Universal Distributed Feature Extraction Complementary to Side Information via H-Score

Yuan Xinjie
Tsinghua Shenzhen International Graduate School, Tsinghua University
Shenzhen, China
yuanxj23@mails.tsinghua.edu.cn

Dier Tang
Department of Mathematics The University of Hong Kong
Hong Kong, China
tangde_math@connect.hku.hk

Shao-Lun Huang
Tsinghua Shenzhen International Graduate School, Tsinghua University
Shenzhen, China
twn2gold@gmail.com

***Abstract*—This paper considers a distributed estimation system where edge nodes transmit averaged features to a center node for parameter estimation. A key challenge arises when the center node possesses auxiliary side information: how to design local extractors that are universally optimal for any downstream task, while avoiding redundancy with the side information. To tackle this problem, we orthogonally decompose the joint function space and demonstrate in the weak-dependence regime that the modal decomposition of the canonical dependence kernel's conditional-additive component yields optimal, task-independent features. We construct a distributed nested H-score to learn this decomposition and show that, under a limited feature budget, the mean-squared error (MSE) gap to the best task-specific features is determined solely by the task's alignment with the discarded, least informative modes, and vanishes to leading order for every square-integrable task once the budget is sufficient.**

***Index Terms*—distributed estimation, side information, universal feature extraction, local information geometry, nested H-score**

I. Introduction

Distributed systems like integrated sensing and communication (ISAC) [1]–[3], IoT, and sensor networks typically comprise edge nodes and a center node. Due to rate and energy constraints, edge nodes transmit block averages of low-dimensional features rather than raw observations. The center node then estimates latent parameters, often aided by side information unavailable at the edge.

For such systems, a primary question is how to design local feature extraction mappings and central estimators. This has long been studied in multi-terminal inference [4]–[6], where local messages are optimized for a fixed task. However, existing designs face two limitations. First, task-oriented designs [7], [8], e.g., supervised learning with a mean-squared-error (MSE) loss, are optimal only for the training task and must be redesigned whenever the task changes. Second, applying universal features [9], [10] directly is sub-optimal with side information S: local features capturing information already in S are redundant, wasting communication resources.

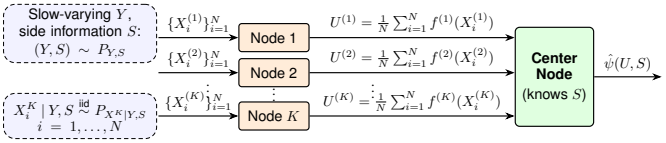


Fig. 1. System model: node k transmits the block average $U^{(k)}$ of its local features, and the center node, which alone knows S, forms the estimate $\hat{\psi}(U,S)$.

A universal feature-extraction scheme is thus desired that is complementary to S and independent of the specific task.

Two further ingredients inform our design. On the physical layer, computation over multiple-access channels [11] and over-the-air aggregation [12], [13] directly deliver sums of node-side preprocessings—an additive structure we formalize at the level of statistical dependence; on the learning side, the H-score objective underlying our approach originates from the centralized Soft-HGR framework [14]. None of the above lines, however, meets both requirements at once.

In this paper, we design such universal features from an information-geometric perspective, building upon the feature geometry framework [15]. Operating in the weak-dependence regime, we first orthogonally decompose the joint function space into a Markov subspace, a conditional subspace, and an interaction subspace. By projecting the canonical dependence kernel (CDK) onto them, we successfully isolate the dependence that is neither captured by the side information nor invisible to distributed implementations. We then show that a distributed nested H-score simultaneously realizes this subspace decomposition and the modal decomposition of the conditional-additive component, whose left modes provide the desired task-independent local features. Finally, we derive the center-node estimator based on these features and characterize its exact leading-order MSE gap to the best task-specific features: the gap is determined by the task's alignment with the discarded modes and vanishes for every square-integrable task once all modes are retained.

II. System Model and Preliminaries

A. System Model

We consider a distributed estimation system comprising K nodes and a single center node, as illustrated in Fig. 1, where

the objective is to estimate a target function $\psi(Y)$ of a latent parameter Y at the center node, based on locally extracted features and a global side information variable S available exclusively at the center node. Let $Y \in \mathcal{Y}$ and $S \in \mathcal{S}$ denote the latent parameter and the side information. We adopt a block-constant model wherein (Y, S) remains invariant over blocks of N time slots, with N large enough, and is treated as fixed in the analysis below. Let $X^K = (X^{(1)}, \dots, X^{(K)})^\mathsf{T} \in \mathcal{X}^K \triangleq \mathcal{X}^{(1)} \times \cdots \times \mathcal{X}^{(K)}$ denote the joint observation, governed by a joint distribution $P_{X^K,S,Y}$. Conditioned on $(Y, S) = (y, s)$, the nodes observe N independent and identically distributed (i.i.d.) samples $\{X_i^K\}_{i=1}^N$ from $P_{X^K|Y=y,S=s}$, and each node $k \in \{1, \dots, K\}$ observes only its local component sequence $\{X_i^{(k)}\}_{i=1}^N$.

Subject to communication constraints, the nodes cannot transmit raw observations. Instead, each node k applies a local feature mapping $f^{(k)} : \mathcal{X}^{(k)} \to \mathbb{R}^d$, where d denotes the per-node feature dimension budget, and transmits the block average

$$U^{(k)} = \frac{1}{N} \sum_{i=1}^{N} f^{(k)}(X_i^{(k)}) \in \mathbb{R}^d. \tag{1}$$

The center node concatenates the transmissions into the stacked vector $U = (U^{(1)\mathsf{T}}, \dots, U^{(K)\mathsf{T}})^\mathsf{T} \in \mathbb{R}^{Kd}$ and forms an estimate $\hat{\psi}(U, S)$ of $\psi(Y)$ as an arbitrary measurable function of (U, S), measured by the MSE $\mathbb{E}[(\hat{\psi}(U, S) - \psi(Y))^2]$. Without loss of generality, all local feature mappings $\{f^{(k)}\}$ and the target function ψ are assumed zero-mean, since constant offsets can be absorbed into the center estimator. This paper addresses the joint design problem

$$\min_{\substack{f^{(k)} \in \mathcal{F}^d_{\mathcal{X}^{(k)}},\, k \in [K] \\ \hat{\psi}:\, \mathbb{R}^{Kd} \times \mathcal{S} \to \mathbb{R}}} \mathbb{E}\big[(\hat{\psi}(U, S) - \psi(Y))^2\big], \tag{2}$$

where $[K] \triangleq \{1, \dots, K\}$ and $\mathcal{F}^d_{\mathcal{Z}}$ denotes the set of d-dimensional square-integrable vector functions on $\mathcal{Z}$ (formalized below). Our goal is feature designs that are *universal*, i.e., independent of the task ψ, yet simultaneously optimal for all tasks, and *complementary to the side information*, i.e., spending no feature dimensions on dependence already explained by S.

