DEEP OPERATOR-VALUED SPECTRAL KERNEL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep Spectral Kernel Networks (DSKNs) construct nonstationary representations via paired-frequency maps, but independent phase sampling collapses cross-frequency interactions in expectation at initialization to a stationary covariance. We show that sharing a single phase per frequency pair restores the exact harmonizable integrand. Building on this, we introduce Deep Operator-Valued Spectral Kernel Networks (DOSKNs), a primal framework that extends deep spectral learning to vector-valued outputs while decoupling spectral resolution, layer width, and subspace rank. Grouped output coupling enforces block-rank constraints, regularizing cross-dimension interactions and lowers parameter count. Furthermore, we introduce a trace-constrained squared group regularizer with an exact $\mathcal{O}(M \log M)$ proximal map. Empirical evaluations confirm that DOSKN addresses stationary collapse, cuts coupling parameters by up to 76%, and consistently matches or outperforms unconstrained models.

1 INTRODUCTION

Kernel machines provide a foundation for nonparametric function estimation, combining convex optimization with regularization theory (Schölkopf & Smola, 2001; Boyd & Vandenberghe, 2004; Rasmussen & Williams, 2005). However, standard stationary kernels (such as the Matérn or Squared Exponential kernels) assume shift-invariance, defined on the displacement $\tau = x - x'$. This spatial homogeneity limits their applicability when modeling complex processes exhibiting localized spectral dynamics and input-dependent behavior.

Harmonizable kernels address this limitation by defining covariance through a spectral bimeasure $F(\omega, \omega')$ over frequency pairs $(\omega, \omega') \in \mathbb{R}^d \times \mathbb{R}^d$ (Loeve, 1978; Yaglom, 1987; Remes et al., 2017). When F assigns mass off the diagonal $\omega = \omega'$, the resulting kernel models cross-frequency interactions that produce rich nonstationary covariances (Shen et al., 2019). Deep Spectral Kernel Networks (DSKNs) implement this representation by parameterizing paired-frequency feature maps and stacking them in hierarchical layers (Xue et al., 2019).

Nevertheless, existing spectral architectures remain scalar. For obtaining vector-valued outputs, scalar features are mapped through unconstrained linear transformations, or the output dimensions are modeled independently (Xue et al., 2019; Álvarez et al., 2012). This neglects structured dependencies between the input spectral components and the geometric subspaces of the output. In physical systems, distinct frequency bands excite distinct output modes. For instance, in multi-joint robotics, high-frequency vibrations localize to distal actuators, whereas low-frequency inertial forces drive whole-body kinematics. Unconstrained linear mixing needs to learn these relationships without geometric regularization, increasing parameter count.

Operator-valued kernels define matrix-valued covariances that generate vector-valued Reproducing Kernel Hilbert Spaces (RKHSs) (Micchelli & Pontil, 2005; Carmeli et al., 2010; Álvarez et al., 2012). While operator-valued random features provide explicit representations for stationary kernels (Brault et al., 2016; Minh, 2016), they remain restricted to shallow settings and do not model paired-frequency harmonizable dynamics. Similarly, algebraic generalizations such as Reproducing Kernel Hilbert C^*-modules (**?**) define shallow models.

We introduce Deep Operator-Valued Spectral Kernel Networks (DOSKNs), bridging primal deep representation learning and structured operator-valued covariance theory. We use the term *kernel*

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

network to denote a deep composition of explicit finite feature maps, each realizing a valid operator-valued kernel conditional on its parameters. We first analyze the stochastic phase map of (Xue et al., 2019) to demonstrate that independent phase sampling eliminates all nonstationary cross-frequency interactions in expectation at initialization, and show that a shared-phase construction restores the exact harmonizable integrand. We then define the DOSKN architecture, decouple its internal structural dimensions, parameterize its output coupling via block-rank constraints, and derive an exact amplitude-profiling regularizer.

Our contributions are: (1) We show that independent phase sampling in DSKNs eliminates nonstationary cross-frequency interactions in expectation at initialization, and that sharing phases within frequency pairs recovers the exact harmonizable integrand. (2) We define a finite PSD operator-valued layer that decouples spectral dictionary resolution (M_l), layer width (D_l), and subspace rank (r_l), removing the constraint $D_l = M_l$. (3) We show that grouped subspace sharing enforces a block-rank constraint ($\mathrm{rank}(C_{\mathcal{I}_q}) \leq r_l$), regularizing multi-dimensional output mixing. (4) We derive a group regularizer (Ω) for which we provide an efficient $\mathcal{O}(M \log M)$ exact proximal solver. (5) On synthetic benchmarks and real-world datasets, DOSKN significantly lowers output parameter count relative to unconstrained linear mixing and operator-valued baselines while having competitive predictive performance.

2 Related Work

Harmonizable and Nonstationary Spectral Kernels. Stationary kernels rely on Bochner's theorem, establishing an isomorphism between continuous shift-invariant kernels and Fourier transforms of positive measures (Rahimi & Recht, 2007; Rudi & Rosasco, 2017). To model nonstationarity, harmonizable kernels express covariance via a general spectral bimeasure F over frequency pairs $(\omega, \omega') \in \mathbb{R}^d \times \mathbb{R}^d$ (Loeve, 1978; Yaglom, 1987):

$$k(x, x') = \iint_{\mathbb{R}^d \times \mathbb{R}^d} e^{i(\omega^\top x - \omega'^\top x')} \mathrm{d}F(\omega, \omega'). \tag{1}$$

The off-diagonal spectral mass ($F(\omega, \omega') \neq 0$ for $\omega \neq \omega'$) produces nonstationary covariances depending on both $x - x'$ and $x + x'$ (Remes et al., 2017; Shen et al., 2019). Xue et al. (2019) implemented this in Deep Spectral Kernel Networks (DSKNs). However, as shown in Section 3, independent sampling of the phase variables eliminates off-diagonal terms in expectation at initialization.

Operator-Valued Kernels and Multi-Output Learning. Vector-valued function learning requires modeling dependencies over output coordinates (Micchelli & Pontil, 2005; Álvarez et al., 2012). The Linear Coregionalization Model parameterizes matrix-valued covariances as sums of separable products $K(x, x') = \sum_q B_q k_q(x, x')$ (Bonilla et al., 2007; Chilès & Delfiner, 2012). Operator-valued kernels generalize this by mapping input pairs to positive semidefinite operators $\mathcal{K}(x, x')$, generating vector-valued RKHSs (Carmeli et al., 2010; Brouard et al., 2016), which have also been extended to Hilbert C^*-modules (Hashimoto et al., 2021).

Operator-valued random Fourier features were derived for stationary translation-invariant kernels by Brault et al. (2016) and Minh (2016) for shallow kernel ridge regression. Operator-valued representations have also been explored algebraically via Reproducing Kernel Hilbert C^*-modules (Hashimoto et al., 2021). In multi-output Gaussian processes, spectral mixture kernels capture cross-channel phase delays in the dual form (Parra & Tobar, 2017; Altamirano & Tobar, 2022), and Kadri et al. (2012) formulated multiple operator-valued kernel learning (MOKL).

While exact dual inference in MOGPs has a cubic computational cost in the number of observations, scalable frameworks mitigate this cost using sparse inducing points (Álvarez & Lawrence, 2011; Moreno-Muñoz et al., 2018) or deep kernel learning (Wilson et al., 2016; Matias et al., 2024). DOSKN addresses vector-valued representation from a distinct perspective, translating operator-valued kernels into explicit deep nonstationary feature maps trained end-to-end via standard backpropagation.

3 PHASE ANALYSIS AND THE CORRECTED SPECTRAL MAP

Let $(\omega, \omega') \in \mathbb{R}^d \times \mathbb{R}^d$ be a frequency pair drawn from a symmetric bimeasure. The real-valued paired-frequency kernel component is:

$$\begin{aligned}\mathcal{T}_{\omega,\omega'}(x,x') &= \frac{1}{4}\mathrm{Re}\left\{\left[e^{i\omega^\top x} + e^{i\omega'^\top x}\right]\left[e^{-i\omega^\top x'} + e^{-i\omega'^\top x'}\right]\right\} \\ &= \frac{1}{4}\Big[\cos(\omega^\top(x-x')) + \cos(\omega^\top x - \omega'^\top x') + \cos(\omega'^\top x - \omega^\top x') + \cos(\omega'^\top(x-x'))\Big].\end{aligned} \tag{2}$$

The outer terms are stationary, depending solely on $\tau = x - x'$. The cross-terms $\cos(\omega^\top x - \omega'^\top x')$ and $\cos(\omega'^\top x - \omega^\top x')$ depend on x and x' individually, producing nonstationary covariance. Xue et al. (2019) introduced a two-coordinate deterministic map that matches the paired-frequency integrand:

$$\psi_{\omega,\omega'}(x) = \frac{1}{2}\begin{bmatrix}\cos(\omega^\top x) + \cos(\omega'^\top x) \\ \sin(\omega^\top x) + \sin(\omega'^\top x)\end{bmatrix}, \quad \text{where } \psi_{\omega,\omega'}(x)^\top \psi_{\omega,\omega'}(x') = \mathcal{T}_{\omega,\omega'}(x,x') \quad \forall x, x'. \tag{3}$$

To halve the feature dimension, they proposed a single-coordinate stochastic phase map:

$$\tilde{z}_{\omega,\omega'}(x; b, b') = \frac{1}{\sqrt{2}}\Big[\cos(\omega^\top x + b) + \cos(\omega'^\top x + b')\Big], \quad b, b' \sim \mathrm{Unif}[0, 2\pi]. \tag{4}$$

However, sampling b and b' independently leads to an expected stationary limit at random initialization. Appendix A.1 shows that, when $b, b' \sim \mathrm{Unif}[0, 2\pi]$ are independent random variables, the expected outer product for the feature map in equation 4 is:

$$\mathbb{E}_{b,b'}\left[\tilde{z}_{\omega,\omega'}(x; b, b')\tilde{z}_{\omega,\omega'}(x'; b, b')\right] = \frac{1}{4}\Big[\cos(\omega^\top(x-x')) + \cos(\omega'^\top(x-x'))\Big]. \tag{5}$$

That is, because b, b' are independent, the expectations of the cross-terms decouple, removing all cross-frequency interactions in expectation at initialization.

