for $i \in S_k$. The projected variable Z is univariate, which allows us to employ the AdaDKRR strategy of Lin et al. (2025) to select the regularization parameter and estimate the link function. The key idea is to avoid the over-fitting problem of purely local tuning—where the locally optimal regularization parameter is typically larger than the globally optimal one—by communicating non-sensitive basis coefficients so that each machine can validate against a globally synthesized approximation.

Global projection support and basis generation. Since the projected covariates $Z_i = \widehat{b}_{[-k]}^\top \boldsymbol{X}_i$ on different machines may exhibit heterogeneous ranges, it is essential to align the basis support before generating low-discrepancy sequences. Each machine k computes the local minimum and maximum of its projected training data, denoted $Z_{\min}^{(k)}$ and $Z_{\max}^{(k)}$, and transmits these two scalars to the central server. The server aggregates them into the global strict bounds

$$Z_{\min} = \min_{1\le k\le m} Z_{\min}^{(k)}, \qquad Z_{\max} = \max_{1\le k\le m} Z_{\max}^{(k)},$$

and broadcasts the pair $(Z_{\min}, Z_{\max})$ back to all machines. Each machine then independently generates the same set of basis centers $\Xi_n = \{\xi_\ell\}_{\ell=1}^n$ on the standardized interval $[0,1]$ using a low-discrepancy sequence such as the Sobol sequence or the Halton sequence (Dick and Pillichshammer, 2010; Leobacher and Pillichshammer, 2014), and maps them linearly to the global support $[Z_{\min}, Z_{\max}]$ via

$$\xi_\ell^{\text{raw}} = Z_{\min} + \xi_\ell (Z_{\max} - Z_{\min}), \qquad \ell = 1, \ldots, n.$$

Let $K(\cdot,\cdot)$ be the Mercer kernel on $\mathbb{R}\times\mathbb{R}$ (e.g., the Gaussian RBF kernel or the Matérn kernel). We define the basis function space

$$\mathcal{B}_{n,K} := \Big\{ \sum_{\ell=1}^n a_\ell K(\xi_\ell^{\text{raw}}, \cdot) : a_\ell \in \mathbb{R} \Big\}.$$

Because the projected space is one-dimensional, the star discrepancy of n Sobol points achieves $O(n^{-1})$ up to logarithmic factors, which is substantially faster than the $O(n^{-1/2})$ rate of Monte

Carlo sampling. Moreover, the generation cost is only $O(n \log n)$, and each machine can reproduce the identical sequence independently after receiving the two scalars $(Z_{\min}, Z_{\max})$, thereby respecting data-silo constraints with negligible communication overhead.

Local kernel ridge regression. For each candidate regularization parameter λ in a finite grid Λ, machine k first computes the standard local KRR estimator on its training subset S_k^{tr} (obtained via a hold-out or ι-fold split of S_k):

$$\widehat{f}_{k,[-k],\lambda} = \arg\min_{f \in \mathcal{H}} \Big\{ \frac{1}{n_k^{\mathrm{tr}}} \sum_{i \in S_k^{\mathrm{tr}}} (Y_i - f(Z_i))^2 + \lambda \|f\|_{\mathcal{H}}^2 \Big\}, \tag{5}$$

where $\mathcal{H}$ is the reproducing kernel Hilbert space (RKHS) associated with K. By the representer theorem (Wahba, 1990), the solution admits the form $\widehat{f}_{k,[-k],\lambda}(z) = \sum_{i \in S_k^{\mathrm{tr}}} \widehat{\alpha}_{k,i} K(Z_i, z)$.

Local approximation via basis expansion. Machine k then constructs a noise-free approximation of $\widehat{f}_{k,[-k],\lambda}$ within the basis space $\mathcal{B}_{n,K}$. Specifically, it forms the pseudo-data $\{(Z_i, \widehat{f}_{k,[-k],\lambda}(Z_i))\}_{i \in S_k^{\mathrm{tr}}}$ and solves

$$\widehat{f}^{\mathrm{loc}}_{k,\lambda,n,\mu} = \arg\min_{g \in \mathcal{B}_{n,K}} \Big\{ \frac{1}{n_k^{\mathrm{tr}}} \sum_{i \in S_k^{\mathrm{tr}}} (g(Z_i) - \widehat{f}_{k,[-k],\lambda}(Z_i))^2 + \mu \|g\|_{\mathcal{H}}^2 \Big\}, \tag{6}$$

where $\mu > 0$ is a small auxiliary regularization parameter introduced solely to overcome ill-conditioning. Writing $\widehat{f}^{\mathrm{loc}}_{k,\lambda,n,\mu}(\cdot) = \sum_{\ell=1}^{n} \widehat{a}^{\mathrm{loc}}_{k,\lambda,\ell} K(\xi_\ell^{\mathrm{raw}}, \cdot)$, we obtain the coefficient vector $\widehat{\boldsymbol{a}}^{\mathrm{loc}}_{k,\lambda} = (\widehat{a}^{\mathrm{loc}}_{k,\lambda,1}, \ldots, \widehat{a}^{\mathrm{loc}}_{k,\lambda,n})^\top \in \mathbb{R}^n$.

First round of communication. Each machine k transmits its coefficient matrix $\{\widehat{\boldsymbol{a}}^{\mathrm{loc}}_{k,\lambda}\}_{\lambda \in \Lambda}$ to the central server. This communication is privacy-preserving because the coefficients encode only the regression rule relative to the publicly known basis centers; no raw data or local kernel centers are revealed.

Global synthesis. The central server synthesizes the global coefficients by the sample-size-weighted average

$$\widehat{\boldsymbol{a}}^{\mathrm{glob}}_{\lambda} = \sum_{k=1}^{m} \frac{n_k}{N} \widehat{\boldsymbol{a}}^{\mathrm{loc}}_{k,\lambda}, \tag{7}$$

and constructs the global approximation

$$\widehat{f}^{\mathrm{glob}}_{\lambda,n,\mu}(\cdot) = \sum_{\ell=1}^{n} [\widehat{\boldsymbol{a}}^{\mathrm{glob}}_{\lambda}]_\ell K(\xi_\ell^{\mathrm{raw}}, \cdot) = \sum_{k=1}^{m} \frac{n_k}{N} \widehat{f}^{\mathrm{loc}}_{k,\lambda,n,\mu}(\cdot).$$

If the local approximations are accurate, then $\widehat{f}^{\mathrm{glob}}_{\lambda,n,\mu}$ is a good approximation of the global kernel ridge regression estimator that would be obtained from the full centralized sample. The server then communicates the global coefficient matrix $\{\widehat{\boldsymbol{a}}^{\mathrm{glob}}_{\lambda}\}_{\lambda \in \Lambda}$ back to every machine.

Local validation and adaptive parameter selection. Equipped with the global approximation, each machine k evaluates the validation error on its local validation set S_k^{val}:

$$\mathrm{MSE}_k(\lambda) = \frac{1}{n_k^{\mathrm{val}}} \sum_{i \in S_k^{\mathrm{val}}} \big(\pi_M \widehat{f}^{\mathrm{glob}}_{\lambda,n,\mu}(Z_i) - Y_i\big)^2, \tag{8}$$

where π_M is the truncation operator $\pi_M f(z) = \mathrm{sign}(f(z)) \min\{|f(z)|, M\}$ with M a uniform

bound on $|Y|$. The truncation operator is introduced only to facilitate theoretical analysis and does not alter the practical implementation. Machine k then selects

$$\widehat{\lambda}_k^* = \arg\min_{\lambda \in \Lambda} \mathrm{MSE}_k(\lambda). \tag{9}$$

Because the validation target $\widehat{f}^{\mathrm{glob}}_{\lambda,n,\mu}$ approximates the global rule rather than the local rule, the selected $\widehat{\lambda}_k^*$ is close to the globally optimal regularization parameter, thereby avoiding the systematic over-fitting inherent in purely local tuning.

It is worth noting that the global coefficients $\widehat{\boldsymbol{a}}^{\mathrm{glob}}_{\lambda}$ depend on all local machines, including machine k 's local estimate $\widehat{b}_k$, which contributes to the construction of every other machine's leave-one-out direction $\widehat{b}_{[-j]}(j \neq k)$ and hence enters the global synthesis indirectly. Consequently, the validation error $\mathrm{MSE}_k(\lambda)$ exhibits a weak indirect dependence on S_k through the parameter selection channel. However, because the candidate grid Λ is finite, this dependence is controlled by a standard Union Bound over the $L = \#\Lambda$ candidate models, and does not destroy the asymptotic convergence rate established in Theorem 4.2.

Extension to joint kernel parameter selection. The above procedure selects only the regularization parameter λ, assuming the kernel function K is fixed. When an adaptive kernel is used—for instance, the Gaussian RBF $K_\gamma(z, z') = \exp(-\gamma\|z - z'\|^2)$ with bandwidth γ—the bandwidth can be tuned jointly with λ without incurring any additional communication overhead. Specifically, one replaces the one-dimensional candidate grid Λ with a two-dimensional grid $\Lambda \times \Gamma$, where Γ contains candidate bandwidth values. Consequently, machine k computes local estimators $\widehat{f}_{k,[-k],\lambda,\gamma}$ in (5) and extracts basis coefficients $\widehat{\boldsymbol{a}}^{\mathrm{loc}}_{k,\lambda,\gamma}$ in (6) for every $(\lambda, \gamma) \in \Lambda \times \Gamma$; the server synthesizes $\widehat{\boldsymbol{a}}^{\mathrm{glob}}_{\lambda,\gamma}$ as in (7); machine k evaluates $\mathrm{MSE}_k(\lambda, \gamma)$ as in (8); and the optimal pair is selected by

$$(\widehat{\lambda}_k^*, \widehat{\gamma}_k^*) = \arg\min_{(\lambda,\gamma) \in \Lambda \times \Gamma} \mathrm{MSE}_k(\lambda, \gamma).$$

The communication cost scales as $O(mn \cdot \#\Lambda \cdot \#\Gamma)$, which remains negligible in the univariate setting because n can be taken small (e.g., $n \asymp \log^2 N$).

Final local retraining and aggregation. After the adaptive parameter $\widehat{\lambda}_k^*$ is determined, machine k retrains the local KRR estimator on its full local dataset S_k using $\widehat{\lambda}_k^*$:

$$\widehat{f}^*_{k,[-k]} = \arg\min_{f \in \mathcal{H}} \Big\{ \frac{1}{n_k} \sum_{i \in S_k} (Y_i - f(Z_i))^2 + \widehat{\lambda}_k^* \|f\|^2_{\mathcal{H}} \Big\}. \tag{10}$$

The global link function estimator is formed by averaging the m leave-one-out local fits evaluated at their respective projected coordinates. For a new observation $\boldsymbol{X}$, the prediction is

$$\widehat{f}(\boldsymbol{X}) = \frac{1}{m} \sum_{k=1}^{m} \widehat{f}^*_{k,[-k]}(\widehat{b}^\top_{[-k]} \boldsymbol{X}). \tag{11}$$

This expression explicitly acknowledges that each machine employs its own leave-one-out direction $\widehat{b}_{[-k]}$ for projection; averaging is performed in the response space after evaluating each local model at its personalized projected input.