Remark: This model captures applications such as ISAC, where edge nodes (e.g., dual-functional terminals) observe signals $X_i^{(k)}$ from the base station, and the center node combines side information S (e.g., ambient temperature) with U to estimate a target function $\psi(Y)$ (e.g., target velocity) rather than the high-dimensional Y itself. Transmitting block averages also meets the low-latency requirements of ISAC, which preclude buffering long sequences, and aligns with gradient aggregation in distributed learning.

B. Feature Geometry and Canonical Dependence Kernels

We work in the feature geometry on $\mathcal{F}_{\mathcal{X}^K \times \mathcal{S} \times \mathcal{Y}}$ with the metric distribution $R_{X^K,S,Y} \triangleq P_{X^K} P_{S,Y}$. Here $\mathcal{F}_{\mathcal{Z}} \triangleq L^2(R_Z)$ denotes the space of square-integrable functions on a measurable space $\mathcal{Z}$ under a probability measure R_Z, with inner product $\langle h_1, h_2 \rangle \triangleq \mathbb{E}_{Z \sim R_Z}[h_1(Z) h_2(Z)]$ and norm $\|h\| = \sqrt{\langle h, h \rangle}$; $\mathcal{F}^d_{\mathcal{Z}} \triangleq (\mathcal{F}_{\mathcal{Z}})^d$. For $f \in \mathcal{F}_{\mathcal{X}}$ and $g \in \mathcal{F}_{\mathcal{Y}}$ the tensor product is $(f \otimes g)(x, y) \triangleq f(x) g(y)$; for vector-valued $f = (f_1, \dots, f_d)^\mathsf{T}$, $g = (g_1, \dots, g_d)^\mathsf{T}$, write $f \otimes g \triangleq \sum_{i=1}^d f_i \otimes g_i$; for subspaces $\mathcal{V}, \mathcal{W}$, $\mathcal{V} \otimes \mathcal{W} \triangleq \operatorname{span}\{f \otimes g : f \in \mathcal{V}, g \in \mathcal{W}\}$. $\Pi(h; \mathcal{V})$ denotes the orthogonal projection of h onto a closed subspace $\mathcal{V}$ (elementwise for vector-valued h), $\mathcal{V} \boxplus \mathcal{W}$ the orthogonal direct sum, and $\mathcal{V} \boxminus \mathcal{W}$ the orthogonal complement of $\mathcal{W}$ in $\mathcal{V}$. A local function $f^{(k)} \in \mathcal{F}_{\mathcal{X}^{(k)}}$ is identified with its embedding in $\mathcal{F}_{\mathcal{X}^K}$, $x^K \mapsto f^{(k)}(x^{(k)})$.

Under the metric distribution, the induced metric on $\mathcal{F}_{\mathcal{X}^K}$ is P_{X^K} and that on $\mathcal{F}_{\mathcal{S} \times \mathcal{Y}}$ is $P_{S,Y}$. Since conditional expectation is an orthogonal projection, the side space admits the decomposition $\mathcal{F}_{\mathcal{S} \times \mathcal{Y}} = \mathcal{F}_{\mathcal{S}} \boxplus \mathcal{F}_{\mathcal{Y}|\mathcal{S}}$, where

$$\mathcal{F}_{\mathcal{Y}|\mathcal{S}} \triangleq \{h \in \mathcal{F}_{\mathcal{S} \times \mathcal{Y}} : \mathbb{E}[h(S, Y) \,|\, S] = 0\}, \tag{3}$$

$\mathcal{F}_{\mathcal{S}}$ is identified with its embedding $\{h(s, y) \equiv h(s)\}$, and $\mathcal{F}^d_{\mathcal{Y}|\mathcal{S}} \triangleq (\mathcal{F}_{\mathcal{Y}|\mathcal{S}})^d$.

The *joint CDK* of $(X^K; S, Y)$ is

$$\mathfrak{i}_{X^K;S,Y}(x^K, s, y) \triangleq \frac{P_{X^K,S,Y}(x^K, s, y)}{P_{X^K}(x^K) P_{S,Y}(s, y)} - 1. \tag{4}$$

To isolate the effect of the side information, we invoke the orthogonal decomposition of function spaces [15, Prop. 22], which splits the joint CDK orthogonally into the *Markov component* $\mathfrak{i}_{X^K;S} \in \mathcal{F}_{\mathcal{X}^K} \otimes \mathcal{F}_{\mathcal{S}}$ and the *conditional dependence component* $\mathfrak{i}_{X^K;Y|S} \in \mathcal{F}_{\mathcal{X}^K} \otimes \mathcal{F}_{\mathcal{Y}|\mathcal{S}}$:

$$\begin{aligned} \mathfrak{i}_{X^K;S,Y} &= \mathfrak{i}_{X^K;S} + \mathfrak{i}_{X^K;Y|S}, \\ \left\|\mathfrak{i}_{X^K;S,Y}\right\|^2 &= \left\|\mathfrak{i}_{X^K;S}\right\|^2 + \left\|\mathfrak{i}_{X^K;Y|S}\right\|^2, \end{aligned} \tag{5}$$

where

$$\begin{aligned} \mathfrak{i}_{X^K;S}(x^K, s) &= \frac{P_{X^K|S=s}(x^K)}{P_{X^K}(x^K)} - 1, \\ \mathfrak{i}_{X^K;Y|S}(x^K, s, y) &= \frac{P_{X^K|S=s,Y=y}(x^K) - P_{X^K|S=s}(x^K)}{P_{X^K}(x^K)}. \end{aligned}$$

Throughout, we analyze the weak-dependence regime

$$\sup_{x^K,s,y} \left|\mathfrak{i}_{X^K;S,Y}(x^K, s, y)\right| = O(\epsilon),$$

which both components in (5) inherit, as an average and a difference of the joint CDK.