We restore the full harmonizable integrand by tying the phase variables within each frequency pair. Let $b \sim \mathrm{Unif}[0, 2\pi]$ be a single shared phase variable and the shared-phase paired feature map be defined as:

$$z_{\omega,\omega'}(x; b) = \frac{1}{\sqrt{2}}\Big[\cos(\omega^\top x + b) + \cos(\omega'^\top x + b)\Big]. \tag{6}$$

Then, as shown in Appendix A.2, the expectation over b recovers the complete four-term harmonizable integrand, i.e., $\mathbb{E}_b\left[z_{\omega,\omega'}(x; b) z_{\omega,\omega'}(x'; b)\right] = \mathcal{T}_{\omega,\omega'}(x, x')$. Under fixed spectral parameters, equation 6 defines an unbiased nonstationary estimator. When trained end-to-end via gradient descent, the phases and frequencies adapt to data, transitioning the architecture from a random feature approximation to supervised spectral dictionary learning over a parameterized family of RKHSs.

4 DEEP OPERATOR-VALUED SPECTRAL KERNEL NETWORKS

4.1 PRIMAL OPERATOR-VALUED LAYER FORMULATION

Let $h_{l-1} \in \mathbb{R}^{D_{l-1}}$ denote the input to layer $l \in \{1, \ldots, L\}$. The layer incorporates a dictionary of M_l paired-frequency components parameterized by frequencies $\omega_{lm}, \omega'_{lm} \in \mathbb{R}^{D_{l-1}}$ and shared phases $b_{lm} \sim \mathrm{Unif}[0, 2\pi]$. The scalar spectral feature for component m is given by:

$$z_{lm}(h_{l-1}) = \frac{1}{\sqrt{2M_l}}\Big[\cos(\omega_{lm}^\top h_{l-1} + b_{lm}) + \cos(\omega_{lm}'^\top h_{l-1} + b_{lm})\Big]. \tag{7}$$

We collect these scalar activations into the vector $z_l(h_{l-1}) = [z_{l1}(h_{l-1}), \ldots, z_{lM_l}(h_{l-1})]^\top \in \mathbb{R}^{M_l}$.

To construct an operator-valued mapping, we associate each component m with a positive semi-definite (PSD) output coupling operator $\bar{B}_{lm} \in \mathbb{R}^{D_l \times D_l}$ satisfying $\bar{B}_{lm} \succeq 0$ and $\mathrm{tr}(\bar{B}_{lm}) = 1$. Each matrix $\bar{B}_{lm}$ is factored as $\bar{L}_{lm}\bar{L}_{lm}^\top$ with $\bar{L}_{lm} \in \mathbb{R}^{D_l \times r_l}$ and $\|\bar{L}_{lm}\|_F^2 = 1$. Scaling by an amplitude $s_{lm} \geq 0$, the finite operator-valued kernel $\mathcal{K}_l : \mathbb{R}^{D_{l-1}} \times \mathbb{R}^{D_{l-1}} \to \mathbb{R}^{D_l \times D_l}$ is defined by:

$$\mathcal{K}_l(h_{l-1}, h'_{l-1}) = \sum_{m=1}^{M_l} s_{lm} z_{lm}(h_{l-1}) z_{lm}(h'_{l-1}) \bar{B}_{lm}, \tag{8}$$

which is PSD by construction as a sum of outer products. Any non-uniform frequency weights (such as importance sampling ratios or mixture proportions) can be absorbed into the component amplitudes s_{lm}.

4.2 Forward Evaluation and Block-Rank Structure

Under the operator-valued kernel $\mathcal{K}_l$, any vector-valued function $f_l : \mathbb{R}^{D_{l-1}} \to \mathbb{R}^{D_l}$ in the associated RKHS is parameterized by coordinate vectors $\alpha_{lm} \in \mathbb{R}^{r_l}$ as:

$$f_l(h_{l-1}) = \sum_{m=1}^{M_l} \sqrt{s_{lm}} z_{lm}(h_{l-1}) \bar{L}_{lm} \alpha_{lm}, \tag{9}$$

with squared RKHS norm $\|f_l\|^2_{\mathcal{H}_{\mathcal{K}_l}} = \sum_{m=1}^{M_l} \|\alpha_{lm}\|_2^2$. Reparameterizing the coefficients as $\beta_{lm} := \sqrt{s_{lm}}\alpha_{lm} \in \mathbb{R}^{r_l}$, letting $c_{lm} := \bar{L}_{lm}\beta_{lm} \in \mathbb{R}^{D_l}$ denote the effective output weight vector for component m. Assembling them as rows into the effective weight matrix $C_l \in \mathbb{R}^{M_l \times D_l}$ with $[C_l]_{m,:} = c_{lm}^\top = \beta_{lm}^\top \bar{L}_{lm}^\top$, the forward pass simplifies to $f_l(h_{l-1}) = C_l^\top z_l(h_{l-1})$. With unconstrained coefficients, positive amplitudes s_{lm} do not alter the span of realizable functions, but govern the RKHS geometry and regularization analyzed in Section 5.

Block-Rank Structure of Grouped Coupling. To understand the inductive bias imposed by the coupling factors $\bar{L}_{lm}$, we first consider the case where $Q_l = M_l$, i.e., assigning each component an independent factor $\bar{L}_{lm} \in \mathbb{R}^{D_l \times r_l}$. For any non-zero effective output weight $c_{lm} \in \mathbb{R}^{D_l}$, choosing $\bar{L}_{lm} = \frac{c_{lm}}{\|c_{lm}\|_2} e_1^\top$ and $\beta_{lm} = \|c_{lm}\|_2 e_1$ (where $e_1 \in \mathbb{R}^{r_l}$ is the first basis vector) satisfies the unit Frobenius norm constraint $\|\bar{L}_{lm}\|_F^2 = 1$ while realizing $c_{lm} = \bar{L}_{lm}\beta_{lm}$. Thus, unshared factors place no geometric restriction on individual output weights c_{lm} and just overparameterize free linear mixing.

Functional regularization arises when multiple components share geometric subspaces, i.e., when $Q_l \ll M_l$. We partition the M_l paired-frequency components into Q_l disjoint groups $\mathcal{I}_1, \ldots, \mathcal{I}_{Q_l} \subseteq \{1, \ldots, M_l\}$. All components in group q share an orthonormal basis $U_{l,q} \in \mathbb{R}^{D_l \times r_l}$ ($U_{l,q}^\top U_{l,q} = I_{r_l}$) via the scaled factor $\bar{L}_{lm} = \frac{1}{\sqrt{r_l}} U_{l,q}$, $\forall m \in \mathcal{I}_q$. For each component $m \in \mathcal{I}_q$, the effective output weight is $c_{lm}^\top = \frac{1}{\sqrt{r_l}} \beta_{lm}^\top U_{l,q}^\top$. Stacking $\{c_{lm}^\top\}_{m \in \mathcal{I}_q}$ results in the group submatrix $C_{\mathcal{I}_q} \in \mathbb{R}^{|\mathcal{I}_q| \times D_l}$:

$$C_{\mathcal{I}_q} = \begin{bmatrix} c_{l,i_1}^\top \\ \vdots \\ c_{l,i_{|\mathcal{I}_q|}}^\top \end{bmatrix} = \frac{1}{\sqrt{r_l}} \begin{bmatrix} \beta_{l,i_1}^\top \\ \vdots \\ \beta_{l,i_{|\mathcal{I}_q|}}^\top \end{bmatrix} U_{l,q}^\top = \frac{1}{\sqrt{r_l}} B_q U_{l,q}^\top, \tag{10}$$

where $B_q \in \mathbb{R}^{|\mathcal{I}_q| \times r_l}$ gathers the coefficient vectors as rows. Applying the rank inequality for matrix products:

$$\operatorname{rank}(C_{\mathcal{I}_q}) \le \min\left(\operatorname{rank}(B_q), \operatorname{rank}(U_{l,q}^\top)\right) \le \min(|\mathcal{I}_q|, r_l, D_l) \le r_l. \tag{11}$$

Whenever $r_l < \min(|\mathcal{I}_q|, D_l)$, the effective weight block $C_{\mathcal{I}_q}$ is strictly low-rank. Grouped coupling constrains the multi-dimensional responses excited by frequency group q to an r_l-dimensional output subspace. This enforces a low-rank constraint that removes parameter redundancy.