The leave-one-out cross-fitting design distinguishes our framework from the original AdaD-

Algorithm 1 Adaptive Distributed Kernel Ridge Regression with Leave-One-Out Cross-Fitting

Input: Local datasets $\{(X_i, Y_i)\}_{i \in S_k}$; leave-one-out directions $\{\widehat{b}_{[-k]}\}_{k=1}^m$; kernel K; candidate grid Λ; auxiliary parameter $\mu > 0$.

Output: Global link function estimator $\widehat{f}$.

1: **Global Support.** Each machine computes local projected extrema and transmits them to the server. The server broadcasts global bounds $(Z_{\min}, Z_{\max})$, and all machines generate common basis centers $\{\xi_\ell^{\text{raw}}\}_{\ell=1}^n$.

2: **Local KRR and Basis Approximation.** For each $\lambda \in \Lambda$, machine k computes the local KRR estimator $\widehat{f}_{k,[-k],\lambda}$ on its training set, then projects it onto the basis space to obtain coefficients $\widehat{a}_{k,\lambda}^{\text{loc}}$. Transmit the coefficient matrix to the server.

3: **Global Synthesis.** The server aggregates coefficients by the sample-size-weighted average $\widehat{a}_\lambda^{\text{glob}} = \sum_{k=1}^m \frac{n_k}{N} \widehat{a}_{k,\lambda}^{\text{loc}}$ and broadcasts the global coefficient matrix back to all machines.

4: **Adaptive Selection.** Each machine validates the global approximation against its local hold-out set and selects $\widehat{\lambda}_k^* = \arg\min_{\lambda \in \Lambda} \text{MSE}_k(\lambda)$.

5: **Final Aggregation.** Each machine retrains $\widehat{f}_{k,[-k]}^*$ on its full local dataset using $\widehat{\lambda}_k^*$. Return the aggregated estimator

$$\widehat{f}(X) = \frac{1}{m} \sum_{k=1}^m \widehat{f}_{k,[-k]}^*(\widehat{b}_{[-k]}^\top X).$$

KRR of Lin et al. (2025) in a manner deeply analogous to the distinction between single-sample and cross-fitted nuisance estimation in semiparametric inference (Chernozhukov et al., 2018). In the original AdaDKRR, each local machine validates the global approximation against its own validation set, but the global approximation is synthesized from local estimates that are trained on the same data subsets (up to a train-validation split) that participate in the aggregation. This creates a weak data-reuse loop: the data on machine k indirectly influence the global tuning target through their contribution to other machines' local approximations. Such *overfitting to the same sample* is precisely the phenomenon that cross-fitting was designed to prevent (Newey and Robins, 2018). In our cross-fitted variant, although the global coefficient synthesis still aggregates information across all machines, the projection direction $\widehat{b}_{[-k]}$ used by machine k is statistically independent of S_k by construction. Consequently, the local KRR estimator $\widehat{f}_{k,[-k],\lambda}$ in (5) is trained on data whose projected coordinates depend only on directions that exclude S_k. This independence eliminates the circularity between direction estimation and link-function estimation, removes the need for uniform oracle inequalities over shrinking neighborhoods, and frees the convergence rate of $\widehat{f}$ from the extraneous $\sqrt{\log N}$ factors that typically arise in coupled two-stage analyses (Chernozhukov et al., 2018). From a practical standpoint, the additional communication of two scalars $(Z_{\min}, Z_{\max})$ per machine and the computation of m leave-one-out eigen-decompositions at the central server are negligible overhead compared with the gains in statistical cleanliness and theoretical sharpness.

The adaptive tuning and distributed aggregation procedure for link function estimation is summarized in Algorithm 1.

3.3 Algorithm and Communication Cost

The complete two-stage estimation procedure is summarized in Algorithm 2, with the detailed adaptive kernel ridge regression step elaborated in Algorithm 1. Algorithm 2 proceeds through

Algorithm 2 Distributed Estimation for Nonlinear Regression (Single-Index Case)

Input: Number of machines m, local datasets $\{(\boldsymbol{X}_i, Y_i)\}_{i \in S_k}$ for $k = 1, \ldots, m$, kernel function $K(\cdot, \cdot)$, SDR method (SIR, PSVM, or DCOV),candidate grid Λ, auxiliary parameter μ, truncation level M.

Output: Estimated index vector $\widehat{b}$ and link function $\widehat{f}$.

1: **Local SDR:** Each machine k computes local moments $\widehat{\mu}_k$, $\widehat{\Sigma}_k$, standardizes the data, and obtains the local direction $\widehat{b}_k = \widehat{\Sigma}_k^{-1/2} \widehat{\eta}_k$ using the specified SDR method, normalized to $\widehat{b}_k^\top \widehat{b}_k = 1$.

2: **Central Aggregation:** Transmit $\{\widehat{b}_k\}_{k=1}^m$ to the central machine. For each k, form $\widehat{\boldsymbol{M}}_{[-k]}$ as in (4) and compute the leave-one-out direction $\widehat{b}_{[-k]}$.

3: **Personalized Broadcast:** Send $\widehat{b}_{[-k]}$ to machine k for each $k = 1, \ldots, m$.

4: **Projection:** Each machine k computes projected covariates $Z_i = \widehat{b}_{[-k]}^\top \boldsymbol{X}_i$ for $i \in S_k$, and generates the same basis centers Ξ_n.

5: **Local Approximation:** For each $\lambda \in \Lambda$, machine k computes $\widehat{f}_{k,[-k],\lambda}$ by (5), then obtains the basis coefficients $\widehat{\boldsymbol{a}}_{k,\lambda}^{\text{loc}}$ by solving (6).

6: **Communication I:** Transmit $\{\widehat{\boldsymbol{a}}_{k,\lambda}^{\text{loc}}\}_{\lambda \in \Lambda}$ to the central server.

7: **Global Synthesis:** The server computes $\widehat{\boldsymbol{a}}_\lambda^{\text{glob}}$ by (7) for all $\lambda \in \Lambda$, and broadcasts the global coefficient matrices back.

8: **Local Validation:** Each machine k computes $\text{MSE}_k(\lambda)$ by (8) and selects $\widehat{\lambda}_k^*$.

9: **Final Local Training:** Each machine k computes $\widehat{f}_{k,[-k]}^*$ by (10) using $\widehat{\lambda}_k^*$.

10: **Final Aggregation:** Compute the global link function $\widehat{f}(\boldsymbol{X}) = \frac{1}{m} \sum_{k=1}^m \widehat{f}_{k,[-k]}^*(\widehat{b}_{[-k]}^\top \boldsymbol{X})$.

four sequential communication rounds: **Round 1** (SDR upload): each machine transmits its local direction $\widehat{b}_k$ to the center ($O(mp)$); **Round 2** (SDR broadcast): the center computes and sends back the leave-one-out directions $\widehat{b}_{[-k]}$ ($O(mp)$); **Round 3** (AdaDKRR upload): each machine transmits its basis coefficient matrices $\{\widehat{\boldsymbol{a}}_{k,\lambda}^{\text{loc}}\}_{\lambda \in \Lambda}$ ($O(mnL)$); **Round 4** (AdaDKRR broadcast): the center synthesizes and broadcasts the global coefficients $\{\widehat{\boldsymbol{a}}_\lambda^{\text{glob}}\}_{\lambda \in \Lambda}$ ($O(mnL)$). Additionally, to align the global projection support before basis generation, each machine communicates two scalars $(Z_{\min}^{(k)}, Z_{\max}^{(k)})$ to the center, and the center broadcasts the aggregated bounds $(Z_{\min}, Z_{\max})$ back; this incurs $O(m)$ overhead, which is absorbed by the $O(mp)$ term. The total communication complexity is

$$O(mp) + O(mnL) = O(mp + mnL).$$

These four rounds constitute a non-iterative pipeline: each round serves a distinct function, and there are no repeated optimization loops or feedback cycles. This stands in sharp contrast to iterative distributed methods such as IDC (Jiang and Peng, 2023), where the number of communication rounds scales with the iteration count required for convergence.

4 Theoretical Properties

We establish the statistical guarantees for the distributed estimators under the *leave-one-out cross-fitting* design described in Section 3. Recall that for each machine k, the direction estimator $\widehat{b}_{[-k]}$ is constructed from all machines except k, and is therefore statistically independent of the

local data S_k used for kernel ridge regression. This independence holds for every $k = 1, \ldots, m$ simultaneously.

Throughout this section, we assume the full dataset is partitioned across m machines with local sample sizes $\{n_k\}_{k=1}^m$ satisfying $\sum_{k=1}^m n_k = N$ and $n_0 = \min_{1\le k\le m} n_k \to \infty$. We write $N_{-k} = \sum_{j\neq k} n_j$ and $N_- = \min_{1\le k\le m} N_{-k}$.

4.1 Regularity Conditions

Assumption 4.1 (Moment Condition)**.** $\mathbb{E}[\|\boldsymbol{X}\|^2] < \infty$, $\mathbb{E}[|Y|^2] < \infty$, and the predictor $\boldsymbol{X}$ has a sub-Gaussian tail, i.e., $\mathbb{E}[\exp(\tau\|\boldsymbol{X}\|^2)] < \infty$ for some $\tau > 0$.

Assumption 4.2 (Kernel Regularity)**.** The kernel $K(\cdot,\cdot)$ is continuous, positive definite, and satisfies the Lipschitz condition $|K(z,z') - K(z,z'')| \le L_K|z' - z''|$ for some constant $L_K > 0$. Furthermore, the eigenvalues $\{\mu_j\}$ of the integral operator associated with K satisfy $\mu_j = O(j^{-2\alpha})$ for some smoothness parameter $\alpha > 1/2$. In addition, $\sup_z K(z,z) \le \kappa < \infty$.

Assumption 4.3 (Noise Condition)**.** The error term ε satisfies $\mathbb{E}[\varepsilon \mid \boldsymbol{X}] = 0$ and $\mathrm{Var}(\varepsilon \mid \boldsymbol{X}) \le \sigma^2 < \infty$ almost surely. Moreover, there exists $M > 0$ such that $|Y| \le M$ almost surely.

Remark 4.1 (On the truncation operator π_M)**.** The truncation π_M is introduced solely to ensure boundedness for Bernstein's inequality in the proof; it is omitted in the practical implementation (Algorithm 1), as the RKHS regularization already guards against extreme fitted values.

Assumption 4.4 (Spectral Gap and Identifiability)**.** The true index vector b satisfies $\|b\| = 1$, and its first nonzero coordinate is constrained to be positive. Let $\boldsymbol{B} = (b_1, \ldots, b_m) \in \mathbb{R}^{p\times m}$ with $b_1 = \ldots = b_m = b$ and define the population weighted aggregate matrix $\boldsymbol{M}_{\mathrm{pop}} = \sum_{k=1}^m \frac{n_k}{N} b_k b_k^\top = bb^\top \in \mathbb{R}^{p\times p}$. The largest eigenvalue $\lambda_1(\boldsymbol{M}_{\mathrm{pop}}) = 1$ is simple with spectral gap $\lambda_1(\boldsymbol{M}_{\mathrm{pop}}) - \lambda_2(\boldsymbol{M}_{\mathrm{pop}}) = 1$, where $\lambda_2(\boldsymbol{M}_{\mathrm{pop}}) = 0$.