Writing $\mathbb{E}[f \mid s] \triangleq \mathbb{E}[f(X^K) \mid S = s]$ and $\mathbb{E}[f \mid s, y] \triangleq \mathbb{E}[f(X^K) \mid S = s, Y = y]$, the following pairing identities, which connect the CDKs with conditional expectations, hold for any $f \in \mathcal{F}_{\mathcal{X}^K}$ (and hence for every embedded local function $f^{(k)}$):

$$\left\langle \mathfrak{i}_{X^K;S}(\cdot, s),\, f \right\rangle_{P_{X^K}} = \mathbb{E}[f \mid s] - \mathbb{E}[f], \tag{6}$$

$$\left\langle \mathfrak{i}_{X^K;Y|S}(\cdot, s, y),\, f \right\rangle_{P_{X^K}} = \mathbb{E}[f \mid s, y] - \mathbb{E}[f \mid s]. \tag{7}$$

C. Modal Decomposition and H-Score

Suppose $\gamma \in \mathcal{F}_{\mathcal{X} \times \mathcal{Y}}$ admits a finite modal decomposition

$$\gamma = \sum_{i=1}^{\operatorname{rank}(\gamma)} \zeta_i(\gamma) = \sum_{i=1}^{\operatorname{rank}(\gamma)} \sigma_i \, (f_i^* \otimes g_i^*), \tag{8}$$

with singular values $\sigma_1 \geq \sigma_2 \geq \cdots > 0$, orthonormal modes $\langle f_i^*, f_j^* \rangle = \langle g_i^*, g_j^* \rangle = \delta_{ij}$, and $\operatorname{rank}(\gamma)$ the number of nonzero singular values. The truncation $\zeta_{\leq i}(\gamma) \triangleq \sum_{j \leq i} \zeta_j(\gamma)$, with $\zeta_j(\gamma) \triangleq 0$ for $j > \operatorname{rank}(\gamma)$, is the optimal rank-$i$ approx-

imation, unique if $\sigma_i > \sigma_{i+1}$ (Eckart–Young–Mirsky) [15]. When γ is the CDK of a pair of variables, the modes are their Hirschfeld–Gebelein–Rényi maximal correlation functions [16].

The H-score for features $f \in \mathcal{F}_{\mathcal{X}}^d$, $g \in \mathcal{F}_{\mathcal{Y}}^d$ is

$$\begin{aligned}\mathscr{H}(f,g) &= \tfrac{1}{2}\big(\|\mathfrak{i}_{X;Y}\|^2 - \|\mathfrak{i}_{X;Y} - f \otimes g\|^2\big) \\ &= \mathbb{E}_{P_{X,Y}}[f^{\mathsf{T}}(X)g(Y)] - \tfrac{1}{2}\operatorname{tr}(\Lambda_f \Lambda_g),\end{aligned} \tag{9}$$

with $\Lambda_f = \mathbb{E}_{P_X}[ff^{\mathsf{T}}]$, $\Lambda_g = \mathbb{E}_{P_Y}[gg^{\mathsf{T}}]$, where $\mathfrak{i}_{X;Y}$ denotes the CDK of a generic pair $(X;Y)$; its maximizers satisfy $f \otimes g = \zeta_{\le d}(\mathfrak{i}_{X;Y})$, and its nested version sums the H-scores of the coordinate prefixes to yield the ordered modes with automatic orthogonality [15].

III. Additive Geometry and Triple Orthogonal Decomposition

In this section, we first characterize how the center node estimates with given features, which motivates the additive space, and then establish the orthogonal decomposition of the joint function space based on it.

A. Estimation with Given Feature Extractors

Let $f \triangleq (f^{(1)\mathsf{T}}, \dots, f^{(K)\mathsf{T}})^{\mathsf{T}} : \mathcal{X}^K \to \mathbb{R}^{Kd}$ denote the stacked feature map, so that $U = \frac{1}{N}\sum_{i=1}^N f(X_i^K)$. Define the feature covariance matrix $\Lambda_f \in \mathbb{R}^{Kd\times Kd}$, the conditional feature variation $\beta(y,s) \in \mathbb{R}^{Kd}$, and the task–feature correlation $c_\psi(s) \in \mathbb{R}^{Kd}$ as

$$\begin{aligned}\Lambda_f &\triangleq \mathbb{E}_{P_{X^K}}[ff^{\mathsf{T}}], \quad \beta(y,s) \triangleq \mathbb{E}[f \mid s,y] - \mathbb{E}[f \mid s], \\ c_\psi(s) &\triangleq \mathbb{E}_{Y|S=s}[\psi(Y)\beta(Y,s)],\end{aligned} \tag{10}$$

where Λ_f is assumed nonsingular on the retained feature span, which can always be arranged by removing redundant coordinates. Since N is large, we model $U \mid (Y,S) = (y,s)$ as Gaussian $\mathcal{N}\big(\mathbb{E}[f \mid s,y], \frac{1}{N}\Lambda_f\big)$, neglecting the $O(\epsilon)$ dependence of the covariance on (s,y), whose effect on the posterior is a factor $\sqrt{N}$ smaller than that of the mean. We then have the following characterization; the proof is deferred to Appendix A.

Lemma 1. *For any given local features f, the posterior distribution of Y at the center node is approximated by*

$$p_{Y|U,S}(y \mid u,s) = p_{Y|S}(y \mid s)\big[1 + N\, u^{\mathsf{T}}\Lambda_f^{-1}\beta(y,s)\big] + O(\epsilon^2). \tag{11}$$

Consequently, for any $\psi \in \mathcal{F}_{\mathcal{Y}}$, the conditional expectation of $\psi(Y)$ given U and S is approximated by

$$\mathbb{E}[\psi(Y) \mid U,S] = \mathbb{E}[\psi(Y) \mid S] + N \cdot c_\psi(S)^{\mathsf{T}}\Lambda_f^{-1}\, U + O(\epsilon^2). \tag{12}$$

Thus, under weak dependence, both the posterior of Y and the conditional expectation of $\psi(Y)$ are linear in U. Hence, for given extractors f, the minimum mean-squared error (MMSE) estimator of $\psi(Y)$ at the center node is the leading-order conditional expectation

$$\hat{\psi}^*(U,S;f) = \mathbb{E}[\psi(Y) \mid S] + N \cdot c_\psi(S)^{\mathsf{T}}\Lambda_f^{-1}U, \tag{13}$$

and the corresponding achievable MSE is

$$\begin{aligned}\mathsf{MSE}_\psi(f) &\triangleq \min_{\hat{\psi}} \mathbb{E}\big[(\hat{\psi}(U,S) - \psi(Y))^2\big] \\ &= \mathbb{E}[\operatorname{Var}(\psi(Y) \mid S)] - N \cdot \mathbb{E}\big[c_\psi(S)^{\mathsf{T}}\Lambda_f^{-1}c_\psi(S)\big] + o(\epsilon^2).\end{aligned} \tag{14}$$