Group Assignment and Orthonormal Subspace Enforcement. To avoid non-differentiable clustering during training, components are assigned to groups at initialization via frequency-norm quantile binning. For each component m, we calculate $\nu_{lm} = \|\omega_{lm}\|_2 + \|\omega'_{lm}\|_2$, sort components by ν_{lm}, and partition them into Q_l balanced bins. The resulting assignment $g_l(m) \in \{1, \ldots, Q_l\}$ is stored as a non-trainable label. Each subspace basis $U_{l,q}$ is initialized with orthonormal columns via QR decomposition. During end-to-end training, gradient updates are followed by QR retraction with positive diagonal normalization on R, preserving column orthonormality ($U_{l,q}^\top U_{l,q} = I_{r_l}$) throughout optimization.

4.3 Output-Coupling Hierarchy and Trace Recovery

Table 1 outlines the spectrum of output coupling geometries supported by DOSKN, ranging from unconstrained full-rank factorizations to rigid rank-one and dimension-specific mappings. Grouped

Table 1: Normalized DOSKN Output Coupling Geometries.

Model	Factorization $\overline{L}_{lm}$	Rank	Params	Geometric Semantics
Uniform	$\frac{1}{\sqrt{D_l}}\mathbf{1}_{D_l}$	1	M_l	Isotropic uniform scaling
Diagonal	$\mathrm{diag}(\sqrt{\pi_{lm}})$	D_l	$M_l D_l$	Axis-aligned diagonal responses
Subspace-r	$\frac{1}{\sqrt{r_l}}U_{l,g_l(m)}$	r_l	$Q_l D_l r_l + M_l r_l$	Tied ($Q=1$), Grouped ($Q=Q_l$), or Free ($Q=M_l$)
Anisotropic	$U_{lm}\,\mathrm{diag}(\sqrt{\pi_{lm}})$	r_l	$M_l D_l r_l + 2M_l r_l$	Rotated low-rank with internal variances

Enforcing $\mathrm{tr}(B_{lm})=1$, $\pi_{lm}\geq 0$ with $\sum_k \pi_{lm,k}=1$ denotes a normalized coordinate variance vector.

coupling saves parameters over unconstrained linear mixing ($M_l D_l$) only when $r_l < \frac{M_l D_l}{Q_l D_l + M_l}$. At $r_l = D_l$, the rank constraint disappears, making the layer slightly larger than free mixing.

Canonical Coordinate Embedding and Trace Recovery. Scalar DSKN is recovered exactly as a fixed-parameter embedding of the feature map. Specifically, let $M_l = D_l$, fix unit amplitudes $s_{lm} = 1$, and configure DOSKN with coordinate-wise coupling using rank-one basis factors $\overline{L}_{lm} = e_m \in \mathbb{R}^{D_l \times 1}$ ($r_l = 1$) with fixed unit scalar coefficients $\beta_{lm} = 1 \in \mathbb{R}$. Substituting these into equation 9 leads to $f_l(h_{l-1}) = \sum_{m=1}^{D_l} z_{lm}(h_{l-1}) e_m = z_l(h_{l-1})$, recovering the corrected DSKN activation map. Furthermore, each output coupling operator evaluates to $\overline{B}_{lm} = e_m e_m^\top$, leading to the diagonal operator-valued kernel $\mathcal{K}_l^{\mathrm{diag}}(h_{l-1}, h'_{l-1}) = \mathrm{diag}\big(z_{l1}(h_{l-1}) z_{l1}(h'_{l-1}), \ldots, z_{lD_l}(h_{l-1}) z_{lD_l}(h'_{l-1})\big)$. Taking the matrix trace recovers the scalar DSKN kernel:

$$\mathrm{tr}\big(\mathcal{K}_l^{\mathrm{diag}}(h_{l-1}, h'_{l-1})\big) = \sum_{m=1}^{D_l} z_{lm}(h_{l-1}) z_{lm}(h'_{l-1}) = z_l(h_{l-1})^\top z_l(h'_{l-1}). \tag{12}$$

Thus, while DOSKN coefficient vectors β_{lm} are generally learned parameters in vector-valued RKHSs, the corrected DSKN is realized as an unrotated, coordinate-wise canonical embedding.

4.4 Network Architecture and Computational Complexity

A depth-L DOSKN is defined by recursive composition $h_0 = x,\ h_l = f_l(h_{l-1})$, where each f_l is a selected function in a finite operator-valued RKHS conditional on its layer parameters. Each layer parameterizes frequencies $\omega_{lm}, \omega'_{lm} \in \mathbb{R}^{D_{l-1}}$ ($2M_l D_{l-1}$ parameters), shared phase offsets $b_{lm} \in \mathbb{R}$ (M_l parameters), subspace basis matrices $U_{l,q} \in \mathbb{R}^{D_l \times r_l}$ ($Q_l D_l r_l$ parameters), and coefficient vectors $\beta_{lm} \in \mathbb{R}^{r_l}$ ($M_l r_l$ parameters). Since component amplitudes s_{lm} are absorbed into $\beta_{lm} = \sqrt{s_{lm}}\alpha_{lm}$, and layers operate without additive output bias vectors, the total layer parameter count is $P_l = 2M_l D_{l-1} + M_l + Q_l D_l r_l + M_l r_l$, accounting for all learnable spectral, phase, and output parameters.

For an input batch of size N, paired projections require $\mathcal{O}(NM_l D_{l-1})$ operations, assembling coefficient matrices requires $\mathcal{O}(M_l D_l r_l)$ operations, forward activation evaluation requires $\mathcal{O}(NM_l D_l)$, and maintaining orthonormality via QR factorizations on the Q_l group bases adds a training overhead of $\mathcal{O}(Q_l D_l r_l^2)$. At test time, the effective output weights $C_l \in \mathbb{R}^{M_l \times D_l}$ are static and cached. This reduces inference complexity to $\mathcal{O}(NM_l D_{l-1} + NM_l D_l)$ independent of internal rank r_l.

5 Optimization and Group Regularization

5.1 Fixed-Dictionary RKHS Norm and Amplitude Optimization

Let the spectral dictionary $\{(\omega_{lm}, \omega'_{lm}, b_{lm})\}_{m=1}^M$ and normalized output coupling operators $\{\bar{B}_{lm}\}_{m=1}^M$ be fixed. The vector-valued RKHS $\mathcal{H}_{\mathcal{K}_l}$ associated with the kernel $\mathcal{K}_l(h_{l-1}, h'_{l-1}) = \sum_m s_{lm} z_{lm}(h_{l-1}) z_{lm}(h'_{l-1}) \bar{B}_{lm}$ possesses the minimum-norm quotient representation (Carmeli et al., 2010):

$$\|f\|^2_{\mathcal{H}_{\mathcal{K}_l}} = \inf_{\{\beta_{lm}\}} \left\{ \sum_{m: s_{lm}>0} \frac{\|\beta_{lm}\|_2^2}{s_{lm}} \ :\ f = \sum_{m=1}^M z_{lm}(\cdot) \bar{L}_{lm} \beta_{lm} \right\}. \tag{13}$$

We adopt the extended-value convention (Boyd & Vandenberghe, 2004):

$$\frac{\|\beta_{lm}\|_2^2}{s_{lm}} = \begin{cases} 0, & \text{if } s_{lm} = 0 \text{ and } \beta_{lm} = 0, \\ +\infty, & \text{if } s_{lm} = 0 \text{ and } \beta_{lm} \neq 0. \end{cases} \tag{14}$$

As shown in Appendix A.3, optimizing component amplitudes under a simplex trace budget $\sum_{m=1}^{M} s_{lm} = \tau$ with $(s_{lm} \geq 0)$ leads to an exact closed-form solution:

$$\inf_{\substack{s_{lm} \geq 0 \\ \sum_{m=1}^{M} s_{lm} = \tau}} \sum_{m=1}^{M} \frac{\|\beta_{lm}\|_2^2}{s_{lm}} = \frac{1}{\tau} \left(\sum_{m=1}^{M} \|\beta_{lm}\|_2 \right)^2 =: \Omega(\beta_l), \tag{15}$$

and the unique optimal amplitude for component m is given by $s_{lm}^* = \tau \frac{\|\beta_{lm}\|_2}{\sum_{j=1}^{M_l} \|\beta_{lj}\|_2}$.

Connection to Multiple Kernel Learning. Equation 15 connects directly to the MKL formulations of Micchelli & Pontil (2005) and Bach (2008). Rather than solving a bilevel optimization problem with explicit simplex variables ρ_m (Kadri et al., 2012), DOSKN optimizes β_l directly under the regularizer $\Omega(\beta_l)$. Furthermore, for orthonormal factors $\bar{L}_{lm} = \frac{1}{\sqrt{r}} U_{lm}$, because each U_{lm} preserves Euclidean norm i.e., $\|U_{lm}\beta_{lm}\|_2 = \|\beta_{lm}\|_2$, the effective output weights satisfy $\|c_{lm}\|_2 = \frac{1}{\sqrt{r}} \|\beta_{lm}\|_2$. Thus, the regularizer satisfies $\Omega(\beta_l) = \frac{1}{\tau} \left(\sum_{m=1}^{M} \|\beta_{lm}\|_2 \right)^2 = \frac{r}{\tau} \left(\sum_{m=1}^{M} \|c_{lm}\|_2 \right)^2 = \frac{r}{\tau} \|C_l\|_{2,1}^2$, showing that optimization of component scales under a trace budget directly penalizes the squared row-group norm of the effective output weight matrix C.