Assumption 4.5 (Link Function Smoothness)**.** The true link function $f_0(\cdot)$ is Lipschitz continuous: there exists a constant $L_f < \infty$ such that $|f_0(z_1) - f_0(z_2)| \le L_f|z_1 - z_2|$ for all z_1, z_2 in the support of $b^\top \boldsymbol{X}$.

Assumption 4.6 (Marginal Distribution and Quadrature)**.** For each machine k, the projected variable $Z = \widehat{b}_{[-k]}^\top \boldsymbol{X}$ (after min-max standardization to $[0,1]$) has a marginal distribution ρ_Z such that the basis centers $\Xi_n = \{\xi_\ell\}_{\ell=1}^n$ satisfy the quadrature property: there exist constants $c, \beta > 0$ with

$$\sup_{\|f\|_K\le 1, \|g\|_K\le 1} \Big| \int_0^1 f(z)g(z)\, d\rho_Z(z) - \frac{1}{n}\sum_{\ell=1}^n f(\xi_\ell)g(\xi_\ell) \Big| \le cn^{-\beta}.$$

Moreover, the basis size n and the auxiliary parameter μ satisfy $\mu n^\beta \ge 2c$.

Remark 4.2 (Practical realization of the quadrature condition)**.** In practice, if the projected marginal distribution ρ_Z is relatively smooth over the global support $[Z_{\min}, Z_{\max}]$, the simple min-max linear mapping of low-discrepancy sequences (e.g., Sobol or Halton) already yields satisfactory empirical performance. A theoretically perfect quadrature would require communicating the global empirical cumulative distribution function (ECDF) to perform a quantile

transform, which incurs significantly higher communication costs and violates the data-silo constraint. Algorithm 1 adopts the min-max linear mapping as a communication-efficient empirical approximation; the quadrature error observed in simulations is indistinguishable from the theoretical bound under moderate smoothness.

Assumption 4.7 (Local Approximation Stability and Grid Resolution)**.** The auxiliary regularization parameter μ in the local approximation step satisfies

$$\mu \in \left[(8C_1^*(\log(1+\kappa)+2))^2 \max_{k=1,\dots,m} \left\{ \frac{\log^4 n_k}{n_k} \right\},\ N^{-\frac{2\alpha}{2\alpha+1}} \right],$$

where $C_1^* := 4\max\{(\kappa^2+1)/3, 2\sqrt{\kappa^2+1}\}$. The candidate set Λ forms a logarithmic equispaced grid with cardinality $L = \#\Lambda$, covering the range $[N^{-\frac{2\alpha}{2\alpha+1}-\Delta},\ N^{-\frac{2\alpha}{2\alpha+1}+\Delta}]$ for some fixed $\Delta > 0$.

Assumption 4.8 (Homogeneous and Non-Adversarial Environment)**.** The data partitions $\{S_k\}_{k=1}^m$ are drawn from the same underlying population distribution, and all local machines are honest (non-adversarial). Consequently, the data are independent and identically distributed (i.i.d.) across and within machines.

Assumptions 4.1–4.3 are standard in the SDR and nonparametric regression literature (Li, 1991; Wahba, 1990; Zhang et al., 2015). Assumption 4.4 ensures that the spectral aggregation method can effectively distinguish the true direction from noise. Assumption 4.5 is a mild regularity condition used to control the approximation error induced by replacing the true projection $b^\top \boldsymbol{X}$ with the estimated leave-one-out projection $\widehat{b}_{[-k]}^\top \boldsymbol{X}$. Assumption 4.6 is the quadrature condition required for the low-discrepancy basis approximation in the AdaDKRR framework (Lin et al., 2025); in the univariate setting it is readily satisfied by Sobol or Halton sequences when ρ_Z is uniform (or after a preliminary quantile transform). Assumption 4.7 guarantees the well-conditionedness of the local basis approximation and ensures that the candidate set captures the globally optimal regularization parameter.

4.2 Convergence of the Leave-One-Out Direction Estimator

In the leave-one-out cross-fitting protocol, the direction estimator transmitted to machine k is constructed by excluding the k-th machine's local estimate from the spectral aggregation. Formally, the central server forms the weighted leave-one-out aggregate matrix $\widehat{\boldsymbol{M}}_{[-k]}$ as in (4) and computes $\widehat{b}_{[-k]}$ as the leading eigenvector of $\widehat{\boldsymbol{M}}_{[-k]}$, normalized to unit length. A crucial consequence is that $\widehat{b}_{[-k]}$ is statistically independent of the data residing on machine k, because the latter is excluded from the spectral aggregation. This independence is the cornerstone that enables the decoupled analysis of the subsequent link-function estimation.

The following theorem establishes the parametric convergence of $\widehat{b}_{[-k]}$ under both fixed and diverging numbers of machines, as well as its asymptotic normality. The proof is given in Appendix A.1.

Theorem 4.1 (Convergence of Leave-One-Out Distributed SDR)**.** *Let $N_{-k} = \sum_{j\neq k} n_j$ denote the total sample size of the $m-1$ machines used to construct $\widehat{b}_{[-k]}$. Under Assumptions 4.1*

and 4.4, the leave-one-out direction estimator satisfies the universal consistency rate

$$\|\widehat{b}_{[-k]} - b\| = O_p\Big(N_{-k}^{-1/2}\Big), \qquad k = 1, \ldots, m. \tag{12}$$

The asymptotic behavior under the two canonical regimes is as follows.

(i) ***Fixed*** m**.** *When m is fixed and $N_{-k} \asymp N$ for all k (e.g., balanced partitions $n_j = N/m$), the optimal parametric rate holds:*

$$\|\widehat{b}_{[-k]} - b\| = O_p\Big(N^{-1/2}\Big). \tag{13}$$

Moreover, if $\max_{j\neq k} N/n_j = O(1)$, then

$$\sqrt{N}\,(\widehat{b}_{[-k]} - b) \xrightarrow{d} \mathcal{N}\left(\mathbf{0},\ \sum_{j\neq k} \frac{N\,n_j}{N_{-k}^2}\,\mathbf{\Sigma}_j\right), \tag{14}$$

where $\mathbf{\Sigma}_j = \lim_{n_j\to\infty} n_j \cdot \mathrm{Cov}(\widehat{b}_j)$. In the balanced case $n_j = N/m$, the asymptotic covariance simplifies to $\frac{m}{(m-1)^2}\sum_{j\neq k}\mathbf{\Sigma}_j$.

(ii) ***Diverging*** m**.** *When $m \to \infty$ with $n_j \asymp N/m$ (balanced in order), if $m = o(\sqrt{N})$, then the rate becomes*

$$\|\widehat{b}_{[-k]} - b\| = O_p(N^{-1/2}).$$

If, in addition, $\sum_{j\neq k}(n_j/N)^2 \to 0$, then

$$\sqrt{N}\,(\widehat{b}_{[-k]} - b) \xrightarrow{d} \mathcal{N}\left(\mathbf{0},\ \lim_{m\to\infty}\sum_{j\neq k} \frac{N\,n_j}{N_{-k}^2}\,\mathbf{\Sigma}_j\right). \tag{15}$$

Under exact balance $n_j = N/m$, the asymptotic covariance converges to $\lim_{m\to\infty}\frac{1}{m}\sum_{j\neq k}\mathbf{\Sigma}_j$.

Remark 4.3 (Independence and decoupling via cross-fitting)**.** By construction, $\widehat{b}_{[-k]} \perp\!\!\!\perp S_k$. Consequently, the projection direction is exogenous to the regression data on machine k, eliminating the circularity and uniform oracle inequalities required by coupled two-stage estimators.

Remark 4.4 (Rate uniformity and sample-size weighting)**.** The universal bound (12) reveals that the convergence of $\widehat{b}_{[-k]}$ is governed by the *total* sample size N_{-k} of the participating machines, rather than by the weakest single machine. Under balanced partitions, $N_{-k} \asymp N$ and all leave-one-out estimators share the same rate $O_p(N^{-1/2})$ uniformly over k, irrespective of whether m is fixed or diverging. This demonstrates that the sample-size-weighted spectral aggregation eliminates the "weakest-machine" bottleneck and achieves genuine robustness to unbalanced partitions.

Remark 4.5 (Comparison with existing distributed single-index estimators)**.** Existing distributed methods for single-index models predominantly rely on either iterative refinement or naive Euclidean averaging of local directions. The iterative divide-and-conquer (IDC) method of Jiang and Peng (2023) achieves asymptotic normality through multiple rounds of communication, incurring synchronization costs that scale linearly with the iteration count and requiring

repeated access to local data. The DC-rOPG method of Li and Yang (2025) employs one-shot Euclidean averaging of local outer-product-gradient directions, which is vulnerable to sign cancellation when local machines report opposite orientations of the same direction and requires ad-hoc sign-alignment pre-processing. In contrast, our spectral aggregation on the Grassmann manifold via $\widehat{\boldsymbol{M}}_{[-k]}$ intrinsically resolves sign ambiguity by seeking the dominant mode of variation across local estimates. Moreover, the leave-one-out design guarantees statistical independence between the aggregated direction and the local regression data—a property that neither IDC nor DC-rOPG possesses. This independence not only simplifies the theoretical analysis but also removes the restrictive balanced-sample assumption. Indeed, our universal rate $O_p(N_{-k}^{-1/2})$ depends on the total sample size of the participating machines and eliminates the weakest-machine bottleneck.

4.3 Convergence of Link Function Estimation

Under the leave-one-out cross-fitting design, each machine k receives a direction $\widehat{b}_{[-k]}$ that is independent of its local data S_k. This independence allows us to decompose the local error cleanly via the L^2–orthogonality of conditional expectation, and then exploit the disjointness of local subsamples to obtain a sharp variance-reduction bound for the aggregated estimator. The following theorem establishes the minimax optimal convergence rate for the link function estimator under the adaptive parameter selection strategy. The proof is given in Appendix A.2.

Theorem 4.2 (Optimal Convergence of Adaptive DKRR with Leave-One-Out Cross-Fitting)**.** *Under Assumptions 4.1–4.7, suppose the number of machines satisfies*

$$m \leq \min\Big\{N^{\frac{2\alpha-1}{4\alpha+2}} \log^{-4} N,\ N^{\frac{1}{2\alpha+1}} \log^{-1} L\Big\},$$

where $L = \#\Lambda$. Then the aggregated link function estimator $\widehat{f}$ obtained by Algorithm 2 satisfies

$$\|\widehat{f} - f_0\|_{L^2} = O_p\Big(N^{-\frac{\alpha}{2\alpha+1}} + N_-^{-\frac{1}{2}} + L^{-1}\Big), \tag{16}$$

where $N_- = \min_{1\leq k\leq m} N_{-k}$. Under both fixed and diverging numbers of machines (with $N_- \asymp N$ in either regime, e.g. balanced partitions $n_k = N/m$), the rate reduces to

$$\|\widehat{f} - f_0\|_{L^2} = O_p\Big(N^{-\frac{\alpha}{2\alpha+1}} + N^{-\frac{1}{2}} + L^{-1}\Big).$$

Provided $\alpha \geq 1/2$ and $L \asymp \log N$, the parametric projection error and the grid discretization error are both dominated by the nonparametric oracle error, yielding the optimal minimax rate

$$\|\widehat{f} - f_0\|_{L^2} = O_p(N^{-\alpha/(2\alpha+1)}).$$

Remark 4.6 (Rate decomposition and the absence of logarithmic factors)**.** The rate $N^{-\alpha/(2\alpha+1)}$ is the minimax optimal rate for estimating a univariate function with smoothness α in the RKHS (Stone, 1982; Zhang et al., 2015). The second term $N_-^{-1/2}$ reflects the cost of estimating the projection direction: each local machine must use its own leave-one-out direction $\widehat{b}_{[-k]}$ rather than the true index vector b. Because $\widehat{b}_{[-k]}$ is statistically independent of the local regression

data on machine k, this projection error is governed solely by the parametric convergence of the direction estimator and carries *no additional logarithmic factor*. This stands in sharp contrast to coupled two-stage analyses where the dependence between $\widehat{b}$ and the kernel matrices necessitates a uniform oracle inequality over shrinking neighborhoods, which typically induces an extraneous $\sqrt{\log N}$ penalty in the final rate. This mirrors the "fast remainder rates" phenomenon in cross-fitted semiparametrics (Chernozhukov et al., 2018; Newey and Robins, 2018): the empirical process term is evaluated on an independent sample, so the second-stage estimator attains the oracle minimax rate without $\sqrt{\log N}$ penalties.