Alternatively, by defining the *task image*

$$\bar{\psi}(x^K, s) \triangleq \mathbb{E}_{Y|S=s}\big[\psi(Y)\, \mathfrak{i}_{X^K;Y|S}(x^K,s,Y)\big] \in \mathcal{F}_{\mathcal{X}^K \times \mathcal{S}}, \tag{15}$$

the MSE admits the projection form

$$\begin{aligned}\mathsf{MSE}_\psi(f) &= \mathbb{E}[\operatorname{Var}(\psi(Y) \mid S)] \\ &\quad - N \cdot \big\|\Pi\big(\bar{\psi};\, \operatorname{span}\{f\} \otimes \mathcal{F}_{\mathcal{S}}\big)\big\|^2_{P_{X^K}P_{\mathcal{S}}} + o(\epsilon^2),\end{aligned} \tag{16}$$

where $\operatorname{span}\{f\} = \operatorname{span}\{f_i : i \in [Kd]\} \subseteq \mathcal{F}_{\mathcal{X}^K}$ denotes the span of all stacked components; see Appendix B.

Since the center node estimates by linearly combining the received features, the performance of a feature design depends only on the subspace spanned by the local features. This motivates the *additive space*

$$\mathcal{A} \triangleq \mathcal{F}_{\mathcal{X}^{(1)}} + \cdots + \mathcal{F}_{\mathcal{X}^{(K)}} = \overline{\textstyle\bigcup_f \operatorname{span}\{f\}} \subseteq \mathcal{F}_{\mathcal{X}^K}, \tag{17}$$

where the local spaces are embedded in $\mathcal{F}_{\mathcal{X}^K}$ as above, and the closure is taken over all finite-dimensional feasible subspaces. Moreover, for any side space $\mathcal{V} \subseteq \mathcal{F}_{\mathcal{S}\times\mathcal{Y}}$, $\sum_{k=1}^K \big(\mathcal{F}_{\mathcal{X}^{(k)}} \otimes \mathcal{V}\big) = \mathcal{A} \otimes \mathcal{V} \subseteq \mathcal{F}_{\mathcal{X}^K\times\mathcal{S}\times\mathcal{Y}}$.

B. Triple Orthogonal Decomposition

Combining the orthogonal decomposition of the X^K-side, $\mathcal{F}_{\mathcal{X}^K} = \mathcal{A} \boxplus \big(\mathcal{F}_{\mathcal{X}^K} \boxminus \mathcal{A}\big)$, with the side-space decomposition (3), the full space splits into three mutually orthogonal subspaces:

$$\begin{aligned}\mathcal{F}_{\mathcal{X}^K\times\mathcal{S}\times\mathcal{Y}} \;=\; &\underbrace{\big(\mathcal{A} \otimes \mathcal{F}_{\mathcal{S}}\big)}_{\text{Markov subspace}} \;\boxplus\; \underbrace{\big(\mathcal{A} \otimes \mathcal{F}_{\mathcal{Y}|\mathcal{S}}\big)}_{\text{conditional subspace}} \\ &\boxplus\; \underbrace{\big((\mathcal{F}_{\mathcal{X}^K} \boxminus \mathcal{A}) \otimes \mathcal{F}_{\mathcal{S}\times\mathcal{Y}}\big)}_{\text{interaction subspace}}.\end{aligned} \tag{18}$$

The first two pieces form the *accessible subspace* $\mathcal{A} \otimes \mathcal{F}_{\mathcal{S}\times\mathcal{Y}}$ realizable under the communication constraint; the third piece is invisible to any distributed implementation. Projecting the joint CDK onto these three subspaces gives the following decomposition.

Definition 1. *The* Markov-additive*,* conditional-additive*, and* interaction components *of the joint CDK are its projections onto the three subspaces in* (18)*:*

$$\begin{aligned}\pi_{\mathsf{M}} &\triangleq \Pi\big(\mathfrak{i}_{X^K;S,Y};\, \mathcal{A} \otimes \mathcal{F}_{\mathcal{S}}\big) = \Pi\big(\mathfrak{i}_{X^K;S};\, \mathcal{A} \otimes \mathcal{F}_{\mathcal{S}}\big), \\ \pi_{\mathsf{C}} &\triangleq \Pi\big(\mathfrak{i}_{X^K;S,Y};\, \mathcal{A} \otimes \mathcal{F}_{\mathcal{Y}|\mathcal{S}}\big) = \Pi\big(\mathfrak{i}_{X^K;Y|S};\, \mathcal{A} \otimes \mathcal{F}_{\mathcal{Y}|\mathcal{S}}\big), \\ \pi_{\mathsf{I}} &\triangleq \mathfrak{i}_{X^K;S,Y} - \pi_{\mathsf{M}} - \pi_{\mathsf{C}}.\end{aligned}$$

The alternative expressions of π_{M} and π_{C} are due to the split (5): $\mathfrak{i}_{X^K;Y|S} \perp \mathcal{A} \otimes \mathcal{F}_{\mathcal{S}}$ and $\mathfrak{i}_{X^K;S} \perp \mathcal{A} \otimes \mathcal{F}_{\mathcal{Y}|\mathcal{S}}$. Moreover, $\pi_{\mathsf{I}} = \Pi\big(\mathfrak{i}_{X^K;S,Y};\, \big(\mathcal{F}_{\mathcal{X}^K} \boxminus \mathcal{A}\big) \otimes \mathcal{F}_{\mathcal{S}\times\mathcal{Y}}\big)$. By the orthogonality of (18), the three components are mutually orthogonal, i.e.,

$$\begin{aligned}\mathfrak{i}_{X^K;S,Y} &= \pi_{\mathsf{M}} + \pi_{\mathsf{C}} + \pi_{\mathsf{I}}, \\ \big\|\mathfrak{i}_{X^K;S,Y}\big\|^2 &= \|\pi_{\mathsf{M}}\|^2 + \|\pi_{\mathsf{C}}\|^2 + \|\pi_{\mathsf{I}}\|^2.\end{aligned} \tag{19}$$

Intuitively, π_{M} is dependence that the center node already knows through S, hence useless for estimating $\psi(Y)$; π_{C} is the dependence on Y beyond S that additive features can extract, hence crucial for estimation; and π_{I} is invisible to any distributed implementation.