5.2 Optimization and Training Settings

To optimize objectives regularized by $\Omega(\beta_l)$, we derive its exact proximal operator. For step size $\gamma > 0$ and input $V_l \in \mathbb{R}^{M_l \times r_l}$, the proximal operator is defined as:

$$\text{prox}_{\gamma\Omega}(V_l) = \arg\min_{\beta_l} \frac{1}{2} \sum_{m=1}^{M} \|\beta_{lm} - V_{lm}\|_2^2 + \frac{\gamma}{\tau} \left(\sum_{m=1}^{M} \|\beta_{lm}\|_2 \right)^2. \tag{16}$$

Utilizing piecewise linear root-finding over sorted group norms, Algorithm 1 in Appendix A.4 evaluates this operator exactly with $\mathcal{O}(M \log M)$ complexity.

We consider three training settings: (1) Fixed-Kernel based Learning: spectral features and output coupling are frozen, leading to a convex optimization problem for convex empirical losses, (2) Fixed-Dictionary based MOKL: amplitudes are learned implicitly through $\Omega(\beta_l)$, and (3) Joint Nonconvex Dictionary Learning: frequencies, phases, subspace bases, and coefficients are trained end-to-end via gradient updates.

6 Experiments

We evaluate DOSKN on controlled synthetic tasks and real-world datasets. Our experiments evaluate four properties: (1) restoring the harmonizable integrand via shared-phase sampling and verifying the phase gap when using gradient descent, (2) output subspace recovery and parameter efficiency, (3) predictive accuracy and parameter count reduction, and (4) decoupling of dictionary resolution and layer width, under exact proximal regularization. Architectural configurations and extended experimental results are provided in Appendix B.

6.1 Phase Configuration and Harmonizable Kernel Recovery

We first evaluate the properties established in Section 3 on kernel approximation and regression performance within the target RKHS. We compare four feature representations, i.e., the deterministic two-coordinate map ($2M$ coordinates), shared-phase paired features (M coordinates), independent-phase paired features (M coordinates), and stationary random Fourier features (M coordinates, scaled with amplitude $1/\sqrt{M}$ to match the marginal spectrum). We draw inputs $X \sim \mathcal{N}(0, I_{d_{\text{in}}})$

Table 2: Kernel recovery error and nonstationarity diagnostics for dictionary size M (mean $\pm$ std for 5 seeds).

Method	M	$E_{\text{full}} \downarrow$	$E_{\text{cross}} \downarrow$	$S_{\text{nonstat}} \to 1$	**Scale**
Deterministic ($2M$)	64	0.0000 ± 0.0000	0.0000 ± 0.0000	1.0000 ± 0.0000	1.88
Shared-Phase (M)	64	0.1268 ± 0.0221	0.2613 ± 0.0456	1.0459 ± 0.1566	1.90
Independent-Phase (M)	64	0.4578 ± 0.0581	0.9428 ± 0.1189	0.2776 ± 0.0575	1.08
Stationary RFF (M)	64	0.5008 ± 0.0193	1.0313 ± 0.0377	0.2256 ± 0.0076	0.99
Shared-Phase (M)	256	0.0560 ± 0.0020	0.1158 ± 0.0042	0.9827 ± 0.0255	1.86
Independent-Phase (M)	256	0.4623 ± 0.0331	0.9556 ± 0.0689	0.1447 ± 0.0325	1.05
Stationary RFF (M)	256	0.4876 ± 0.0185	1.0077 ± 0.0379	0.1150 ± 0.0085	1.00
Shared-Phase (M)	1024	$\mathbf{0.0331} \pm 0.0056$	$\mathbf{0.0685} \pm 0.0116$	$\mathbf{0.9935} \pm 0.0420$	1.88
Independent-Phase (M)	1024	0.4772 ± 0.0249	0.9856 ± 0.0511	0.0732 ± 0.0122	1.01
Stationary RFF (M)	1024	0.4873 ± 0.0125	1.0065 ± 0.0255	0.0598 ± 0.0049	1.00

Table 3: Fixed-dictionary coupling hierarchy on synthetic data (mean $\pm$ std for 5 seeds).

Coupling Mode	**Test nRMSE** $\downarrow$	**Coupling Params**	**Latency (ms)**
Free Linear Mixing	0.4661 ± 0.0207	4096	0.114
Uniform Rank-1	0.9959 ± 0.0102	256	0.113
Diagonal Coupling	0.5788 ± 0.0126	4096	0.112
Tied Subspace ($Q = 1, r = 3$)	0.7467 ± 0.0306	816	0.140
Grouped Subspace ($Q = 4, r = 3$)	$\mathbf{0.4534} \pm 0.0260$	**960**	0.137
Per-Component ($Q = 256, r = 3$)	0.4453 ± 0.0217	13056	0.112

with $d_{\text{in}} = 8$, $N = 600$, and sample frequencies $\omega_m, \omega'_m \sim \mathcal{N}(0, I_{d_{\text{in}}}/d_{\text{in}})$. For a dictionary of M frequency pairs, we denote by $\mathcal{K}_{\text{full}} \in \mathbb{R}^{N\times N}$ the exact harmonizable Gram matrix $[\mathcal{K}_{\text{full}}]_{ij} = \frac{1}{M}\sum_{m=1}^{M} \mathcal{T}_{\omega_m,\omega'_m}(x_i, x_j)$, and by $\mathcal{K}_{\text{stat}} \in \mathbb{R}^{N\times N}$ the stationary component corresponding to the expectation of independent phases in equation 5. Then, $\mathcal{K}_{\text{cross}} = \mathcal{K}_{\text{full}} - \mathcal{K}_{\text{stat}}$ isolates the nonstationary cross-terms. In Table 2, we evaluate the errors $E_{\text{full}} = \|\mathcal{K}_{\text{full}} - \hat{\mathcal{K}}\|_F / \|\mathcal{K}_{\text{full}}\|_F$ and $E_{\text{cross}} = \|\mathcal{K}_{\text{cross}} - (\hat{\mathcal{K}} - \mathcal{K}_{\text{stat}})\|_F / \|\mathcal{K}_{\text{cross}}\|_F$, the nonstationarity preservation ratio $S_{\text{nonstat}} = \|\hat{\mathcal{K}} - \mathcal{K}_{\text{stat}}\|_F / \|\mathcal{K}_{\text{cross}}\|_F$, and the scale expressing the ratio $\langle \hat{\mathcal{K}}, \mathcal{K}_{\text{stat}} \rangle / \|\mathcal{K}_{\text{stat}}\|_F^2$ for 5 random seeds. While the deterministic map matches $\mathcal{K}_{\text{full}}$ exactly, the shared-phase map achieves consistent error decay driving E_{full} from 0.1268 ($M = 64$) down to 0.0331 ($M = 1024$) while maintaining $S_{\text{nonstat}} \approx 1.0$. On the other hand, the independent-phase map remains trapped at the ranges of ($E_{\text{full}} \to 0.46$, $E_{\text{cross}} \to 0.96$), with S_{nonstat} vanishing to 0.0732.

6.2 Coupling Hierarchy and Subspace Recovery

We evaluate the output-coupling geometries from Section 4.3 on synthetic data generated from a known block-rank ground truth ($N = 1500, d_{\text{in}} = 8, D_{\text{out}} = 16, Q^* = 4, r^* = 3, M = 256$, noise $\sigma = 0.05$). Table 3 shows test nRMSE of the different coupling modes. Grouped subspace coupling matches the performance of free linear mixing while lowering coupling parameters from 4096 to 960 (a 76.6% reduction). Uniform rank-1 and tied subspace models fail due to underparameterization, while diagonal coupling cannot model cross-dimensional correlations. All models evaluate within 0.11–0.14 ms.

As discussed in Section 4.2, grouped coupling acts as a regularizer whose advantage emerges when sample observations are constrained relative to output dimensions. Table 4 reports results for $M = 256$, sample sizes $N \in \{125, 250, 500, 1000\}$, and output dimensions $D_{\text{out}} \in \{16, 64\}$. Grouped coupling consistently leads to lower error, with the difference between the performance of linear and grouped coupling increases for lower training data sizes.

6.3 Regression Performance Evaluation

We evaluate DOSKN on regression tasks and compare it to Multilayer Perceptron (MLP), stationary RFF network, original DSKN with independent phases, and corrected DSKN with unconstrained free linear mixing. For fair evaluation, the hyperparameters of each method are selected by ap-

Table 4: Test nRMSE delta ($\Delta = \text{Grouped} - \text{Free}$) for training size N and output dimension D_{out} on synthetic data (mean $\pm$ std for 5 seeds).

N	D_{out}	Free nRMSE	Grouped nRMSE	$\Delta \downarrow$
125	16	0.6382	0.5742	$\mathbf{-0.0640} \pm 0.0420$
250	16	0.5360	0.5064	$\mathbf{-0.0295} \pm 0.0309$
500	16	0.4838	0.4639	$\mathbf{-0.0199} \pm 0.0184$
1000	16	0.4600	0.4457	$\mathbf{-0.0143} \pm 0.0114$
125	64	0.8575	0.7829	$\mathbf{-0.0746} \pm 0.0149$
250	64	0.7944	0.7316	$\mathbf{-0.0628} \pm 0.0105$
500	64	0.7473	0.7015	$\mathbf{-0.0457} \pm 0.0059$
1000	64	0.7152	0.6867	$\mathbf{-0.0285} \pm 0.0015$

Table 5: Regression performance evaluation (mean $\pm$ std across 5 seeds/splits). All models are selected on validation data. Coupling parameters denote weights in the output mapping.