Remark 4.7 (Comparison with existing distributed single-index and kernel methods)**.** Theorem 4.2 provides the first minimax optimal convergence rate for the distributed nonparametric link function. Prior works (Jiang and Peng, 2023; Li and Yang, 2025) treat only the parametric index vector. In the broader kernel literature, Zhang et al. (2015) require an identical oracle λ across machines, while Xu et al. (2019) incur heavy communication via complete kernel matrices. Our AdaDKRR framework circumvents both by communicating only low-dimensional basis coefficients, and the cross-fitted design eliminates the $\sqrt{\log N}$ penalties inherent in coupled analyses.

5 Numerical Studies

We conduct extensive Monte Carlo experiments to evaluate the finite-sample performance of the proposed distributed estimation framework. The first three simulation studies focus primarily on the accuracy of the direction estimation under varying distributed scenarios: Simulation I examines the effect of the number of local machines, Simulation II investigates the impact of predictor dimension, and Simulation III assesses robustness under skew-normal predictors that violate the linearity condition. Simulation IV then provides a comprehensive comparison with existing distributed approaches, evaluating both the index vector estimation error and the out-of-sample prediction accuracy of the estimated link function.

All simulations are performed on a workstation with a 12th Gen Intel Core i9-12900K 3.20 GHz processor and 128 GB RAM. Each experiment is replicated 100 times. We assess the accuracy of the index vector estimator $\widehat{b}$ via the Frobenius norm distance

$$D = \left\| \widehat{b}\widehat{b}^\top - bb^\top \right\|_F,$$

and evaluate the link function estimation by the root mean squared error (RMSE) of prediction. We consider the following four single-index models:

$$\begin{aligned}
&\text{Model 1:} \quad Y = \exp(b^\top \boldsymbol{X}) + \cos(5b^\top \boldsymbol{X}) + \varepsilon;\\
&\text{Model 2:} \quad Y = -3b^\top \boldsymbol{X} + 0.5(b^\top \boldsymbol{X})^3 + \varepsilon;\\
&\text{Model 3:} \quad Y = \sin(b^\top \boldsymbol{X}) + 0.1\varepsilon;\\
&\text{Model 4:} \quad Y = \frac{\exp(4b^\top \boldsymbol{X})}{1+\exp(4b^\top \boldsymbol{X})} + 0.1\varepsilon,
\end{aligned}$$

where $\varepsilon \sim N(0,1)$ is independent of $\boldsymbol{X}$.

In all distributed experiments, the total sample size N is randomly partitioned across m machines such that each local sample size n_k fluctuates within $\pm 30\%$ of the average allocation N/m, i.e., $n_k \in [0.7N/m,\, 1.3N/m]$ with $\sum_{k=1}^{m} n_k = N$. For SIR and PSVM, the number of slices (or dichotomization cut-off points) is set equal to the predictor dimension p. For DCOV, the initial direction is constructed by computing the distance covariance between each predictor component and the response Y, with the sign of each component determined by the corresponding Pearson correlation coefficient; the resulting vector is then normalized to unit length.

5.1 Simulation I: Varying Number of Machines

We fix the total sample size at $N = 2^{14} = 16384$ and the predictor dimension at $p = 5$. The covariates are generated from $\boldsymbol{X} \sim N_p(\boldsymbol{0}, \boldsymbol{I}_p)$ and the index vector is set to $b = (1, \ldots, 1)^\top/\sqrt{p}$. We vary the number of machines $m \in \{1, 20, 40, 60, 80, 100\}$. Tables 1 and 2 report the averaged values of D and the computational time (in seconds), respectively.

Table 1: Index vector estimation error D for Simulation I.

Model	Method	$m = 1$	$m = 20$	$m = 40$	$m = 60$	$m = 80$	$m = 100$
1	DC-SIR	0.0248	0.0255	0.0246	0.0248	0.0252	0.0259
	DC-PSVM	0.0177	0.0189	0.0190	0.0193	0.0204	0.0206
	DC-DCOV	0.0145	0.0156	0.0157	0.0161	0.0168	0.0180
2	DC-SIR	0.0190	0.0200	0.0206	0.0202	0.0205	0.0209
	DC-PSVM	0.0135	0.0140	0.0137	0.0141	0.0147	0.0139
	DC-DCOV	0.0130	0.0134	0.0133	0.0133	0.0133	0.0131

Table 2: Computational time (seconds) for Simulation I.

Model	Method	$m = 1$	$m = 20$	$m = 40$	$m = 60$	$m = 80$	$m = 100$
1	DC-SIR	0.0010	0.0031	0.0030	0.0034	0.0042	0.0050
	DC-PSVM	15.7918	1.0031	0.8074	0.8345	0.9384	1.0596
	DC-DCOV	19.3235	1.5428	0.9663	0.6635	0.7254	0.8106
2	DC-SIR	0.0010	0.0031	0.0029	0.0035	0.0043	0.0051
	DC-PSVM	11.9566	0.8939	0.7646	0.8051	0.9157	1.0483
	DC-DCOV	19.4559	1.4861	0.8796	0.5769	0.6159	0.6796

Several observations emerge from Tables 1 and 2. First, for a fixed total sample size, the estimation accuracy of all three SDR methods remains remarkably stable as m increases from 1 to 100. This confirms that the spectral aggregation procedure in Algorithm 2 successfully pools information across local machines without incurring appreciable efficiency loss, corroborating the $\sqrt{N}$-consistency established in Theorem 4.1. Second, the computational time is drastically reduced in the distributed setting. For instance, DCOV requires approximately 19 seconds on a single machine but only about 0.8 seconds with $m = 100$, achieving nearly linear speed-up. SIR is consistently the fastest, while PSVM offers an attractive balance between accuracy and speed. Third, PSVM and DCOV generally yield smaller D than SIR, reflecting their robustness to the link function.

5.2 Simulation II: Varying Predictor Dimension

We fix $N = 5000$ and $m = 10$, and vary the predictor dimension $p \in \{3, 5, 10, 15, 20\}$. The covariates are drawn from a multivariate normal distribution with an AR(1) covariance structure $\Sigma_{ij} = 0.25^{|i-j|}$, and the index vector is generated as $b_j = 0.5^j$ with normalization $\|b\| = 1$. We report results for Models 1 and 2 in Tables 3 and 4.

Table 3: Index vector estimation error D for Simulation II.

Model	m	Method	$p = 3$	$p = 5$	$p = 10$	$p = 15$	$p = 20$
1	1	DC-SIR	0.0378	0.0467	0.0696	0.0845	0.0981
		DC-PSVM	0.0295	0.0332	0.0498	0.0574	0.0617
		DC-DCOV	0.0192	0.0284	0.0464	0.0589	0.0685
	10	DC-SIR	0.0396	0.0510	0.0762	0.0991	0.1191
		DC-PSVM	0.0325	0.0372	0.0534	0.0647	0.0737
		DC-DCOV	0.0208	0.0318	0.0529	0.0701	0.0845
2	1	DC-SIR	0.0407	0.0564	0.0838	0.1030	0.1209
		DC-PSVM	0.0243	0.0356	0.0483	0.0599	0.0701
		DC-DCOV	0.0190	0.0296	0.0428	0.0548	0.0633
	10	DC-SIR	0.0432	0.0629	0.0974	0.1283	0.1528
		DC-PSVM	0.0263	0.0395	0.0533	0.0691	0.0801
		DC-DCOV	0.0202	0.0315	0.0461	0.0596	0.0685

Table 4: Computational time (seconds) for Simulation II.

Model	m	Method	$p = 3$	$p = 5$	$p = 10$	$p = 15$	$p = 20$
1	1	DC-SIR	0.0005	0.0005	0.0007	0.0008	0.0009
		DC-PSVM	0.8245	1.9386	6.5336	14.6503	27.7598
		DC-DCOV	1.8766	2.2139	3.0373	3.8780	4.6497
	10	DC-SIR	0.0010	0.0014	0.0021	0.0033	0.0043
		DC-PSVM	0.1192	0.2447	0.7616	1.6910	3.1107
		DC-DCOV	0.3076	0.3706	0.5177	0.6610	0.7983
2	1	DC-SIR	0.0004	0.0005	0.0006	0.0007	0.0009
		DC-PSVM	0.9789	1.7520	6.4055	14.5461	27.2526
		DC-DCOV	2.1446	2.4373	3.1663	3.8935	4.6233
	10	DC-SIR	0.0010	0.0012	0.0019	0.0029	0.0038
		DC-PSVM	0.1249	0.2279	0.7541	1.6731	3.1266
		DC-DCOV	0.3028	0.3598	0.4796	0.6043	0.7190

As expected, the estimation error D increases with p for all methods, reflecting the curse of dimensionality inherent in nonparametric sufficient dimension reduction. Nevertheless, the distributed estimators ($m = 10$) track the centralized benchmarks ($m = 1$) closely across all dimensions, confirming that communication-efficient aggregation does not sacrifice statistical accuracy. The computational gains of distributed estimation become more pronounced as p grows.

5.3 Simulation III: Skew-Normal Predictors

To examine robustness to departures from elliptical symmetry, we generate the predictor vector $\boldsymbol{X}$ from a multivariate skew-normal distribution. We partition $\boldsymbol{X} = (\boldsymbol{X}_{(1)}^\top, \boldsymbol{X}_{(2)}^\top)^\top$, where the

first two components $\boldsymbol{X}_{(1)} \in \mathbb{R}^2$ follow a bivariate skew-normal distribution with density

$$2\Phi_2(\boldsymbol{x}_{(1)}; \boldsymbol{\mu}, \boldsymbol{\Sigma})\, \Phi\big(r\, \boldsymbol{c}^\top \mathrm{diag}(\boldsymbol{\Sigma})^{-1/2}(\boldsymbol{x}_{(1)} - \boldsymbol{\mu})\big),$$

where $\Phi_2(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the normal distribution $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ density, $\Phi(\cdot)$ is the standard normal cumulative distribution function, $\boldsymbol{\mu} = (0, 0)^\top$, $\boldsymbol{\Sigma} = \boldsymbol{I}_2$, and $\boldsymbol{c} = (1, 1)^\top$. The skewness is controlled by the parameter $r \in \{0, 0.2, 0.5, 1, 2\}$; when $r = 0$ the distribution reduces to the ordinary bivariate normal, and larger values induce stronger asymmetry that progressively violates the linearity condition required by SIR and PSVM. The remaining eight components $\boldsymbol{X}_{(2)} \in \mathbb{R}^8$ are drawn independently from $N(\boldsymbol{0}_8, \boldsymbol{I}_8)$. The true index vector is set to $b = (1, -1, \boldsymbol{0}_8^\top)^\top/\sqrt{2}$. We set $N = 5000$, $m = 10$, and $p = 10$, and report results for Models 3 and 4 in Tables 5 and 6.