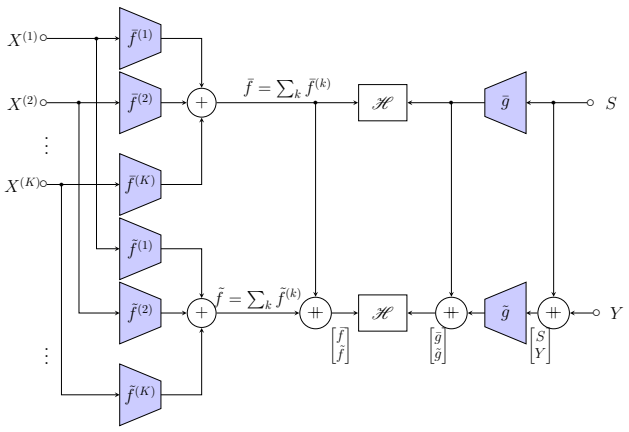


Fig. 2. The nested H-score network for the modal decomposition of the Markov-additive component π_{M} and the conditional-additive component π_{C}.

IV. Modal Decomposition with Nested H-Score

This section first shows how the modal decomposition of the two accessible components is obtained by a distributed nested H-score, and then develops the center-node estimator and its MSE gap.

A. Deriving the Modal Decomposition via Nested H-Scores

Since π_{I} is invisible to any distributed implementation, we decompose only π_{M} and π_{C}, with feature functions of dimensions $\bar{d}$ and $\tilde{d}$:

$$\bar{f} \in \mathcal{A}^{\bar{d}}, \quad \bar{g} \in \mathcal{F}_{\mathcal{S}}^{\bar{d}}, \qquad \tilde{f} \in \mathcal{A}^{\tilde{d}}, \quad \tilde{g} \in \mathcal{F}_{\mathcal{S}\times\mathcal{Y}}^{\tilde{d}}, \tag{20}$$

where $\mathcal{A}^d$ denotes the set of d-dimensional vector functions whose coordinates lie in $\mathcal{A}$, i.e., $\bar{f} = \sum_{k=1}^K \bar{f}^{(k)}$ with $\bar{f}^{(k)} \in \mathcal{F}_{\mathcal{X}^{(k)}}^{\bar{d}}$, and similarly for $\tilde{f}$. The left branches are subject to the distributed (additive) constraint, whereas the right branches, realized at the center node during offline training, are unconstrained. Following the nesting technique of [15], we define $\mathscr{H}_{\mathrm{MC}}$ as the sum of two H-scores:

$$\mathscr{H}_{\mathrm{MC}}\left(\begin{bmatrix}\bar{f}\\ \tilde{f}\end{bmatrix}, \begin{bmatrix}\bar{g}\\ \tilde{g}\end{bmatrix}\right) \triangleq \mathscr{H}(\bar{f},\bar{g}) + \mathscr{H}\left(\begin{bmatrix}\bar{f}\\ \tilde{f}\end{bmatrix}, \begin{bmatrix}\bar{g}\\ \tilde{g}\end{bmatrix}\right), \tag{21}$$

where the first H-score is between $(X^K; S)$ and the second between $(X^K; S, Y)$, both under the metric distribution $R_{X^K,S,Y}$. The corresponding nested H-score network structure is shown in Fig. 2.

Maximizing $\mathscr{H}_{\mathrm{MC}}$ over $\bar{f}, \bar{g}, \tilde{f}, \tilde{g}$ yields the modal decomposition of both components, as formalized below (proof in Appendix C).

Proposition 1. *For $\bar{f}, \bar{g}, \tilde{f}, \tilde{g}$ defined in (20) and $\bar{d} \geq \mathrm{rank}(\pi_{\mathsf{M}})$, the score (21) is maximized if and only if*

$$\bar{f} \otimes \bar{g} = \pi_{\mathsf{M}}, \quad \tilde{f} \otimes \tilde{g} = \begin{cases} \zeta_{\leq \tilde{d}}(\pi_{\mathsf{C}}), & \tilde{d} < \mathrm{rank}(\pi_{\mathsf{C}}), \\ \pi_{\mathsf{C}}, & \tilde{d} \geq \mathrm{rank}(\pi_{\mathsf{C}}). \end{cases} \tag{22}$$

B. Estimation with the Modal Decomposition

Let $\pi_{\mathsf{C}} = \sum_{i=1}^r \sigma_i (f_i^* \otimes g_i^*)$, $r \triangleq \mathrm{rank}(\pi_{\mathsf{C}})$, be the modal decomposition (8) of π_{C}. For a given budget $\tilde{d}$, let $\tilde{f}^* = \sum_{k=1}^K \tilde{f}^{(k)*} \in \mathcal{A}^{\tilde{d}}$ and $\tilde{g}^* \in \mathcal{F}_{\mathcal{Y}|\mathcal{S}}^{\tilde{d}}$ be maximizers of $\mathscr{H}_{\mathrm{MC}}$, so that, by Proposition 1,

$$\tilde{f}^* \otimes \tilde{g}^* = \begin{cases} \zeta_{\leq \tilde{d}}(\pi_{\mathsf{C}}), & \tilde{d} < r, \\ \pi_{\mathsf{C}}, & \tilde{d} \geq r. \end{cases} \tag{23}$$

Each node k uses $f^{(k)} = \tilde{f}^{(k)*}$ (so $d = \tilde{d}$), and the center node uses the aggregate

$$V \triangleq \sum_{k=1}^K U^{(k)} = \frac{1}{N}\sum_{i=1}^N \tilde{f}^*(X_i^K) \in \mathbb{R}^{\tilde{d}},$$

which over-the-air aggregation [12], [13] delivers directly. Lemma 1 applies verbatim to the feature map $\tilde{f}^*$, with $\tilde{\beta}(y,s) \triangleq \mathbb{E}[\tilde{f}^* \mid s, y] - \mathbb{E}[\tilde{f}^* \mid s] = \langle \tilde{f}^*, \mathrm{i}_{X^K;Y|S}(\cdot, s, y)\rangle_{P_{X^K}}$ by (7). Since the coordinates of $\tilde{f}^*$ lie in $\mathcal{A}$, only π_{C} contributes to this inner product; moreover, $\pi_{\mathsf{C}} - \tilde{f}^* \otimes \tilde{g}^* = \sum_{i>\tilde{d}} \sigma_i (f_i^* \otimes g_i^*)$ is orthogonal to the coordinates of $\tilde{f}^*$, whose span is $\mathrm{span}\{f_1^*, \ldots, f_{\tilde{d}}^*\}$ when $\tilde{d} < r$. Hence

$$\tilde{\beta}(y,s) = \Lambda_{\tilde{f}^*}\, \tilde{g}^*(y,s), \qquad \Lambda_{\tilde{f}^*} \triangleq \mathbb{E}\big[\tilde{f}^* \tilde{f}^{*\mathsf{T}}\big], \tag{24}$$

which gives the following characterization of the distributed estimator.