Model Family	**SARCOS**	**Parkinson's**
MLP	0.1007 ± 0.0037	1.2699 ± 0.2495
Stationary RFF	0.0989 ± 0.0008	1.3012 ± 0.3488
Original DSKN (independent)	$\mathbf{0.0912} \pm 0.0014$	1.2466 ± 0.2493
Corrected DSKN (free mixing)	0.0916 ± 0.0011	1.2900 ± 0.2993
Proposed DOSKN	0.0923 ± 0.0002	$\mathbf{1.1655} \pm 0.0801$
DOSKN Coupling Parameters	1,704 ($r = 3$)	258–528 ($r \in \{1, 2\}$)
Free Mixing Coupling Parameters	3,584 ($M = 512$)	512 ($M = 256$)
Coupling Parameter Reduction	**52.5%**	**Up to 49.6%**

plying a grid search on the validation set. For DOSKN, we restrict the candidate subspace rank to low-rank geometries $r \in \{1, 2, 3\}$ across group counts $Q \in \{1, 2, 3, 4, 6, 8\}$ to directly evaluate the inductive bias of structured, parameter-compressed output coupling. We use the SARCOS Robot Arm Dynamics dataset describing the inverse dynamics of a seven-joint anthropomorphic manipulator ($D_{\text{in}} = 21, D_{\text{out}} = 7$), and the Parkinson's Telemonitoring dataset tracking clinical progression scores ($D_{\text{in}} = 19, D_{\text{out}} = 2$). For the SACROS dataset we train on the canonical split ($N_{\text{tr}} = 37{,}812, N_{\text{val}} = 6{,}672, N_{\text{te}} = 4{,}449$) with hyperparameter selection based on validation MSE. To prevent subject leakage and evaluate true cross-patient transfer on the Parkinson's Telemonitoring dataset, evaluation uses 5-fold subject-isolated GroupKFold splits across 42 patients.

Table 5 summarizes regression performance. On SARCOS, nonstationary spectral methods substantially outperform MLPs and stationary RFFs. DOSKN achieves a similar performance with unconstrained free linear mixing while reducing output coupling parameters by 52.5%. On Parkinson's telemonitoring, strict subject isolation creates a severe estimation-limited setting where unregularized nonlinear models overfit individual patient variance. DOSKN's block-rank regularizer suppresses this multi-output overfitting, achieving the best performance while cutting coupling parameters by up to 49.6% when selecting $r = 1$.

6.4 Depth, Architectural Decoupling, and Optimization Ablations

We isolate the components of DOSKN using random subsets of the SARCOS robot dynamics dataset ($N_{\text{sub}} = 5000, d_{\text{in}} = 21, D_{\text{out}} = 7$). We evaluate DOSKNs for different depths ($L \in \{1, 2, 3\}$) while independently ablating residual connections and LayerNorm. To verify that performance gains do not arise just from linear skip parameters, we include a skip-only control with all spectral layers removed. Table 6 shows that performance degrades without normalization ($L = 2$ reaches 0.1721, $L = 3$ degrades to 0.1968). Introducing LayerNorm restores activation scales, while residual paths preserve stable gradient flow, making the $L = 2$ network to achieve the best performance. The skip-only control remains almost unchanged, confirming that nonstationary spectral layers drive the performance gain.

In original DSKNs, layer width is constrained to the number of frequency pairs ($D_l = M_l$). DOSKN decouples these dimensions. In Table 7, we evaluate a two-layer network ($21 \to D \to 7$) on a grid of dictionary sizes $M_1 \in \{64, 128, 256, 512\}$ and layer widths $D \in \{16, 64, 128\}$. The

Table 6: Depth, residual, and normalization ablation on SARCOS (3 seeds).

Architecture	Residual	LayerNorm	Test nRMSE ↓	Coupling P	Total P
$L = 1$: $21 \to 7$	False	False	0.1525 ± 0.0028	831	11839
$L = 1$: $21 \to 7$	True	False	0.1531 ± 0.0029	831	11986
Control: skip path only	—	—	0.3431 ± 0.0006	0	147
$L = 2$: $21 \to 64 \to 7$	False	False	0.1721 ± 0.0037	2879	46911
$L = 2$: $21 \to 64 \to 7$	False	True	0.1568 ± 0.0024	2879	47039
$L = 2$: $21 \to 64 \to 7$	True	False	0.1610 ± 0.0018	2879	48703
$L = 2$: $21 \to 64 \to 7$	True	True	$\mathbf{0.1432} \pm 0.0020$	2879	48831
Control: skip path only	—	—	0.3464 ± 0.0010	0	1792
$L = 3$: $21 \to 128 \to 64 \to 7$	False	False	0.1968 ± 0.0120	9023	118847
$L = 3$: $21 \to 128 \to 64 \to 7$	False	True	0.1624 ± 0.0041	9023	119231
$L = 3$: $21 \to 128 \to 64 \to 7$	True	False	0.1677 ± 0.0054	9023	130175
$L = 3$: $21 \to 128 \to 64 \to 7$	True	True	0.1472 ± 0.0014	9023	130559
Control: skip path only	—	—	0.3495 ± 0.0016	0	11328

Table 7: Test nRMSE $\pm$ std (for 3 seeds) and total parameter count for different values of M and D. Asterisks (*) mark the constrained $M_1 = D$ configurations required by original DSKN.

$M_1 \backslash D$	$D = 16$	$D = 64$	$D = 128$
$M_1 = 64$	0.1516 ± 0.0016 (13k)	$0.1461 \pm 0.0028^*$ (41k)	0.1553 ± 0.0021 (77k)
$M_1 = 128$	0.1511 ± 0.0031 (16k)	0.1458 ± 0.0040 (44k)	$0.1520 \pm 0.0029^*$ (81k)
$M_1 = 256$	0.1479 ± 0.0009 (23k)	0.1447 ± 0.0014 (50k)	0.1464 ± 0.0019 (87k)
$M_1 = 512$	$\mathbf{0.1415} \pm 0.0011$ (36k)	0.1423 ± 0.0013 (63k)	0.1470 ± 0.0017 (100k)

results show that expanding layer width without expanding spectral resolution increases parameter count without improving accuracy. Increasing spectral resolution while keeping layer width compact ($M_1 = 512, D = 16$) achieves the lowest error using only 36k parameters.

Finally, we evaluate Contribution (4) on the convex fixed-dictionary problem where the theoretical guarantees for Ω and Algorithm 1 hold. With spectral features and subspace bases U frozen, we optimize β using Algorithm 1 or with gradient descent for the same number of iterations (400 steps) and regularizer strengths $\gamma \in \{10^{-4}, 10^{-3}, 10^{-2}\}$. Table 8 confirms that Algorithm 1 converges to lower objective values for all regularization strengths. Extended analysis is provided in Appendix B.3.

Table 8: Exact proximal operator (Algorithm 1) versus gradient descent (GD) (mean for 3 seeds).

γ	Optimization Method	Final Objective ↓	Test nRMSE ↓
10^{-4}	Algorithm 1	**0.6015**	**0.3810**
	GD	1.3954	0.6336
10^{-3}	Algorithm 1	**1.0327**	**0.4770**
	GD	1.5159	0.6422
10^{-2}	Algorithm 1	**1.8142**	**0.6387**
	GD	2.0987	0.7111

7 Conclusion

We addressed a degeneracy in deep spectral kernel networks, i.e., that independent phase sampling causes the feature covariance to collapse into a stationary kernel at initialization. By tying the phase offsets within each frequency pair, the shared-phase feature map restores the harmonizable integrand and provides an unbiased estimator of nonstationary covariance. We then proposed Deep Operator-Valued Spectral Kernel Networks which extend deep spectral learning to vector-valued outputs by decoupling dictionary resolution (M_l), layer width (D_l), and subspace rank (r_l), removing the rigid $D_l = M_l$ constraint of Deep Spectral Kernel Networks. Grouping frequencies into quantile-aligned subspaces enforces a block-rank inductive bias that regularizes cross-dimension dependencies and cuts redundant parameters. Finally, variational amplitude profiling under a trace budget provides an exact squared group regularizer Ω, solved efficiently via a closed-form proximal operator.

AI USE STATEMENT

(This section is required and does not count toward the page limit.)

In this work, we used generative AI tools to create and edit code, and as an additional way to find related work. We have not used generative AI tools to generate synthetic datasets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, assist with translation, clean and reformat dataset, support qualitative and thematic data analysis, interpret results. We have reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness, and it is runnable in the uploaded supplementary material.

REPRODUCIBILITY STATEMENT

(This section is recommended and does not count toward the page limit.)

We uploaded the code to reproduce the results in the paper as supplementary materials.

REFERENCES

Matías Altamirano and Felipe Tobar. Nonstationary multi-output Gaussian processes via harmonizable spectral mixtures. *International Conference on Artificial Intelligence and Statistics*, 2022.

Mauricio A Álvarez and Neil D Lawrence. Computationally efficient multi-output Gaussian processes. *Journal of Machine Learning Research*, 12(41):1459–1500, 2011.

Mauricio A Álvarez, Lorenzo Rosasco, and Neil D Lawrence. Kernels for vector-valued functions: A review. *Foundations and Trends® in Machine Learning*, 4(3):195–266, 2012.

Francis R Bach. Consistency of the group lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9:1179–1225, 2008.

Edwin V Bonilla, Kian MA Chai, and Christopher KI Williams. Multi-task Gaussian process prediction. In *Advances in Neural Information Processing Systems*, 2007.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

Romain Brault, Markus Heinonen, and Florence d'Alché Buc. Random Fourier features for operator-valued kernels. In *Asian Conference on Machine Learning*, 2016.