Table 5: Index vector estimation error D under skew-normal predictors (Simulation III).

Model	m	Method	$r = 0$	$r = 0.2$	$r = 0.5$	$r = 1$	$r = 2$
3	1	DC-SIR	0.0255	0.0297	0.0487	0.2322	0.3870
		DC-PSVM	0.0111	0.0120	0.0185	0.1193	0.2285
		DC-DCOV	0.0220	0.0240	0.0253	0.0286	0.0389
	10	DC-SIR	0.0281	0.0320	0.0539	0.3451	1.1497
		DC-PSVM	0.0131	0.0142	0.0209	0.1307	0.2749
		DC-DCOV	0.0240	0.0259	0.0278	0.0315	0.0429
4	1	DC-SIR	0.0295	0.0327	0.0385	0.0664	0.1698
		DC-PSVM	0.0146	0.0150	0.0176	0.0259	0.0499
		DC-DCOV	0.0290	0.0319	0.0326	0.0460	0.0823
	10	DC-SIR	0.0322	0.0348	0.0418	0.0721	0.2236
		DC-PSVM	0.0162	0.0171	0.0203	0.0292	0.0636
		DC-DCOV	0.0317	0.0343	0.0358	0.0509	0.0989

Table 6: Computational time (seconds) under skew-normal predictors (Simulation III).

Model	m	Method	$r = 0$	$r = 0.2$	$r = 0.5$	$r = 1$	$r = 2$
3	1	DC-SIR	0.0007	0.0006	0.0006	0.0006	0.0006
		DC-PSVM	4.0823	4.6807	6.3752	6.2985	5.8678
		DC-DCOV	2.6711	2.6691	2.6810	2.9221	2.9281
	10	DC-SIR	0.0025	0.0025	0.0025	0.0026	0.0024
		DC-PSVM	0.6554	0.6784	0.7888	0.7600	0.6940
		DC-DCOV	0.4173	0.4161	0.4345	0.5038	0.5406
4	1	DC-SIR	0.0006	0.0006	0.0006	0.0006	0.0006
		DC-PSVM	3.6363	3.8677	4.3336	5.1061	5.5515
		DC-DCOV	2.6176	2.6419	2.7764	2.8722	2.9533
	10	DC-SIR	0.0020	0.0020	0.0019	0.0019	0.0019
		DC-PSVM	0.5705	0.5967	0.6090	0.6583	0.7047
		DC-DCOV	0.4229	0.4272	0.4405	0.4832	0.5973

For $r = 0$ and mild skewness ($r = 0.2$), all three methods perform comparably, as the linearity condition underlying SIR and PSVM is approximately satisfied. As r increases, the accuracy of SIR and PSVM deteriorates substantially because the linearity condition is violated. In contrast, DCOV exhibits strong robustness. This aligns with the model-free nature of distance covariance and underscores its reliability under distributional contamination. The computational patterns

are similar to those in Simulation I, with DCOV becoming slightly slower for large r due to the increased complexity of the distance matrix computation.

5.4 Simulation IV: Comparison with Alternative Methods

We compare the proposed distributed framework against the IDC method of Jiang and Peng (2023) and the AdaDKRR without dimension reduction. While we report the index vector estimation error for completeness, our primary focus is on evaluating and comparing the out-of-sample prediction accuracy of the estimated link function, as measured by the RMSE. We set $N = 5000$, $m = 10$, $p = 3$, and generate 500 additional observations for out-of-sample prediction. The index vector is set to $b = (1, -2, 1)^\top/\sqrt{6}$.

The ridge parameter λ is selected from a composite grid

$$\Lambda_{\text{grid}} = [\text{logspace}(-4, -2, 6),\ \text{logspace}(-1.5, 0, 4)],$$

formed by concatenating two logarithmically spaced segments: a strong-regularization segment $\text{logspace}(-4, -2, 6)$ and a weak-regularization segment $\text{logspace}(-1.5, 0, 4)$. The left segment, $\lambda \in [10^{-4}, 10^{-2}]$, provides strong regularization. This is essential for local machines with relatively small sample sizes—in our unbalanced setting each n_k ranges between $0.7N/m$ and $1.3N/m$—to prevent overfitting in the individual KRR fits. The right segment, $\lambda \in [10^{-1.5}, 10^{0}]$, covers weak regularization, reflecting the fact that after global aggregation the effective sample size increases and the optimal penalty should decrease accordingly.

The Gaussian kernel bandwidth γ is selected from the grid

$$\Gamma_{\text{grid}} = \text{logspace}(-2, 1, 10),$$

which spans from an extremely narrow kernel ($\gamma = 10^{-2}$) to a very wide kernel ($\gamma = 10^{1}$). This range covers the full spectrum from high-variance local interpolation to high-bias global smoothing, ensuring that the GCV minimum lies well inside the search range.

Here, $\text{logspace}(a, b, n)$ denotes a logarithmically spaced vector of n points between 10^a and 10^b, i.e., $\{10^{a+\frac{(b-a)k}{n-1}} : k = 0, \ldots, n-1\}$. Such logarithmic sampling is standard in kernel ridge regression because both λ and γ exhibit sensitivity on a multiplicative rather than an additive scale; a linear grid would either oversample uninformative large values or miss critical small values. The above ranges are consistent with the tuning strategies adopted in the distributed kernel ridge regression literature. Specifically, Zhang et al. (2015) and Xu et al. (2019) recommend scanning λ over $[10^{-4}, 1]$ and the Gaussian RBF bandwidth over $[10^{-2}, 10^{1}]$ for standardized predictors. The optimal pair (λ, γ) is chosen by minimizing the MSE criterion described in Section 3.2. Table 7 summarizes the index vector error D, the prediction RMSE, and the computational time.

The proposed distributed SDR methods uniformly outperform IDC in both direction estimation and prediction accuracy. Moreover, AdaDKRR without dimension reduction produces the highest RMSE, confirming that reducing the predictor to a single index effectively mitigates the curse of dimensionality. In terms of computational cost, the one-shot aggregation of our approach is substantially lower than that of the iterative IDC method, demonstrating the

Table 7: Comparison of distributed SDR with IDC and AdaDKRR (Simulation IV).

Model	Metric	m	Proposed			IDC	AdaDKRR
			DC-SIR	DC-PSVM	DC-DCOV		
3	D	1	0.0148	0.0086	0.0094	0.0090	–
		10	0.0148	0.0096	0.0101	0.0356	–
	RMSE	1	0.1004	0.1001	0.1002	0.1313	0.1024
		10	0.1014	0.1012	0.1012	0.1215	0.1039
	Time	1	74.1179	78.8089	82.0850	152.0221	17.3136
		10	1.6118	1.7211	1.9509	2.6362	1.1118
4	D	1	0.0145	0.0099	0.0124	0.0344	–
		10	0.0151	0.0105	0.0133	0.1210	–
	RMSE	1	0.1003	0.1002	0.1003	0.1262	0.1209
		10	0.1017	0.1016	0.1017	0.1485	0.1231
	Time	1	68.3882	69.1502	71.5836	106.8917	17.3914
		10	1.6376	1.7511	1.9969	2.4941	1.1233

communication efficiency promised by Theorem 4.1.

5.5 Real Data Analysis: Airline On-Time Performance

We illustrate the practical utility of our method on the airline on-time performance data from the 2009 ASA Data Expo (http://stat-computing.org/dataexpo/2009/the-data.html). The dataset comprises flight records from October 1987 to April 2008. We focus on the year 2008, which contains 2,389,217 observations and 29 variables. After removing 70,096 incomplete cases (2.93%), 2,319,121 records remain. We use the first 2,310,000 observations for estimation and the remaining 9,121 for out-of-sample prediction, distributing the estimation sample across $m =$ 10,000 machines.

We predict the arrival delay (Y, in minutes) using three continuous predictors: departure delay (X_1), departure time (X_2, scaled to $[0, 24]$), and flight distance (X_3, in thousands of miles). Let $\boldsymbol{X} = (X_1, X_2, X_3)^\top$. We compare three modeling strategies:

(i) **LS:** The linear model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$ fitted by ordinary least squares using the full centralized data as a baseline.

(ii) **KRR:** Nonparametric kernel ridge regression $Y = f(X_1, X_2, X_3) + \varepsilon$ without dimension reduction.

(iii) **Proposed:** The single-index model $Y = f(b^\top \boldsymbol{X}) + \varepsilon$ estimated by distributed SDR (SIR, PSVM, or DCOV) followed by adaptive DKRR.

Table 8 reports the prediction RMSE and the total computational time, averaged over 100 random splits of the data.

The results in Table 8 lead to several important observations. First, all three distributed single-index methods achieve substantially lower RMSE than both the linear model and the full nonparametric KRR. The linear model fails to capture the pronounced nonlinearities inherent in flight delay dynamics, whereas the distributed single-index models reduce the prediction error.

Table 8: Prediction RMSE and computational time on the airline on-time data.

	DC-SIR	DC-PSVM	DC-DCOV	DC-KRR	LS
RMSE	0.2667	0.2194	0.2177	3.3813	1.3716
Time (sec)	1001.4864	1253.2445	1272.8289	931.3594	0.0093

This confirms that the single-index structure effectively encodes the relevant nonlinear features while avoiding the curse of dimensionality.

Second, KRR without dimension reduction exhibits the worst predictive performance, even though it uses the same adaptive distributed kernel ridge regression machinery. The severe deterioration stems from the fact that, despite the large total sample size ($N \approx 2.3 \times 10^6$), nonparametric regression in three dimensions requires an exponentially larger effective sample size to achieve comparable accuracy. By projecting the three-dimensional predictor onto a single direction, our framework transforms the estimation problem into a univariate nonparametric regression, where the adaptive DKRR attains near-oracle performance.

Third, among the distributed SDR variants, DC-DCOV and DC-PSVM yield almost identical RMSE, both slightly outperforming DC-SIR. This is consistent with the simulation findings: DCOV and PSVM rely on weaker structural assumptions and are more robust to departures from the linearity condition that may be present in this real dataset. The small margin between them suggests that the underlying dependence structure is adequately captured by all three methods, with the model-free DC-DCOV providing a safe default choice in the absence of prior knowledge about the predictor distribution. Finally, the computational times show that the distributed SDR methods require approximately 1000–1300 seconds, which is only marginally higher than the 931 seconds consumed by the distributed KRR without dimension reduction. The additional overhead is entirely attributable to the initial distributed SDR step (eigen-decompositions or quadratic programming on local machines), while the subsequent adaptive DKRR stage is shared by all nonparametric variants. Crucially, with $m = 10000$ machines, the communication load remains bounded by $O(mp + mnL)$, which is negligible relative to the size of the raw data. This demonstrates that the proposed framework scales gracefully to massive distributed environments, delivering substantial gains in predictive accuracy at an acceptable and predictable computational cost.