Proposition 2. *With the local feature extractors $\tilde{f}^{(k)*}$, the posterior distribution and the conditional expectation at the center node are*

$$p_{Y|V,S}(y \mid v, s) = p_{Y|S}(y \mid s)\big[1 + N v^{\mathsf{T}} \tilde{g}^*(y,s)\big] + O(\epsilon^2), \tag{25}$$

$$\mathbb{E}[\psi(Y) \mid V, S] = \mathbb{E}[\psi(Y) \mid S] + N \cdot V^{\mathsf{T}} \tilde{c}_\psi(S) + O(\epsilon^2), \tag{26}$$

where $\tilde{c}_\psi(S) \triangleq \mathbb{E}\big[\tilde{g}^*(Y,S)\,\psi(Y)\,\big|\,S\big]$. Hence the MMSE estimator of ψ in the leading order is

$$\hat{\psi}^*(V,S) = \mathbb{E}[\psi(Y) \mid S] + N \cdot V^{\mathsf{T}} \tilde{c}_\psi(S), \tag{27}$$

and its MSE for any task $\psi \in \mathcal{F}_{\mathcal{Y}}$ is

$$\begin{aligned} \mathsf{MSE}_\psi^{\mathrm{univ}} &= \mathbb{E}\big[\mathrm{Var}(\psi(Y) \mid S)\big] \\ &\quad - N\, \mathbb{E}\big[\tilde{c}_\psi(S)^{\mathsf{T}} \Lambda_{\tilde{f}^*} \tilde{c}_\psi(S)\big] + o(\epsilon^2). \end{aligned} \tag{28}$$

Proof. Substitute $\Lambda_{\tilde{f}^*}^{-1} \tilde{\beta} = \tilde{g}^*$ from (24) into (11) and (12), and substitute $c_\psi = \Lambda_{\tilde{f}^*} \tilde{c}_\psi$, which follows from (24), into (14). □

Once the modes are learned offline, e.g., by maximizing the empirical nested H-score, the node-side maps $\tilde{f}^{(k)*}$ are fixed and task-independent: when the task ψ changes, only the coefficients $\tilde{c}_\psi$ and the baseline $\mathbb{E}[\psi(Y) \mid S]$ are recomputed at the center node, without retraining the nodes.

C. Optimality and Universality of the Feature Extractors

Let $\mathsf{MSE}_\psi^* \triangleq \inf_f \mathsf{MSE}_\psi(f)$ be the best MSE achievable by task-specific extractors of arbitrary dimension. Since $\mathrm{span}\{f\}$ ranges over all finite-dimensional subspaces of $\mathcal{A}$, (16) and the closure characterization (17) give

$$\begin{aligned} \mathsf{MSE}_\psi^* &= \mathbb{E}[\mathrm{Var}(\psi(Y) \mid S)] \\ &\quad - N \cdot \big\|\Pi\big(\bar{\psi};\, \mathcal{A} \otimes \mathcal{F}_{\mathcal{S}}\big)\big\|^2_{P_{X^K} P_S} + o(\epsilon^2), \end{aligned} \tag{29}$$

attained when $\mathrm{span}\{f\}$ contains the left singular space of $\Pi\big(\bar{\psi};\, \mathcal{A} \otimes \mathcal{F}_{\mathcal{S}}\big)$. Comparing (29) with (28) gives the gap.

Theorem 1. *For any budget $\tilde{d}$ and any task $\psi \in \mathcal{F}_{\mathcal{Y}}$, the universal features in* (23) *satisfy*

$$\mathsf{MSE}_{\psi}^{\mathrm{univ}} - \mathsf{MSE}_{\psi}^{*} = N \sum_{i=\tilde{d}+1}^{r} \sigma_i^2 \, \mathbb{E}\big[\rho_{\psi,i}(S)^2\big] + o(\epsilon^2), \quad (30)$$

where $\rho_{\psi,i}(s) \triangleq \mathbb{E}[g_i^*(Y,s)\psi(Y) \mid S = s]$ is the conditional correlation of the task with the i-th right mode.

Proof. Write $\mathsf{i}_{X^K;Y|S} = \pi_{\mathsf{C}} + \pi_{\mathsf{IC}}$ with $\pi_{\mathsf{IC}} \triangleq \Pi\big(\mathsf{i}_{X^K;Y|S};\, (\mathcal{F}_{\mathcal{X}^K} \boxminus \mathcal{A}) \otimes \mathcal{F}_{\mathcal{Y}|\mathcal{S}}\big)$. In (15), the contribution of π_{IC} lies in $(\mathcal{F}_{\mathcal{X}^K} \boxminus \mathcal{A}) \otimes \mathcal{F}_{\mathcal{S}}$ and is annihilated by the projection onto $\mathcal{A} \otimes \mathcal{F}_{\mathcal{S}}$, whereas that of π_{C} already lies in $\mathcal{A} \otimes \mathcal{F}_{\mathcal{S}}$. Hence

$$\begin{aligned}\Pi\big(\bar{\psi};\, \mathcal{A} \otimes \mathcal{F}_{\mathcal{S}}\big)\,(x^K, s) &= \big\langle \pi_{\mathsf{C}}(x^K, s, \cdot), \psi \big\rangle_{P_{Y|S=s}} \\ &= \textstyle\sum_{i=1}^{r} \sigma_i \, \rho_{\psi,i}(s)\, f_i^*(x^K),\end{aligned}$$

whose squared norm is $\sum_{i=1}^{r} \sigma_i^2 \, \mathbb{E}[\rho_{\psi,i}(S)^2]$ by the orthonormality of $\{f_i^*\}$. Similarly, the definition of $\tilde{c}_\psi$ and (23) give

$$\begin{aligned}\tilde{f}^{*\mathsf{T}}(x^K)\, \tilde{c}_\psi(s) &= \mathbb{E}\big[(\tilde{f}^* \otimes \tilde{g}^*)(x^K, s, Y)\, \psi(Y) \;\big|\; S = s\big] \\ &= \textstyle\sum_{i=1}^{\min\{\tilde{d},r\}} \sigma_i \, \rho_{\psi,i}(s)\, f_i^*(x^K),\end{aligned}$$

so that

$$\mathbb{E}\big[\tilde{c}_\psi(S)^{\mathsf{T}} \Lambda_{\tilde{f}^*} \tilde{c}_\psi(S)\big] = \sum_{i=1}^{\min\{\tilde{d},r\}} \sigma_i^2 \, \mathbb{E}[\rho_{\psi,i}(S)^2].$$