Céline Brouard, Marie Szafranski, and Florence d'Alché Buc. Input Output Kernel Regression: Supervised and Semi-Supervised Structured Output Prediction with Operator-Valued Kernels. *Journal of Machine Learning Research*, 17:1–48, 2016.

Claudio Carmeli, Ernesto De Vito, Alessandro Toigo, and Veronica Umanità. Vector valued reproducing kernel Hilbert spaces and universality. *Analysis and Applications*, 8(1):19–61, 2010.

Jean-Paul Chilès and Pierre Delfiner. *Geostatistics: Modeling Spatial Uncertainty*. John Wiley & Sons, 2012.

Yuka Hashimoto, Isao Ishikawa, Masahiro Ikeda, Fuyuta Komura, Takeshi Katsura, and Yoshinobu Kawahara. Reproducing kernel hilbert C^*-module and kernel mean embeddings. *Journal of Machine Learning Research*, 22(267):1–56, 2021.

Hachem Kadri, Alain Rakotomamonjy, Philippe Preux, and Francis R Bach. Multiple operator-valued kernel learning. In *Advances in Neural Information Processing Systems*, 2012.

Michel Loeve. *Probability Theory II*. Springer-Verlag, New York, 4th edition, 1978.

Alan LS Matias, Cesar L Mattos, Joao PP Gomes, and Diego Mesquita. Amortized variational deep kernel learning. In *International Conference on Machine Learning*, 2024.

Charles A Micchelli and Massimiliano Pontil. On learning vector-valued functions. *Neural Computation*, 17(1):177–204, 2005.

Hà Quang Minh. Operator-valued Bochner theorem, Fourier feature maps for operator-valued kernels, and vector-valued learning. *arXiv:1608.05639*, 2016.

Pablo Moreno-Muñoz, Antonio Artés-Rodríguez, and Mauricio A Álvarez. Heterogeneous multi-output gaussian process prediction. In *Advances in Neural Information Processing Systems*, 2018.

Gabriel Parra and Felipe Tobar. Spectral mixture kernels for multi-output Gaussian processes. In *Advances in Neural Information Processing Systems*, 2017.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, 2007.

Carl Edward Rasmussen and Christopher KI Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA, 2005.

Sami Remes, Markus Heinonen, and Samuel Kaski. Non-stationary spectral kernels. In *Advances in Neural Information Processing Systems*, 2017.

Alessandro Rudi and Lorenzo Rosasco. Generalization properties of learning with random features. In *Advances in Neural Information Processing Systems*, 2017.

Bernhard Schölkopf and Alexander J Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, 2001.

Zheyang Shen, Markus Heinonen, and Samuel Kaski. Harmonizable mixture kernels with variational Fourier features. In *International Conference on Artificial Intelligence and Statistics*, 2019.

Andrew Gordon Wilson, Zhiting Hu, Ruslan Salakhutdinov, and Eric P Xing. Deep kernel learning. In *International Conference on Artificial Intelligence and Statistics*, 2016.

Hui Xue, Zheng-Fan Wu, and Wei-Xiang Sun. Deep spectral kernel learning. In *International Joint Conference on Artificial Intelligence*, 2019.

Akiva M Yaglom. *Correlation Theory of Stationary and Related Random Functions*. Springer-Verlag, New York, 1987.

A Mathematical Analyses

A.1 Stationary Expectation under Independent Phases

Let $(\omega, \omega') \in \mathbb{R}^d \times \mathbb{R}^d$ and $x, x' \in \mathbb{R}^d$. The stochastic feature map with two independent phase variables is defined as:

$$\tilde{z}(x; b, b') = \frac{1}{\sqrt{2}} \left[\cos(\omega^\top x + b) + \cos(\omega'^\top x + b') \right]. \tag{17}$$

Multiplying $\tilde{z}(x; b, b')\tilde{z}(x'; b, b')$ we get four terms:

$$\tilde{z}(x; b, b')\tilde{z}(x'; b, b') = \frac{1}{2} \left[I_{11} + I_{22} + I_{12} + I_{21} \right], \tag{18}$$

where:

$$\begin{aligned} I_{11} &= \cos(\omega^\top x + b)\cos(\omega^\top x' + b), \\ I_{22} &= \cos(\omega'^\top x + b')\cos(\omega'^\top x' + b'), \\ I_{12} &= \cos(\omega^\top x + b)\cos(\omega'^\top x' + b'), \\ I_{21} &= \cos(\omega'^\top x + b')\cos(\omega^\top x' + b). \end{aligned}$$

Using the identity $\cos(A)\cos(B) = \frac{1}{2}[\cos(A - B) + \cos(A + B)]$:

$$\begin{aligned} I_{11} &= \frac{1}{2}\cos(\omega^\top (x - x')) + \frac{1}{2}\cos(\omega^\top (x + x') + 2b), \\ I_{22} &= \frac{1}{2}\cos(\omega'^\top (x - x')) + \frac{1}{2}\cos(\omega'^\top (x + x') + 2b'). \end{aligned}$$

Since $b \sim \mathrm{Unif}[0, 2\pi]$, the oscillating terms vanish in expectation:

$$\mathbb{E}_b[\cos(\omega^\top (x + x') + 2b)] = \frac{1}{2\pi} \int_0^{2\pi} \cos(\omega^\top (x + x') + 2b) \mathrm{d}b = 0.$$

Thus, $\mathbb{E}_b[I_{11}] = \frac{1}{2}\cos(\omega^\top (x - x'))$ and $\mathbb{E}_{b'}[I_{22}] = \frac{1}{2}\cos(\omega'^\top (x - x'))$.

For the cross-term I_{12}, because b and b' are independent random variables, the joint expectation becomes:

$$\begin{aligned} \mathbb{E}_{b,b'}[I_{12}] &= \mathbb{E}_{b,b'}[\cos(\omega^\top x + b)\cos(\omega'^\top x' + b')] \\ &= \left(\frac{1}{2\pi} \int_0^{2\pi} \cos(\omega^\top x + b) \mathrm{d}b \right) \left(\frac{1}{2\pi} \int_0^{2\pi} \cos(\omega'^\top x' + b') \mathrm{d}b' \right) = 0 \cdot 0 = 0. \end{aligned}$$

Similarly, $\mathbb{E}_{b,b'}[I_{21}] = 0$. Summing the expectations of all four terms yields:

$$\mathbb{E}_{b,b'}\left[\tilde{z}(x; b, b')\tilde{z}(x'; b, b')\right] = \frac{1}{4} \left[\cos(\omega^\top (x - x')) + \cos(\omega'^\top (x - x')) \right], \tag{19}$$

which depends only on the spatial displacement $x - x'$.

A.2 Shared-Phase Unbiasedness

For a single shared phase variable $b \sim \mathrm{Unif}[0, 2\pi]$, the cross-terms in the product expansion become:

$$\begin{aligned} I_{12} &= \cos(\omega^\top x + b)\cos(\omega'^\top x' + b) = \frac{1}{2}\cos(\omega^\top x - \omega'^\top x') + \frac{1}{2}\cos(\omega^\top x + \omega'^\top x' + 2b), \\ I_{21} &= \cos(\omega'^\top x + b)\cos(\omega^\top x' + b) = \frac{1}{2}\cos(\omega'^\top x - \omega^\top x') + \frac{1}{2}\cos(\omega'^\top x + \omega^\top x' + 2b). \end{aligned}$$

Taking expectations over b, the terms containing $2b$ integrate to zero:

$$\mathbb{E}_b[\cos(\omega^\top x + \omega'^\top x' + 2b)] = 0, \quad \mathbb{E}_b[\cos(\omega'^\top x + \omega^\top x' + 2b)] = 0.$$

Thus, the expectations of the cross-terms become:

$$\mathbb{E}_b[I_{12}] = \frac{1}{2}\cos(\omega^\top x - \omega'^\top x'), \quad \mathbb{E}_b[I_{21}] = \frac{1}{2}\cos(\omega'^\top x - \omega^\top x').$$

Combining these with the expectations of I_{11} and I_{22} we get:

$$\mathbb{E}_b[z(x;b)z(x';b)] = \frac{1}{4}\Big[\cos(\omega^\top(x-x')) + \cos(\omega^\top x - \omega'^\top x') + \cos(\omega'^\top x - \omega^\top x') + \cos(\omega'^\top(x-x'))\Big] = \mathcal{T}_{\omega,\omega'}(x,x'),$$

which recovers the full harmonizable integrand $\mathcal{T}_{\omega,\omega'}(x,x')$.

A.3 Exact Amplitude Profiling

To analyze the regularizer in Section 5 without notational clutter, we consider an arbitrary fixed layer and drop the layer index l. Let $\tau > 0$ be a fixed trace budget. For any non-zero coefficient set $\beta = \{\beta_m\}_{m=1}^M \neq 0$, and for any feasible amplitude vector s, the Cauchy–Schwarz inequality (Boyd & Vandenberghe, 2004) gives:

$$\left(\sum_{m=1}^M \|\beta_m\|_2\right)^2 = \left(\sum_{m=1}^M \frac{\|\beta_m\|_2}{\sqrt{s_m}} \cdot \sqrt{s_m}\right)^2 \leq \left(\sum_{m=1}^M \frac{\|\beta_m\|_2^2}{s_m}\right)\left(\sum_{m=1}^M s_m\right).$$

Since $\sum_{m=1}^M s_m = \tau$, dividing by τ provides the lower bound:

$$\sum_{m=1}^M \frac{\|\beta_m\|_2^2}{s_m} \geq \frac{1}{\tau}\left(\sum_{m=1}^M \|\beta_m\|_2\right)^2.$$

Choosing $s_m^* = \tau \frac{\|\beta_m\|_2}{\sum_{j=1}^M \|\beta_j\|_2}$ satisfies $\sum_{m=1}^M s_m^* = \tau$ and achieves this lower bound with equality. If $\beta = 0$, every feasible amplitude vector s achieves the infimum, and $\Omega(0) = 0$.