6 Conclusions

In this paper, we have proposed a communication-efficient one-shot distributed estimation framework for single-index models based on sufficient dimension reduction. By leveraging the geometric structure inherent in SDR estimation, our method transmits only p-dimensional projection directions from local machines to the central server, where sample-size-weighted spectral aggregation on the Grassmann manifold yields a $\sqrt{N}$-consistent estimator of the index vector. The subsequent adaptive distributed kernel ridge regression step estimates the link function at the optimal univariate nonparametric rate, effectively circumventing the curse of dimensionality. Theoretical guarantees in terms of convergence rates and asymptotic normality have been established, and extensive numerical studies confirm that the proposed approach achieves sta-

tistical accuracy comparable to centralized methods while drastically reducing computational and communication burdens.

A distinctive feature of our analysis is that it accommodates heterogeneous partitions without requiring $n_k = N/m$. The sample-size-weighted spectral aggregation eliminates the weakest-machine bottleneck, yielding the universal direction rate $O_p(N_{-k}^{-1/2})$. By decoupling the parametric and nonparametric stages via cross-fitting, we establish, for the first time in this literature, a complete theoretical chain: $\sqrt{N}$-consistency and asymptotic normality for the direction (Theorem 4.1), followed by the minimax optimal rate $N^{-\alpha/(2\alpha+1)}$ for the distributed link function estimator under adaptive parameter selection (Theorem 4.2). This two-stage optimality rigorously justifies the use of distributed single-index models as a communication-efficient alternative to fully nonparametric distributed regression, whose rate $N^{-\alpha/(2\alpha+p)}$ would otherwise deteriorate rapidly with the predictor dimension p.

Furthermore, the adaptive parameter selection mechanism, built upon the AdaDKRR principle of Lin et al. (2025), resolves the over-fitting problem of purely local tuning by allowing each machine to validate against a globally synthesized basis approximation. This design requires only two rounds of communication of non-sensitive basis coefficients, thereby circumventing data-silo constraints while preserving the oracle statistical efficiency of the full-sample kernel ridge regression. The leave-one-out cross-fitting variant eliminates the statistical dependence between the direction estimator and the local regression data, removing the extraneous logarithmic factors that typically arise from data reuse in coupled two-stage analyses.

Beyond this theoretical completeness, the proposed framework offers distinct advantages in communication architecture and geometric aggregation. The IDC method of Jiang and Peng (2023) requires multiple rounds of refinement, incurring synchronization delays that scale with the iteration count. The DC-rOPG method of Li and Yang (2025), while one-shot in aggregation, relies on simple Euclidean averaging of local index vectors and is therefore exposed to sign ambiguity: if local machines report opposite signs, the average collapses toward the origin. Our spectral aggregation on the Grassmann manifold via $\widehat{\boldsymbol{M}}_{[-k]}$ intrinsically eliminates sign dependence and achieves consensus through the dominant mode of variation. This geometric perspective not only yields a cleaner theoretical analysis (via Davis–Kahan and rank-one perturbation) but also removes the need for ad-hoc sign-alignment pre-processing across local machines.

Finally, the proposed framework provides greater modeling flexibility and robustness. The DC-rOPG mechanism is tied to the OPG methodology, which requires the linearity condition and can fail when $\mathbb{E}[f'(b^\top \boldsymbol{X})] = 0$ or when the predictor distribution deviates from elliptical symmetry. Our unified SDR treatment accommodates moment-based SIR, machine-learning-enhanced PSVM, and model-free DCOV within a single distributed protocol. As demonstrated in our numerical studies, the DCOV estimator retains robustness under skew-normal predictors where OPG-based methods suffer substantial bias, because distance covariance measures dependence without imposing linearity or parametric link specifications.

Although the current exposition focuses on the single-index model ($d = 1$), the extension to multiple-index models ($d > 1$) is immediate and straightforward. For a general sufficient dimension reduction problem with structural dimension d, each local machine simply estimates the

d-dimensional central subspace via any of the three SDR paradigms reviewed in Section 2. The central server then collects the local basis matrices $\widehat{\boldsymbol{B}}_k \in \mathbb{R}^{p\times d}$ and forms the aggregate matrix $\widehat{\boldsymbol{B}} = (\widehat{\boldsymbol{B}}_1, \ldots, \widehat{\boldsymbol{B}}_m) \in \mathbb{R}^{p\times(md)}$. The final estimator is obtained as the d leading eigenvectors of $\widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^\top$, which span the estimated central subspace. The distributed kernel ridge regression step is similarly adapted by projecting the covariates onto the d-dimensional subspace and employing multivariate kernel smoothing. All theoretical derivations in Section 4 carry over with minor modifications, preserving the optimal convergence rates under general partition heterogeneity. The condition $n_0 \to \infty$ remains the sole requirement for consistency, ensuring that the extension is valid for both balanced and unbalanced distributed architectures.

Limitations and future directions. The current framework presupposes a homogeneous, non-adversarial environment (Assumption 4.8). Extending it to heterogeneous (Non-IID) partitions and Byzantine-robust spectral aggregation—e.g., via median-of-means or trimmed eigenvector procedures—is a critical next step. Other avenues include nonlinear SDR, high-dimensional settings with p growing with N, and adaptive tuning that accounts for local heterogeneity in the basis approximation.

References

Rajendra Bhatia. *Matrix analysis*. Springer Science & Business Media, 2013.

Peter J Bickel. On adaptive estimation. *The Annals of Statistics*, 10(3):647–671, 1982.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, Oxford, UK, 2013.

Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational mathematics*, 7(3):331–368, 2007.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

R Dennis Cook. *Regression graphics: Ideas for studying regressions through graphics*. John Wiley & Sons, 2009.

R Dennis Cook and Sanford Weisberg. Discussion of sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):328–332, 1991.

Josef Dick and Friedrich Pillichshammer. *Digital nets and sequences: discrepancy theory and quasi–Monte Carlo integration*. Cambridge University Press, 2010.

Rong Jiang and Yexun Peng. A short note on fitting a single-index model with massive data. *Statistical Theory and Related Fields*, 7(1):49–60, 2023.

Rong Jiang, Meng-Fan Guo, and Xin Liu. Composite quasi-likelihood for single-index models with massive datasets. *Communications in Statistics-Simulation and Computation*, 51(9): 5024–5040, 2022.

Gunther Leobacher and Friedrich Pillichshammer. *Introduction to quasi-Monte Carlo integration and applications*. Springer, 2014.

Bing Li and Shaoli Wang. On directional regression for dimension reduction. *Journal of the American Statistical Association*, 102(479):997–1008, 2007.

Bing Li, Andreas Artemiou, and Lexin Li. Principal support vector machines for linear and nonlinear sufficient dimension reduction. *The Annals of Statistics*, 39(6):3182–3210, 2011.

Ker-Chau Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327, 1991.

Na Li and Jing Yang. A fast divide-and-conquer strategy for single-index model with massive data. *Computational Statistics*, 40(5):2367–2387, 2025.

Shao-Bo Lin, Xiaotong Liu, Di Wang, Hai Zhang, and Ding-Xuan Zhou. Adaptive distributed kernel ridge regression: A feasible distributed learning scheme for data silos. *Journal of Machine Learning Research*, 26(108):1–54, 2025.

Whitney K Newey and James R Robins. Cross-fitting and fast remainder rates for semiparametric estimation. *arXiv preprint arXiv:1801.09138*, 2018.

Anton Schick. On asymptotically efficient estimation in semiparametric models. *The Annals of Statistics*, 14(3):1139–1151, 1986.

Wenhui Sheng and Xiangrong Yin. Direction estimation in single-index models via distance covariance. *Journal of Multivariate Analysis*, 122:148–161, 2013.

Wenhui Sheng and Xiangrong Yin. Sufficient dimension reduction via distance covariance. *Journal of Computational and Graphical Statistics*, 25(1):91–104, 2016.

Charles J. Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10(4):1040–1053, 1982.

Gábor J. Székely, Maria L. Rizzo, and Nail K. Bakirov. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769 – 2794, 2007.

Joel A Tropp. An introduction to matrix concentration inequalities. *Foundations and trends® in machine learning*, 8(1-2):1–230, 2015.

Grace Wahba. *Spline models for observational data*. SIAM, 1990.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Yingcun Xia. A constructive approach to the estimation of dimension reduction directions. *The Annals of Statistics*, 35(6):2654–2690, 2007.

Yingcun Xia, Howell Tong, Wai Keung Li, and Li-Xing Zhu. An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 64 (3):363–410, 2002.

Ganggang Xu, Zuofeng Shang, and Guang Cheng. Distributed generalized cross-validation for divide-and-conquer kernel ridge regression and its asymptotic optimality. *Journal of computational and graphical statistics*, 28(4):891–908, 2019.

Yi Yu, Tengyao Wang, and Richard J Samworth. A useful variant of the davis–kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015.

Yuchen Zhang, John Duchi, and Martin Wainwright. Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. *The Journal of Machine Learning Research*, 16(1):3299–3340, 2015.

Li-Ping Zhu, Li-Xing Zhu, and Zheng-Hui Feng. Dimension reduction in regressions through cumulative slicing estimation. *Journal of the American Statistical Association*, 105(492):1455–1466, 2010.

A Technical Proofs

This appendix provides the detailed proofs for the theoretical results presented in Section 4. Throughout the proofs, C, C', c_1, c_2 denote generic absolute constants that may change from line to line. We use the notation $a_n \lesssim b_n$ to denote that $a_n \le C b_n$ for some constant $C > 0$ independent of n.

A.1 Proof of Theorem 4.1

Proof. Let $N_{-k} = \sum_{j \neq k} n_j$ be the total sample size among the $m-1$ machines used to compute the leave-one-out direction $\widehat{b}_{[-k]}$. Let $\widehat{\boldsymbol{M}}_{[-k]} = \sum_{j \neq k} \frac{n_j}{N_{-k}} \widehat{b}_j \widehat{b}_j^\top$ denote the weighted aggregate matrix, and $\boldsymbol{M}_{[-k]} = bb^\top$ be its population counterpart.