Substituting into (29) and (28) and subtracting yields (30). □

Thus, the penalty of a limited budget is the task's σ_i^2-weighted energy on the discarded right modes. Since the nested H-score discards the modes with the smallest singular values, the loss is small unless the task concentrates on these least informative modes; in particular, any task with $\rho_{\psi,i} \equiv 0$ for all $i > \tilde{d}$ is estimated optimally even when $\tilde{d} < r$. When all modes are retained, the sum in (30) is empty:

Corollary 1. *If $\tilde{d} \geq r$, then for every task $\psi \in \mathcal{F}_{\mathcal{Y}}$,*

$$\mathsf{MSE}_{\psi}^{\mathrm{univ}} = \mathsf{MSE}_{\psi}^{*} + o(\epsilon^2),$$

i.e., the features are simultaneously optimal for all tasks.

Since π_{M} is absorbed by $(\bar{f}, \bar{g})$, which are used only in training and never transmitted, no transmitted dimension is spent on dependence already explained by S. Without side information, i.e., for constant S, we have $\pi_{\mathsf{M}} = 0$ and $\pi_{\mathsf{C}} = \Pi\big(\mathsf{i}_{X^K;Y};\, \mathcal{A} \otimes \mathcal{F}_{\mathcal{Y}}\big)$, so the scheme reduces to universal distributed features without side information.

V. Conclusion

We studied universal local feature design for distributed estimation with side information at the center node. Under weak dependence, a triple orthogonal decomposition shows that only the conditional-additive component of the CDK is useful. Its modal decomposition, learned by a distributed nested H-score without knowledge of the task, yields features whose MSE gap to the best task-specific features is determined by the task's alignment with the discarded modes and vanishes when all modes are retained. We believe this offers a principled basis for reusable, side-information-aware feature design in distributed sensing.

Appendix A
Proof of Lemma 1

Under the Gaussian model of Section III-A, with $\mu_f(y,s) \triangleq \mathbb{E}[f(X^K) \mid s, y]$,

$$\begin{aligned}&p_{Y|U,S}(y \mid u, s) \propto \\ &p_{Y|S}(y \mid s) \cdot \exp\Big(-\tfrac{N}{2}\big(u - \mu_f(y,s)\big)^{\mathsf{T}} \Lambda_f^{-1} \big(u - \mu_f(y,s)\big)\Big).\end{aligned}$$

Expanding the quadratic form and using $\mu_f(y,s) = \mathbb{E}[f \mid s] + \beta(y,s)$ with $\beta(y,s) = O(\epsilon)$ and $\mathbb{E}[f \mid s] = O(\epsilon)$ (by (6) and the weak-dependence bound), the y-dependent part of the exponent is $N\, u^{\mathsf{T}} \Lambda_f^{-1} \beta(y,s) + O(\epsilon^2)$, since $\mu_f^{\mathsf{T}} \Lambda_f^{-1} \mu_f = O(\epsilon^2)$. Exponentiating to first order gives (11), and integrating $\psi(y)$ against it under $P_{Y|S=s}$ yields (12). For the MSE, the orthogonality principle conditioned on S gives

$$\mathsf{MSE}_{\psi}(f) = \mathbb{E}[\mathrm{Var}(\psi(Y) \mid S)] - \mathbb{E}\big[\mathrm{Var}(\mathbb{E}[\psi(Y) \mid U, S] \mid S)\big],$$

where, by (12),

$$\begin{aligned}&\mathrm{Var}(\mathbb{E}[\psi(Y) \mid U, S] \mid S) \\ &\quad = N^2 c_\psi(S)^{\mathsf{T}} \Lambda_f^{-1}\, \mathbb{E}[UU^{\mathsf{T}} \mid S]\, \Lambda_f^{-1} c_\psi(S) + o(\epsilon^2),\end{aligned}$$

with $\mathbb{E}[UU^{\mathsf{T}} \mid S] = \frac{1}{N}\Lambda_f + O(\epsilon^2)$ (using $\mathbb{E}[U \mid S] = O(\epsilon)$ and the law of total variance). Substitution gives (14).

Appendix B
Proof of the Projection Form (16)

Define $\eta(x^K, s) \triangleq f(x^K)^{\mathsf{T}} \Lambda_f^{-1} c_\psi(s) \in \mathrm{span}\{f\} \otimes \mathcal{F}_{\mathcal{S}}$, whose squared norm under $P_{X^K} P_S$ is $\mathbb{E}_S\big[c_\psi(S)^{\mathsf{T}} \Lambda_f^{-1} c_\psi(S)\big]$, precisely the correction term in (14). It remains to show $\eta = \Pi\big(\bar{\psi};\, \mathrm{span}\{f\} \otimes \mathcal{F}_{\mathcal{S}}\big)$; since the tensors $f_i \otimes c$ ($i \in [Kd]$, $c \in \mathcal{F}_{\mathcal{S}}$) span this space, it suffices to verify $\langle \bar{\psi} - \eta, f_i \otimes c \rangle = 0$. By (15), Fubini, and (7),

$$\mathbb{E}_{P_{X^K}}[\bar{\psi}(\cdot, s)\, f_i] = \mathbb{E}_{Y|S=s}\big[\psi(Y)\, \beta_i(Y,s)\big] = c_{\psi,i}(s),$$

and $\mathbb{E}_{P_{X^K}}[\eta(\cdot,s) f_i] = (\Lambda_f \Lambda_f^{-1} c_\psi(s))_i = c_{\psi,i}(s)$. Hence η satisfies the defining orthogonality of the projection, and substituting its norm into (14) proves (16).