A.4 Exact Proximal Operator

We consider an arbitrary fixed layer and drop the layer index l. The proximal map for $\Omega(\beta) = \frac{1}{\tau}\big(\sum_{m=1}^M \|\beta_m\|_2\big)^2$ with step size $\gamma > 0$, $\alpha = \frac{2\gamma}{\tau}$, and input $V \in \mathbb{R}^{M\times r}$ solves:

$$\min_{\beta\in\mathbb{R}^{M\times r}} \frac{1}{2}\sum_{m=1}^M \|\beta_m - V_m\|_2^2 + \frac{\alpha}{2}\left(\sum_{m=1}^M \|\beta_m\|_2\right)^2. \tag{20}$$

Expanding $\|\beta_m - V_m\|_2^2 = \|\beta_m\|_2^2 - 2\langle\beta_m, V_m\rangle + \|V_m\|_2^2$, Cauchy–Schwarz requires optimal blocks to align with V_m, i.e., $\beta_m^* = \frac{u_m^*}{v_m} V_m$ when $v_m := \|V_m\|_2 > 0$ (and $\beta_m^* = 0$ if $v_m = 0$), reducing equation 20 to minimizing $\psi(u) = \frac{1}{2}\sum_{m=1}^M (u_m - v_m)^2 + \frac{\alpha}{2}(\sum_{m=1}^M u_m)^2$ over $u \geq 0$.

Associating multipliers $\mu_m \geq 0$ with the non-negativity constraints $-u_m \leq 0$, the Lagrangian is $\mathcal{L}(u,\mu) = \psi(u) - \sum_{m=1}^M \mu_m u_m$. The Karush–Kuhn–Tucker (KKT) stationarity condition requires $\nabla_{u_m}\mathcal{L} = \nabla_{u_m}\psi - \mu_m = u_m - v_m + \alpha t - \mu_m = 0$, where $t = \sum_{j=1}^M u_j \geq 0$ is the total active norm, alongside complementary slackness $\mu_m u_m = 0$. For $u_m > 0$, complementary slackness yields $\mu_m = 0 \implies u_m = v_m - \alpha t$; for $v_m \leq \alpha t$, $u_m = 0$ with $\mu_m = \alpha t - v_m \geq 0$. Thus, coordinate norms satisfy:

$$u_m^*(t) = [v_m - \alpha t]_+ := \max(0, v_m - \alpha t). \tag{21}$$

Summing equation 21 provides the 1D fixed-point equation $t = \phi(t) := \sum_{m=1}^M [v_m - \alpha t]_+$. If $V = 0$, then $t^* = 0$ and $\beta^* = 0$. For $V \neq 0$, $\phi(t)$ is continuous and non-increasing with $\phi(0) = \sum_m v_m > 0$, guaranteeing a unique root $t^* > 0$.

Algorithm 1 Exact Proximal Operator $\mathrm{prox}_{\gamma\Omega}(V)$

Require: $V \in \mathbb{R}^{M\times r}$, $\gamma > 0$, $\tau > 0$. **Ensure:** $\beta^* = \mathrm{prox}_{\gamma\Omega}(V) \in \mathbb{R}^{M\times r}$.

$\alpha \leftarrow 2\gamma/\tau$, $v_m \leftarrow \|V_m\|_2$ for $m \in \{1,\dots,M\}$.
if $\max_m v_m = 0$ **then**
  return V
end if
Sort $v_{(1)} \geq \cdots \geq v_{(M)}$, set $v_{(M+1)} \leftarrow 0$, and compute prefix sums $S_k \leftarrow \sum_{j=1}^{k} v_{(j)}$.
Find smallest $k^* \in \{1,\dots,M\}$ such that $\alpha\frac{S_{k^*}}{1+\alpha k^*} \geq v_{(k^*+1)}$ (or set $k^* = M$).
$t^* \leftarrow \frac{S_{k^*}}{1+\alpha k^*}$.
for $m = 1$ **to** M **do**
  if $v_m > \alpha t^*$ **then**
    $\beta_m^* \leftarrow \left(1 - \frac{\alpha t^*}{v_m}\right) V_m$
  else
    $\beta_m^* \leftarrow 0$
  end if
end for
return β^*

Table 9: Summary of experimental benchmark tasks and architectural configurations.

Dataset	Type	D_{in}	D_{out}	N_{total}	Split ($N_{\text{tr}}/N_{\text{val}}/N_{\text{te}}$)	Loss Function
Emotions	Multi-label	72	6	593	$415/89/89$ (70/15/15%)	Binary Cross-Entropy (Logits)
Scene	Multi-label	294	6	2,407	$1,684/361/362$ (70/15/15%)	Binary Cross-Entropy (Logits)
Yeast	Multi-label	103	14	2,417	$1,691/362/364$ (70/15/15%)	Binary Cross-Entropy (Logits)
SARCOS	Regression	21	7	48,933	$37,811/6,673/4,449$	Mean Squared Error
Parkinson's	Regression	19	2	5,875	5-fold GroupKFold (42 subjects)	Mean Squared Error

Support Identification in $\mathcal{O}(M \log M)$. Sort norms in descending order $v_{(1)} \geq \cdots \geq v_{(M)} \geq v_{(M+1)} := 0$. On candidate support of size k satisfying $v_{(k)} > \alpha t \geq v_{(k+1)}$, the fixed-point equation simplifies to $t = \sum_{j=1}^{k}(v_{(j)} - \alpha t) = S_k - \alpha k t$, providing the root candidate:

$$t_k = \frac{S_k}{1+\alpha k}, \quad \text{where } S_k = \sum_{j=1}^{k} v_{(j)}. \tag{22}$$

The exact active support k^* is the smallest index $k \in \{1,\dots,M\}$ satisfying $\alpha t_k \geq v_{(k+1)}$ (or $k^* = M$). The lower bound $v_{(k^*)} > \alpha t_{k^*}$ is guaranteed:

- for $k^* = 1$, $v_{(1)} > 0$ gives $\alpha t_1 = \frac{\alpha}{1+\alpha} v_{(1)} < v_{(1)}$,
- for $k^* > 1$, failure at $k^* - 1$ implies $\alpha S_{k^*-1} < (1 + \alpha(k^* - 1))v_{(k^*)}$, and adding $\alpha v_{(k^*)}$ to both sides leads to $\alpha t_{k^*} < v_{(k^*)}$.

Algorithm 1 executes this closed-form update in $\mathcal{O}(M \log M + Mr)$ operations and $\mathcal{O}(M)$ auxiliary memory.

B Experimental Setups and Extended Experimental Results

This appendix provides extended evaluations, numerical validation protocols, complete architectural configurations, and benchmark analyses supporting Section 6.

B.1 Architectural Configurations and Model Selection

We use the following architectures in the experiments:

1. **Multilayer Perceptron (MLP):** A 3-layer feedforward network ($D_{\text{in}} \to h \to h \to D_{\text{out}}$) with ReLU activations. The hidden dimension is selected from $h \in \{128, 256\}$.

2. **Stationary RFF Network:** Single-frequency random Fourier features $z_m(x) = \frac{1}{\sqrt{M}} \cos(\omega_m^\top x + b_m)$ with linear mixing $W \in \mathbb{R}^{M \times D_{\text{out}}}$. Dictionary width is selected from $M \in \{256, 512\}$.
3. **Original DSKN (Independent Phases):** Paired-frequency feature map $z_m(x) = \frac{1}{\sqrt{2M}}[\cos(\omega_m^\top x + b_m) + \cos({\omega'_m}^\top x + b'_m)]$ with independent phase variables $b_m, b'_m \sim \mathcal{U}[0, 2\pi]$, linear mixing, and $M \in \{256, 512\}$.
4. **Corrected DSKN (Free Mixing):** Paired-frequency feature map with shared phases b_m as in equation 7, combined with linear output mixing $W \in \mathbb{R}^{M \times D_{\text{out}}}$, and $M \in \{256, 512\}$.

Hyperparameter Tuning and Model Selection. For classification, models are trained for up to 300 epochs with batch size 128, learning rate $\eta = 5 \times 10^{-3}$, weight decay 10^{-4}, and early stopping patience of 25 epochs based on validation BCE. Decision thresholds $\theta_c \in [0.05, 0.95]$ (grid of 19 values) are tuned per dimension on validation probabilities to maximize label-wise F1 score. For regression, models are trained for up to 350 epochs with batch size 256, learning rate $\eta = 10^{-2}$, weight decay 10^{-4}, and patience of 30 epochs based on validation MSE. For DOSKN, hyperparameter selection evaluates group count $Q \in \{4, 8\}$ and subspace rank $r \in \{4, \min(8, D_{\text{out}}), D_{\text{out}}\}$ on validation loss, including a warm-start where (U, β) are initialized from the singular value decomposition of the unconstrained ridge regression weights. All networks are optimized using AdamW. Subspace parameters U are updated with a base learning rate scaled by $\eta_U = 0.3 \cdot \eta_{\text{base}}$ to prevent unstable rotational updates. Following every gradient step, column orthonormality is restored via QR factorization: $U_q \leftarrow Q_U \operatorname{diag}(\operatorname{sgn}(\operatorname{diag}(R_U)))$, ensuring $U_q^\top U_q = I_r$ with positive diagonal signs on R.