Step 1: Concentration of the weighted aggregate matrix. Write the aggregate perturbation as

$$\boldsymbol{E} \;=\; \widehat{\boldsymbol{M}}_{[-k]} - \boldsymbol{M}_{[-k]} \;=\; \sum_{j \neq k} \frac{n_j}{N_{-k}} \big(\widehat{b}_j \widehat{b}_j^\top - bb^\top\big).$$

Decompose each summand into a zero-mean random fluctuation and a deterministic bias:

$$\frac{n_j}{N_{-k}} \big(\widehat{b}_j \widehat{b}_j^\top - bb^\top\big) \;=\; \underbrace{\frac{n_j}{N_{-k}} \big(\widehat{b}_j \widehat{b}_j^\top - \mathbb{E}[\widehat{b}_j \widehat{b}_j^\top]\big)}_{=:\,\boldsymbol{A}_j} \;+\; \underbrace{\frac{n_j}{N_{-k}} \big(\mathbb{E}[\widehat{b}_j \widehat{b}_j^\top] - bb^\top\big)}_{=:\,\boldsymbol{B}_j}. \tag{17}$$

By the $\sqrt{n_j}$-consistency of the local SDR estimators (Section 2), each local estimator admits the stochastic expansion $\widehat{b}_j - b = \boldsymbol{\xi}_j / \sqrt{n_j} + O_p(n_j^{-1})$ with $\mathbb{E}[\boldsymbol{\xi}_j] = \mathbf{0}$ and $\mathbb{E}\|\boldsymbol{\xi}_j\|^2 = O(1)$. Consequently, the bias satisfies $\|\boldsymbol{B}_j\|_2 = O(1/N_{-k})$, whence

$$\Big\| \sum_{j \neq k} \boldsymbol{B}_j \Big\|_2 \;\le\; \sum_{j \neq k} \|\boldsymbol{B}_j\|_2 \;=\; O\Big(\frac{m-1}{N_{-k}}\Big). \tag{18}$$

The random matrices $\{\boldsymbol{A}_j\}_{j\neq k}$ are independent, centered ($\mathbb{E}[\boldsymbol{A}_j] = \boldsymbol{0}$), and satisfy $\|\boldsymbol{A}_j\|_2 \leq 4n_j/N_{-k}$ almost surely as well as $\mathbb{E}\|\boldsymbol{A}_j\|_2^2 = O(n_j/N_{-k}^2)$. Applying Theorem 1.6.2 of Tropp (2015) to the sum $\sum_{j\neq k} \boldsymbol{A}_j$, the matrix variance statistic

$$v(\boldsymbol{A}) \;=\; \Big\|\sum_{j\neq k} \mathbb{E}[\boldsymbol{A}_j^2]\Big\|_2 \;=\; O\Big(\sum_{j\neq k} \frac{n_j}{N_{-k}^2}\Big) \;=\; O\Big(\frac{1}{N_{-k}}\Big)$$

yields the concentration bound

$$\Big\|\sum_{j\neq k} \boldsymbol{A}_j\Big\|_2 \;=\; O_p\Big(\sqrt{\frac{1}{N_{-k}}}\Big), \tag{19}$$

where the dimensional factor $\log(2p)$ has been absorbed into the O_p notation because p is fixed.

Combining (18) and (19), the total perturbation satisfies the general bound

$$\|\boldsymbol{E}\|_2 \;=\; O_p\Big(\frac{1}{\sqrt{N_{-k}}} \;+\; \frac{m-1}{N_{-k}}\Big). \tag{20}$$

This bound holds universally for both fixed and diverging m without any additional restriction beyond those stated in Theorem 4.1. The relative magnitude of the two terms will be analyzed separately under each regime in the steps below.

Step 2: Eigenvector perturbation and consistency. By Assumption 4.4, the spectral gap of the rank-one population matrix $\boldsymbol{M}_{[-k]}$ is $\lambda_1(\boldsymbol{M}_{[-k]}) - \lambda_2(\boldsymbol{M}_{[-k]}) = 1$. Applying the Davis–Kahan $\sin\Theta$ theorem in the spectral norm (Yu et al., 2015), we first bound the rank-one projection operator:

$$\|\widehat{b}_{[-k]}\widehat{b}_{[-k]}^\top - bb^\top\|_2 \;\leq\; \frac{\|\boldsymbol{E}\|_2}{\lambda_1(\boldsymbol{M}_{[-k]}) - \lambda_2(\boldsymbol{M}_{[-k]})} \;=\; \|\boldsymbol{E}\|_2. \tag{21}$$

Since both $\widehat{b}_{[-k]}$ and b are unit vectors and Assumption 4.4 constrains their first nonzero coordinate to be positive (so that $\widehat{b}_{[-k]}^\top b \geq 0$ and the principal angle lies in $[0, \pi/2]$), the vector ℓ_2-distance is controlled by the projection error via the elementary inequality $\|\widehat{b}_{[-k]} - b\| \leq \sqrt{2}\,\|\widehat{b}_{[-k]}\widehat{b}_{[-k]}^\top - bb^\top\|_2$. Multiplying (21) by $\sqrt{2}$ and substituting (20) yields

$$\|\widehat{b}_{[-k]} - b\| \;=\; O_p\Big(\frac{1}{\sqrt{N_{-k}}} \;+\; \frac{m-1}{N_{-k}}\Big). \tag{22}$$

The asymptotic behavior under the two canonical regimes is as follows.

Step 3: First-order linearization with a sharp remainder. To obtain the asymptotic linear representation required for the CLT, we use the perturbation expansion for the leading eigenvector of a rank-one symmetric matrix (Bhatia, 2013). Because $\boldsymbol{M}_{[-k]} = bb^\top$ has spectral gap $\delta = 1$, the eigenvector map $\boldsymbol{M} \mapsto v_1(\boldsymbol{M})$ has Lipschitz constant $1/\delta$ and its second-order

Fréchet derivative is bounded by C/δ^2 in a neighbourhood of $\boldsymbol{M}_{[-k]}$. Hence

$$\widehat{b}_{[-k]} - b \;=\; (\boldsymbol{I}_p - bb^\top)\boldsymbol{E}b \;+\; \boldsymbol{R}, \qquad \|\boldsymbol{R}\| \;=\; O_p\Big(\|\boldsymbol{E}\|_2^2\Big) \;=\; O_p\Big(\frac{1}{N_{-k}}\Big). \tag{23}$$

For the linear term, observe from (17) that

$$\boldsymbol{E}b \;=\; \sum_{j\neq k} \frac{n_j}{N_{-k}} (\widehat{b}_j\widehat{b}_j^\top - bb^\top)b \;=\; \sum_{j\neq k} \frac{n_j}{N_{-k}} \big[\widehat{b}_j(\widehat{b}_j^\top b) - b\big].$$

Because $\|\widehat{b}_j\| = \|b\| = 1$, we have $\widehat{b}_j^\top b = 1 - \frac{1}{2}\|\widehat{b}_j - b\|^2$, which implies

$$\widehat{b}_j(\widehat{b}_j^\top b) - b \;=\; (\widehat{b}_j - b) \;-\; \frac{1}{2}\|\widehat{b}_j - b\|^2\widehat{b}_j \;=\; (\widehat{b}_j - b) \;+\; O_p(n_j^{-1}).$$

Projecting onto the orthogonal complement of b and noting that $(\boldsymbol{I}_p - bb^\top)(\widehat{b}_j - b) = \widehat{b}_j - b + O_p(n_j^{-1})$ (because $b^\top(\widehat{b}_j - b) = -\frac{1}{2}\|\widehat{b}_j - b\|^2 = O_p(n_j^{-1})$), we obtain from (23) the clean linearization

$$\widehat{b}_{[-k]} - b \;=\; \sum_{j\neq k} \frac{n_j}{N_{-k}} (\widehat{b}_j - b) \;+\; \boldsymbol{R}', \qquad \|\boldsymbol{R}'\| \;=\; \|\boldsymbol{R}\| \;+\; O_p\Big(\frac{1}{N_{-k}}\Big) \;=\; O_p\Big(\frac{1}{N_{-k}}\Big). \tag{24}$$

It remains to show that $\sqrt{N}\,\|\boldsymbol{R}'\| = o_p(1)$ under the theorem assumptions. Since $N_{-k} \asymp N$ in both the fixed-m and the balanced diverging-m regimes, $\sqrt{N}\cdot O_p(1/N_{-k}) = O_p(1/\sqrt{N}) = o_p(1)$. Thus $\sqrt{N}\,\|\boldsymbol{R}'\| = o_p(1)$ for both regimes.

Step 4: Asymptotic normality under fixed m. When m is fixed and $\max_{j\neq k} N/n_j = O(1)$, scaling (24) by $\sqrt{N}$ yields

$$\sqrt{N}(\widehat{b}_{[-k]} - b) \;=\; \sum_{j\neq k} \sqrt{\frac{N\,n_j}{N_{-k}^2}} \cdot \sqrt{n_j}(\widehat{b}_j - b) \;+\; o_p(1). \tag{25}$$

Since $N_{-k} \asymp N$, the remainder is $o_p(1)$ as shown in Step 3. By the $\sqrt{n_j}$-asymptotic normality of each local estimator, $\sqrt{n_j}(\widehat{b}_j - b) \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_j)$ independently across j. The finite weighted sum of independent Gaussian vectors is Gaussian, yielding (14).

Step 5: Asymptotic normality under diverging m. Under the diverging-m regime with $n_j \asymp N/m$ (balanced in order), $N_{-k} \asymp N$ and scaling (24) by $\sqrt{N}$ gives

$$\sqrt{N}(\widehat{b}_{[-k]} - b) \;=\; \sum_{j\neq k} \sqrt{\frac{N\,n_j}{N_{-k}^2}} \cdot \sqrt{n_j}(\widehat{b}_j - b) \;+\; o_p(1). \tag{26}$$

If, in addition, $\sum_{j\neq k}(n_j/N)^2 \to 0$, the weights $w_j = \sqrt{N\,n_j}/N_{-k}$ satisfy the Lyapunov condition for the triangular-array CLT:

$$\sum_{j\neq k} w_j^4\,\mathbb{E}\|\sqrt{n_j}(\widehat{b}_j - b)\|^4 \;=\; O\left(\sum_{j\neq k} \frac{N^2 n_j^2}{N_{-k}^4}\right) \;=\; O\left(\sum_{j\neq k} \Big(\frac{N\,n_j}{N_{-k}^2}\Big)^2\right) \;\to\; 0.$$

Hence the Lindeberg–Feller CLT applies to the weighted sum in (26), and (15) follows. □

A.2 Proof of Theorem 4.2

Proof. We employ a proof by decomposition that explicitly tracks the three sources of error: (i) the oracle nonparametric error from distributed KRR with known direction and optimal parameter; (ii) the basis approximation error induced by the AdaDKRR coefficient synthesis; and (iii) the projection error from using the estimated leave-one-out directions.

Let $\bar{\lambda} \asymp N^{-\frac{2\alpha}{2\alpha+1}}$ be the oracle optimal parameter contained in Λ (Assumption 4.7). Let $\widetilde{f}_{k,\bar{\lambda}}$ denote the oracle local KRR estimator on machine k using the *true* projected covariates $Z_i^* = b^\top \boldsymbol{X}_i$ and the oracle parameter $\bar{\lambda}$. We decompose the global L^2 error as:

$$\begin{aligned} \|\widehat{f} - f_0\|_{L^2} &= \Big\|\frac{1}{m}\sum_{k=1}^{m} \widehat{f}^*_{k,[-k]} - f_0\Big\|_{L^2} \\ &\le \underbrace{\Big\|\frac{1}{m}\sum_{k=1}^{m}(\widehat{f}^*_{k,[-k]} - \widetilde{f}_{k,\bar{\lambda}})\Big\|_{L^2}}_{I_1\text{:Adaptation \& Projection Error}} + \underbrace{\Big\|\frac{1}{m}\sum_{k=1}^{m}\widetilde{f}_{k,\bar{\lambda}} - f_0\Big\|_{L^2}}_{I_2\text{:Oracle Nonparametric Error}} . \end{aligned} \tag{27}$$

Here, $\widehat{f}^*_{k,[-k]}$ is evaluated at the estimated projections $Z_i = \widehat{b}_{[-k]}^\top \boldsymbol{X}_i$ with adaptively selected parameter $\widehat{\lambda}^*_k$, whereas $\widetilde{f}_{k,\bar{\lambda}}$ is evaluated at the true projections $Z_i^* = b^\top \boldsymbol{X}_i$ with the oracle parameter $\bar{\lambda}$.