Appendix C
Proof of Proposition 1

Let $\bar{\gamma} \triangleq \bar{f} \otimes \bar{g}$ and $\tilde{\gamma} \triangleq \tilde{f} \otimes \tilde{g}$, and write $F \triangleq (\bar{f}^{\mathsf{T}}, \tilde{f}^{\mathsf{T}})^{\mathsf{T}}$, $G \triangleq (\bar{g}^{\mathsf{T}}, \tilde{g}^{\mathsf{T}})^{\mathsf{T}}$. Then $\bar{\gamma} \in \mathcal{A} \otimes \mathcal{F}_{\mathcal{S}}$, and the projections of $\tilde{\gamma}$ onto the subspaces in (18) are $\tilde{\gamma}_{\mathsf{I}} = 0$, $\tilde{\gamma}_{\mathsf{M}} = \sum_j \tilde{f}_j \otimes \Pi(\tilde{g}_j;\, \mathcal{F}_{\mathcal{S}})$, and $\tilde{\gamma}_{\mathsf{C}} = \sum_j \tilde{f}_j \otimes \Pi\big(\tilde{g}_j;\, \mathcal{F}_{\mathcal{Y}|\mathcal{S}}\big)$, the latter two of rank at most $\tilde{d}$. Using the geometric form (9) of the H-score and the orthogonality of (5) and (18),

$$\mathscr{H}(\bar{f}, \bar{g}) = \tfrac{1}{2}\,\|\pi_{\mathsf{M}}\|^2 - \tfrac{1}{2}\,\|\pi_{\mathsf{M}} - \bar{\gamma}\|^2, \quad (31)$$

$$\begin{aligned}\mathscr{H}(F, G) = {}& \tfrac{1}{2}\big(\|\pi_{\mathsf{M}}\|^2 + \|\pi_{\mathsf{C}}\|^2\big) \\ & - \tfrac{1}{2}\Big[\|\pi_{\mathsf{M}} - \bar{\gamma} - \tilde{\gamma}_{\mathsf{M}}\|^2 + \|\pi_{\mathsf{C}} - \tilde{\gamma}_{\mathsf{C}}\|^2\Big], \quad (32)\end{aligned}$$

where (31) uses $\mathsf{i}_{X^K;S} - \pi_{\mathsf{M}} \perp \pi_{\mathsf{M}} - \bar{\gamma}$, and π_{I} cancels in (32) since $\bar{\gamma} + \tilde{\gamma}$ is orthogonal to the interaction subspace. Hence maximizing $\mathscr{H}_{\mathrm{MC}}$ is equivalent to minimizing

$$\|\pi_{\mathsf{M}} - \bar{\gamma}\|^2 + \|\pi_{\mathsf{M}} - \bar{\gamma} - \tilde{\gamma}_{\mathsf{M}}\|^2 + \|\pi_{\mathsf{C}} - \tilde{\gamma}_{\mathsf{C}}\|^2,$$

whose three nonnegative terms attain their minima simultaneously iff $\bar{\gamma} = \pi_{\mathsf{M}}$ (feasible since $\bar{d} \geq \mathrm{rank}(\pi_{\mathsf{M}})$), $\tilde{\gamma}_{\mathsf{M}} = 0$, and $\tilde{\gamma}_{\mathsf{C}} = \zeta_{\leq \tilde{d}}(\pi_{\mathsf{C}})$ (by Eckart–Young–Mirsky), i.e., iff (22) holds.

Declaration on the Use of Artificial Intelligence

In this work, the authors used DeepSeek-V4-Pro to assist with the proofs, and used GLM-5.3 and Kimi-K3 to assist with polishing each section of the paper. All AI-assisted content was carefully reviewed and edited by authors, who take full responsibility for the integrity and accuracy of this publication.

References

[1] F. Liu, Y. Cui, C. Masouros, J. Xu, T. X. Han, Y. C. Eldar, and S. Buzzi, "Integrated sensing and communications: Toward dual-functional wireless networks for 6G and beyond," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 6, pp. 1728–1767, 2022.

[2] M. Ahmadipour, M. Kobayashi, M. Wigger, and G. Caire, "An information-theoretic approach to joint sensing and communication," *IEEE Transactions on Information Theory*, vol. 70, no. 2, pp. 1124–1146, 2024.

[3] X. Yuan, T. Peng, H. Fu, and S.-L. Huang, "Unbounded Multi-Letter Marginal Gain for Bistatic Integrated Sensing and Communication," in *2025 Asia Pacific Workshop on Data Science and Information Theory (APWDSIT)*, Oct. 2025, pp. 1–6.

[4] R. R. Tenney and N. R. Sandell, Jr., "Detection with distributed sensors," *IEEE Transactions on Aerospace and Electronic Systems*, vol. AES-17, no. 4, pp. 501–510, 1981.

[5] R. Ahlswede and I. Csiszár, "Hypothesis testing with communication constraints," *IEEE Transactions on Information Theory*, vol. IT-32, no. 4, pp. 533–542, 1986.

[6] T. S. Han and S.-I. Amari, "Statistical inference under multiterminal data compression," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2300–2324, 1998.

[7] J. Shao, Y. Mao, and J. Zhang, "Learning task-oriented communication for edge inference: An information bottleneck approach," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 1, pp. 197–211, 2022.

[8] D. Gündüz *et al.*, "Beyond transmitting bits: Context, semantics, and task-oriented communications," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 1, pp. 5–41, 2023.

[9] X. Xu and S.-L. Huang, "An Information Theoretic Framework for Distributed Learning Algorithms," in *2021 IEEE International Symposium on Information Theory (ISIT)*. Melbourne, Australia: IEEE, Jul. 2021, pp. 314–319.

[10] D. Tang, J. Y. Tan, and G. Han, "Sharp Regret Bounds and a Task-Covariance Correction for Spectral Representation Learning," arXiv:2609.15825, Sep. 2026.

[11] B. Nazer and M. Gastpar, "Computation over multiple-access channels," *IEEE Transactions on Information Theory*, vol. 53, no. 10, pp. 3498–3516, 2007.

[12] M. Gastpar, "Uncoded transmission is exactly optimal for a simple Gaussian "sensor" network," *IEEE Transactions on Information Theory*, vol. 54, no. 11, pp. 5247–5251, 2008.

[13] G. Zhu, Y. Wang, and K. Huang, "Broadband analog aggregation for low-latency federated edge learning," *IEEE Transactions on Wireless Communications*, vol. 19, no. 1, pp. 491–506, 2020.

[14] L. Wang, J. Wu, S.-L. Huang, L. Zheng, X. Xu, L. Zhang, and J. Huang, "An efficient approach to informative feature extraction from multimodal data," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 1, 2019, pp. 5281–5288.

[15] X. Xu and L. Zheng, "Neural Feature Learning in Function Space," *Journal of Machine Learning Research*, vol. 25, no. 142, pp. 1–76, 2024.

[16] X. Xu and S.-L. Huang, "Maximal Correlation Regression," *IEEE Access*, vol. 8, pp. 26 591–26 601, 2020.