B.2 Numerical Verification of Algorithm 1

To verify the correctness of Algorithm 1, we evaluate the solver over 40 synthetic instances with random dimensions $M \in [2, 40]$, $r \in [1, 6]$, and varying parameters $\gamma \in [10^{-4}, 2.0]$, $\tau \in [0.5, 500.0]$. For each trial, we perform two independent checks:

1. **1D Root Bisection Check:** The soft-thresholding scalar fixed-point equation $t = \phi(t) := \sum_{m=1}^{M} [v_m - \alpha t]_+$ involves a continuous, non-increasing function $\phi(t) - t$ with a unique positive root. We locate this root t_{bisect} using 200 iterations of bisection root-finding without using prefix sums. Across all 40 random trials, the maximum relative discrepancy between the algorithmic closed-form root t^* from Line 7 of Algorithm 1 and t_{bisect} was $\max \frac{|t^* - t_{\text{bisect}}|}{\max(10^{-12}, t_{\text{bisect}})} = 3.141 \times 10^{-12}$, confirming exact agreement down to double-precision limits.
2. **Local Optimality Perturbation Test:** We generate 2,000 independent Gaussian random perturbations $\tilde{\beta} = \beta^* + \xi$ around the returned point $\beta^* = \operatorname{prox}_{\gamma\Omega}(V)$ with perturbation scales $\|\xi\| \in [10^{-4}, 0.5]$. For all trials, not a single perturbation achieved a lower objective value. empirically confirming strict local and global minimality.

B.3 Extended Optimization and Orthonormal Retraction Analysis

In non-convex dictionary learning, each subspace basis $U_{l,q} \in \mathbb{R}^{D_l \times r_l}$ resides on the orthonormal manifold $\operatorname{St}(r_l, D_l) = \{U : U^\top U = I_{r_l}\}$. Optimization maintains orthonormality by performing QR retraction with positive diagonal signs on R following gradient steps.

We evaluate the necessity of this retraction on SARCOS dataset ($N_{\text{sub}} = 5000$). We compare unregularized networks against soft penalty ($\lambda\Omega$ via AdamW) and proximal shrinkage (Algorithm 1) for regularization grids $\gamma, \lambda \in [10^{-5}, 10^{-3}]$ with and without QR retraction. For each run, we report test nRMSE, validation loss, effective output weight norm $\sum_{m=1}^{M} \|c_m\|_2$, and the orthonormal deviation error $\Delta_{\text{St}} = \frac{1}{Q} \sum_{q=1}^{Q} \|U_q^\top U_q - I_r\|_F$.

From Table 10 we can observe the following: (1) Without retraction, column orthogonality collapses immediately (Δ_{St} increases from 10^{-7} to 0.55–1.58), violating the isometric embedding constraint $\|U\beta\|_2 = \|\beta\|_2$ and destabilizing operator-valued RKHS geometry. QR retraction enforces $\Delta_{\text{St}} \leq 3.5 \times 10^{-7}$ throughout training with minimal computational overhead. (2) When matched on effective output-weight norm ($\sum \|c_m\|_2 \approx 625$), the exact proximal step reaches an

Table 10: Orthonormal retraction and regularization comparison on SARCOS ($M = 256, r = 3, Q = 3$, mean across 3 seeds).

Configuration	QR Retract	Strength (γ, λ)	Test nRMSE ↓	Val Loss ↓	$\sum \|c_m\|_2$	$\Delta_{\mathrm{St}} = \|U^\top U - I\|_F$
Unregularized	True	0.0	0.1525 ± 0.0028	0.02462	863.69	2.24×10^{-7}
	False	0.0	0.1483 ± 0.0021	0.02381	683.17	7.55×10^{-1}
Soft Penalty	True	1×10^{-5}	0.1471 ± 0.0027	0.02334	517.75	2.47×10^{-7}
	True	1×10^{-4}	0.1768 ± 0.0038	0.03310	422.11	2.02×10^{-7}
	False	1×10^{-5}	0.1492 ± 0.0032	0.02401	386.67	1.36
	False	1×10^{-4}	0.1688 ± 0.0009	0.03055	225.67	1.58
Algorithm 1	True	1×10^{-4}	0.1505 ± 0.0026	0.02410	777.83	2.56×10^{-7}
	True	3×10^{-4}	$\mathbf{0.1469} \pm 0.0002$	**0.02288**	625.52	3.25×10^{-7}
	True	1×10^{-3}	0.1492 ± 0.0026	0.02360	374.43	3.19×10^{-7}
	False	3×10^{-4}	0.1472 ± 0.0013	0.02284	620.00	5.45×10^{-1}

Table 11: Classification performance (mean ± std for 5 random splits).

Task / Family	Macro-F1 ↑	Micro-F1 ↑	mAP ↑	Hamming ↓	Coupling P
Emotions ($D_{\mathrm{in}} = 72, D_{\mathrm{out}} = 6, N = 593, M = 512$)					
MLP	$\mathbf{0.6909} \pm 0.0316$	$\mathbf{0.6970} \pm 0.0241$	$\mathbf{0.7538} \pm 0.0452$	$\mathbf{0.2119} \pm 0.0164$	0
Stationary RFF	0.6795 ± 0.0374	0.6888 ± 0.0374	0.7532 ± 0.0517	0.2148 ± 0.0323	3072
Original DSKN (independent)	0.6876 ± 0.0354	0.6900 ± 0.0363	0.7528 ± 0.0414	0.2189 ± 0.0220	3072
Corrected DSKN (free mixing)	$\mathbf{0.6909} \pm 0.0362$	0.6943 ± 0.0359	0.7510 ± 0.0437	0.2174 ± 0.0230	3072
Proposed DOSKN	0.6839 ± 0.0253	0.6878 ± 0.0296	0.7486 ± 0.0439	0.2259 ± 0.0237	2240–3360
Scene ($D_{\mathrm{in}} = 294, D_{\mathrm{out}} = 6, N = 2407, M = 512$)					
MLP	0.7555 ± 0.0191	0.7436 ± 0.0220	0.8388 ± 0.0300	0.0966 ± 0.0106	0
Stationary RFF	0.7747 ± 0.0272	0.7665 ± 0.0277	0.8538 ± 0.0229	0.0852 ± 0.0106	3072
Original DSKN (independent)	0.7782 ± 0.0259	0.7671 ± 0.0273	0.8539 ± 0.0223	0.0862 ± 0.0112	3072
Corrected DSKN (free mixing)	0.7691 ± 0.0166	0.7586 ± 0.0204	0.8553 ± 0.0258	0.0898 ± 0.0068	3072
Proposed DOSKN	$\mathbf{0.7808} \pm 0.0174$	$\mathbf{0.7739} \pm 0.0210$	$\mathbf{0.8569} \pm 0.0198$	$\mathbf{0.0809} \pm 0.0081$	2144–3216
Yeast ($D_{\mathrm{in}} = 103, D_{\mathrm{out}} = 14, N = 2417, M = 512$)					
MLP	$\mathbf{0.4935} \pm 0.0088$	0.6161 ± 0.0093	0.5136 ± 0.0183	0.3094 ± 0.0071	0
Stationary RFF	0.4819 ± 0.0123	0.5874 ± 0.0083	0.4949 ± 0.0209	0.3453 ± 0.0125	3584
Original DSKN (independent)	0.4837 ± 0.0151	$\mathbf{0.6163} \pm 0.0185$	0.5086 ± 0.0193	$\mathbf{0.2998} \pm 0.0225$	7168
Corrected DSKN (free mixing)	0.4853 ± 0.0151	0.6107 ± 0.0177	0.5018 ± 0.0142	0.3092 ± 0.0226	7168
Proposed DOSKN	0.4851 ± 0.0219	0.6051 ± 0.0273	$\mathbf{0.5176} \pm 0.0295$	0.3100 ± 0.0271	2496–4992

optimal test nRMSE of 0.1469 ± 0.0002 with notably lower variance than gradient penalty updates (0.1471 ± 0.0027).

B.4 Classification Performance Evaluation

We also evaluate DOSKN on three classification tasks and compare it to Multilayer Perceptrons (MLP), stationary RFF networks, original DSKN with independent phases, and corrected DSKN with unconstrained free linear mixing. The hyper-parameters of each method are selected by applying a grid search on the validation set. For DOSKN, we tested $Q \in \{4, 8\}$ and $r \in \{4, \ldots, D_{\mathrm{out}}\}$, including the unconstrained case $r = D_{\mathrm{out}}$.

Table 11 summarizes classification performance. On *Emotions* dataset, where the output dimension is small ($D_{\mathrm{out}} \approx r$), the corrected DSKN provides the best Macro-F1 performance and improves over the original DSKN in most metrics. DOSKN provides competitive performance while while saving 27.1% coupling parameters under $Q = 8, r = 4$. On *Scene* dataset, DOSKN provides the best performance accross all metrics. On *Yeast* dataset, DOSKN achieves the best mean average precision while reducing coupling parameters from 7168 to as low as 2496 (a 65.2% reduction under $Q = 8, r = 4$).