Step 1: Bounding the Oracle Nonparametric Error (I_2). The term I_2 represents the error of a distributed KRR estimator if the true direction b and the optimal oracle parameter $\bar{\lambda}$ were known perfectly. Since the local estimators $\widetilde{f}_{k,\bar{\lambda}}$ are trained on independent subsets S_k with a common parameter $\bar{\lambda}$, and the theorem assumes $m \le N^{\frac{2\alpha-1}{4\alpha+2}} \log^{-4} N$, the conditions of Theorem 1 in Zhang et al. (2015) are satisfied. Therefore, the distributed estimator achieves the minimax optimal rate:

$$I_2 = O_p\Big(N^{-\frac{\alpha}{2\alpha+1}}\Big) . \tag{28}$$

Step 2: Decomposing the Adaptation and Projection Error (I_1). For each machine k, we further decompose the local discrepancy:

$$\widehat{f}^*_{k,[-k]} - \widetilde{f}_{k,\bar{\lambda}} = \underbrace{(\widehat{f}^*_{k,[-k]} - \widehat{f}_{k,[-k],\bar{\lambda}})}_{\text{Parameter adaptation error}} + \underbrace{(\widehat{f}_{k,[-k],\bar{\lambda}} - \widetilde{f}_{k,\bar{\lambda}})}_{\text{Projection error}}, \tag{29}$$

where $\widehat{f}_{k,[-k],\bar{\lambda}}$ is the local KRR estimator using the estimated direction $\widehat{b}_{[-k]}$ but with the oracle parameter $\bar{\lambda}$.

Controlling the projection error. Conditional on $\widehat{b}_{[-k]}$, the data S_k are independent and identically distributed. By Assumption 4.5, the pointwise drift of the target function is bounded by:

$$|f_0(\widehat{b}_{[-k]}^\top \boldsymbol{x}) - f_0(b^\top \boldsymbol{x})| \le L_f \|\widehat{b}_{[-k]} - b\| \cdot \|\boldsymbol{x}\|. \tag{30}$$

Taking the L^2 norm with respect to the marginal distribution of $\boldsymbol{X}$ and applying Theorem 4.1, we obtain:

$$\|f_0(\widehat{b}_{[-k]}^\top \cdot) - f_0(b^\top \cdot)\|_{L^2(\rho)} \leq L_f \|\widehat{b}_{[-k]} - b\| \cdot \mathbb{E}[\|\boldsymbol{X}\|^2]^{1/2} = O_p(N_{-k}^{-1/2}), \tag{31}$$

where the finiteness of $\mathbb{E}[\|\boldsymbol{X}\|^2]$ is guaranteed by Assumption 4.1. This controls the propagation of the parametric error into the L^2 prediction risk *without* incurring any $\sqrt{\log N}$ covering-number penalties, precisely because the leave-one-out design eliminates the need for uniform oracle inequalities over shrinking neighborhoods of b.

By the standard stability of kernel ridge regression (see, e.g., Caponnetto and De Vito 2007), the perturbation in the target function translates to the estimator as:

$$\|\widehat{f}_{k,[-k],\bar{\lambda}} - \widetilde{f}_{k,\bar{\lambda}}\|_{L^2} = O_p(\|f_0(\widehat{b}_{[-k]}^\top \cdot) - f_0(b^\top \cdot)\|_{L^2}) = O_p(N_{-k}^{-1/2}). \tag{32}$$

Controlling the parameter adaptation error. The selected parameter $\widehat{\lambda}_k^*$ minimizes the local validation MSE in (8), which is evaluated against the globally synthesized basis approximation $\widehat{f}_{\lambda,n,\mu}^{\text{glob}}$. To relate this to the oracle risk, we decompose the validation target:

$$\widehat{f}_{\lambda,n,\mu}^{\text{glob}} = \bar{f}_\lambda^{\text{oracle}} + (\widehat{f}_{\lambda,n,\mu}^{\text{glob}} - \bar{f}_\lambda^{\text{oracle}}), \tag{33}$$

where $\bar{f}_\lambda^{\text{oracle}} = \sum_{k=1}^m \frac{n_k}{N} \widehat{f}_{k,[-k],\lambda}$ is the oracle distributed KRR estimator with parameter λ (without basis approximation). By Assumption 4.6 and Assumption 4.7, and following Proposition 17 of Lin et al. (2025), the basis approximation error satisfies:

$$\|\widehat{f}_{\lambda,n,\mu}^{\text{glob}} - \bar{f}_\lambda^{\text{oracle}}\|_{L^\infty} \leq C\Big(\mu^\alpha + \mu^{1/2}\lambda^{-1/2}n^{-\beta/2}\Big), \tag{34}$$

for all $\lambda \in \Lambda$, where C depends on $\kappa, M, \alpha, c, \beta$. Under the conditions $\mu \leq N^{-\frac{2\alpha}{2\alpha+1}}$ and $\mu n^\beta \geq 2c$, the right-hand side of (34) is $O(N^{-\frac{\alpha}{2\alpha+1}})$, which is of the same order as the minimax optimal rate and thus negligible for parameter selection.

Because the candidate grid Λ is finite, the empirical process $\{\text{MSE}_k(\lambda) - \mathbb{E}[\text{MSE}_k(\lambda)]\}_{\lambda\in\Lambda}$ indexed by the regularization parameter constitutes a finite function class. For any fixed λ, the truncated squared errors $(\pi_M \widehat{f}_{\lambda,n,\mu}^{\text{glob}}(Z_i) - Y_i)^2$ are bounded random variables by Assumption 4.3, so that Bernstein's inequality applies to the sample mean $\text{MSE}_k(\lambda)$ (Boucheron et al., 2013, Theorem 2.10). Aggregating these single-point bounds over the $L = \#\Lambda$ candidates via a union bound (Boucheron et al., 2013, Theorem 2.10) and (Wainwright, 2019, Chapter 14) yields the uniform deviation control

$$\sup_{\lambda\in\Lambda} |\text{MSE}_k(\lambda) - \mathbb{E}[\text{MSE}_k(\lambda)]| \leq C\sqrt{\frac{\log L + \log(2/\delta)}{n_k^{\text{val}}}}, \tag{35}$$

with probability at least $1 - \delta$, where $C > 0$ is an absolute constant depending only on the uniform bound M (see also Lin et al., 2025, Proposition 17 for an analogous treatment in the adaptive distributed KRR context).

Choosing $\delta = N^{-1}$ and noting that $n_k^{\text{val}} \asymp n_k$ (since the train-validation split is balanced), the uniform deviation on machine k is $O_p(\sqrt{\log L/n_k})$. Under the theorem assumption $m \leq$

$N^{\frac{1}{2\alpha+1}} \log^{-1} L$, the average sample size satisfies $N/m \geq N^{\frac{2\alpha}{2\alpha+1}} \log L$. Hence, in the regime where local sample sizes are balanced in order ($n_k \asymp N/m$), we obtain:

$$\sqrt{\frac{\log L}{n_k}} \lesssim \sqrt{\frac{\log L}{N^{\frac{2\alpha}{2\alpha+1}} \log L}} = N^{-\frac{\alpha}{2\alpha+1}}. \tag{36}$$

Therefore, the uniform concentration error is dominated by the minimax rate. Combining this with the basis approximation bound (34), the adaptively selected parameter $\widehat{\lambda}_k^*$ satisfies:

$$\|\widehat{f}^*_{k,[-k]} - \widehat{f}_{k,[-k],\bar{\lambda}}\|_{L^2} = O_p\left(\sqrt{\frac{\log L}{n_k}} + N^{-\frac{\alpha}{2\alpha+1}}\right) = O_p\left(N^{-\frac{\alpha}{2\alpha+1}}\right). \tag{37}$$

Controlling the grid discretization error. Because the candidate grid Λ is logarithmic equispaced with cardinality L, the relative spacing between adjacent grid points is $O(1/L)$. By the Lipschitz continuity of the KRR objective and the validation risk with respect to $\log \lambda$ on the compact interval $[N^{-\frac{2\alpha}{2\alpha+1}-\Delta}, N^{-\frac{2\alpha}{2\alpha+1}+\Delta}]$, the oracle parameter $\bar{\lambda}$ and its nearest grid point $\lambda^\dagger \in \Lambda$ induce a risk difference of order

$$\|\widehat{f}_{k,[-k],\lambda^\dagger} - \widehat{f}_{k,[-k],\bar{\lambda}}\|_{L^2} = O_p\left(\frac{N^{-\frac{\alpha}{2\alpha+1}}}{L}\right),$$

and the same bound carries over to the adaptively selected $\widehat{\lambda}_k^*$. Choosing $L \asymp \log N$ guarantees $N^{-\frac{\alpha}{2\alpha+1}}/L = o(N^{-\frac{\alpha}{2\alpha+1}})$, so the discretization error is asymptotically negligible.

Synthesizing the local bounds. Combining the parameter adaptation error and the projection error in (29), we obtain for each k:

$$\|\widehat{f}^*_{k,[-k]} - \widetilde{f}_{k,\bar{\lambda}}\|_{L^2} = O_p\left(N_{-k}^{-1/2} + N^{-\frac{\alpha}{2\alpha+1}}\right). \tag{38}$$

Averaging over m machines and noting that $N_{-k} \geq N_- = \min_{1\leq k\leq m} N_{-k}$ for all k, we have:

$$I_1 \leq \frac{1}{m}\sum_{k=1}^{m} \|\widehat{f}^*_{k,[-k]} - \widetilde{f}_{k,\bar{\lambda}}\|_{L^2} \leq \max_{1\leq k\leq m} \|\widehat{f}^*_{k,[-k]} - \widetilde{f}_{k,\bar{\lambda}}\|_{L^2} = O_p\left(N_-^{-1/2} + N^{-\frac{\alpha}{2\alpha+1}}\right). \tag{39}$$

Step 3: Synthesis of Final Rates. Substituting (28) and (39) back into the decomposition (27) yields the general bound:

$$\|\widehat{f} - f_0\|_{L^2} = O_p\left(N^{-\frac{\alpha}{2\alpha+1}} + N_-^{-\frac{1}{2}} + L^{-1}\right). \tag{40}$$

Under the fixed-m regime with $N_- \asymp N$, the projection error scales as $N^{-1/2}$, which is dominated by, or of the same order as, the nonparametric oracle error for $\alpha \geq 1/2$. Under the diverging-m regime with $n_k \asymp N/m$ (balanced in order), we still have $N_- \asymp N$ and the projection error remains $N^{-1/2}$. Provided $m \leq N^{\frac{1}{2\alpha+1}} \log^{-1} L$ (which in particular implies $m = o(N^{1/(2\alpha+1)})$), the projection error is negligible compared with the minimax rate. The grid discretization error L^{-1} with $L \asymp \log N$ is likewise asymptotically negligible.

This completes the proof. $\square$