
Refining Dual Spectral Sparsity in Transformed Tensor Singular Values

Anonymous Author(s)

Affiliation

Address

email

Abstract

1 The Tensor Nuclear Norm (TNN), derived from the tensor Singular Value Decom-
2 position, is a central low-rank modeling tool that enforces *element-wise sparsity*
3 on frequency-domain singular values and has been widely used in multi-way data
4 recovery for machine learning and computer vision. However, as a direct extension
5 of the matrix nuclear norm, it inherits the assumption of *single-level spectral spar-*
6 *sity*, which strictly limits its ability to capture the *multi-level spectral structures*
7 inherent in real-world data—particularly the coexistence of low-rankness within
8 and sparsity across frequency components. To address this, we propose the tensor
9 ℓ_p -Schatten- q quasi-norm ($p, q \in (0, 1]$), a new metric that enables *dual spectral*
10 *sparsity control* by jointly regularizing both types of structure. While this formula-
11 tion generalizes TNN and unifies existing methods such as the tensor Schatten- p
12 norm and tensor average rank, it differs fundamentally in modeling principle by
13 coupling global frequency sparsity with local spectral low-rankness. This coupling
14 introduces significant theoretical and algorithmic challenges. To tackle these chal-
15 lenges, we provide a theoretical characterization by establishing the first minimax
16 error bounds under dual spectral sparsity, and an algorithmic solution by designing
17 an efficient reweighted optimization scheme tailored to the resulting nonconvex
18 structure. Numerical experiments demonstrate the effectiveness of our method in
19 modeling complex multi-way data.

20 1 Introduction

21 Modeling latent structural patterns in high-dimensional signals is a fundamental challenge across
22 domains such as machine learning and signal processing [28, 63, 30]. Real-world datasets are often
23 inherently multi-modal and high-dimensional (tensor-form), containing intricate dependencies that
24 cannot be adequately captured by naïve modeling or vector/matrix-based representations [8]. A
25 common strategy to uncover these relationships is to impose a *low-rank* prior, which isolates essential
26 information and reduces the degrees of freedom, focusing on the principal components of the signal
27 [32, 2]. Traditional tensor decomposition methods, such as CANDECOMP/PARAFAC (CP) [7],
28 Tucker [47], and Tensor Train [34], have been widely used to model tensor signals [8, 27, 18, 59].
29 While effective in certain scenarios, these methods rely on the assumption of intrinsic low-rankness
30 in the *original domain*, which may fail to hold in complex, real-world applications. This limitation
31 has led to the development of *transformed-domain* modeling, where linear transformations like
32 the Discrete Fourier Transform (DFT) are applied to reveal more pronounced low-rank patterns.
33 Within this paradigm, the tensor Singular Value Decomposition (t-SVD) has emerged as a powerful
34 framework with notable success in applications such as image and video analysis [28, 63, 56, 52].

35 Building on the t-SVD framework, the Tensor Nuclear Norm (TNN) has become an extensively
36 adopted regularizer for low-rank tensor modeling [31, 63, 41, 14, 61, 29, 64]. By extending the

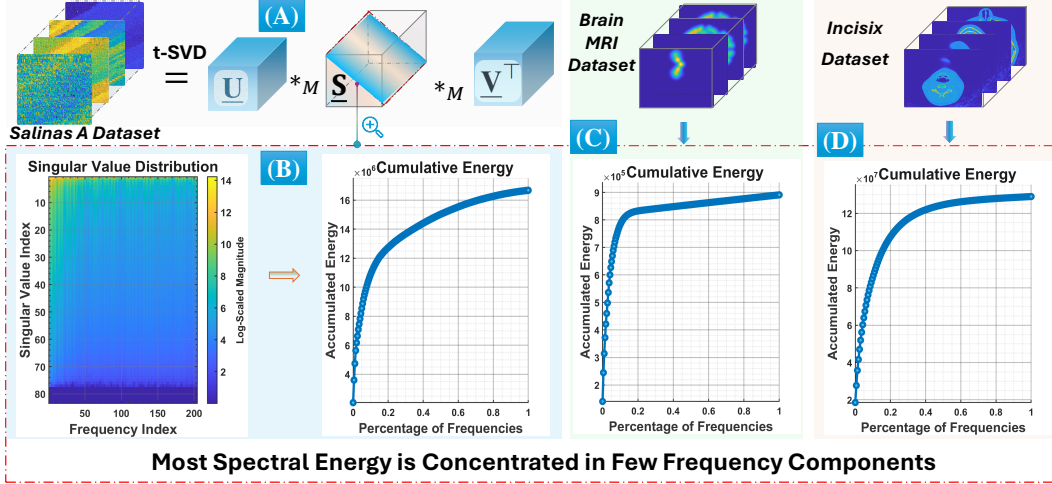


Figure 1: Empirical illustration of dual spectral sparsity patterns in the transformed (DCT) domain via t-SVD. (A) The t-SVD framework decomposes a tensor into frequency-domain singular structures. (B)-Left: Singular value heatmap of the *Salinas A* dataset under t-SVD—each column represents one frequency slice. Vertical decay reveals intra-frequency low-rankness, while horizontal variation indicates sparsity across frequencies. (B)-Right, (C), (D): Cumulative energy curves for *Salinas A*, *Brain MRI*, and *Incisix* datasets show that over 80% of total spectral energy is concentrated in the top 15%–30% frequency components, confirming frequency-wise sparsity. These observations support the presence of a dual-level spectral structure and motivate regularization schemes that go beyond uniform norms like TNN [63, 30] to jointly model frequency sparsity and low-rankness.

matrix nuclear norm to the tensor setting, TNN promotes low-rankness by enforcing *element-wise sparsity* on singular values in the transformed domain [24, 63]. This formulation effectively captures low-rank dependencies within individual frequency components.

However, a long-overlooked limitation of TNN lies in its assumption of *uniform spectral regularization*, which treats all frequency components equally regardless of their relative importance. From a signal processing perspective, this *single-level sparsity* design fails to account for the dual-level structure often observed in transformed tensor data. In particular, real-world tensors may exhibit strong low-rankness within each frequency component along with sparsity across the frequency domain. As illustrated in Fig. 1 and further discussed in §3, empirical analyses of several datasets, including hyperspectral images and medical imaging volumes, indicate that a small subset of frequency slices contributes the majority of spectral energy. In addition, these dominant components often exhibit pronounced low-rank structures within each frequency slice. These observations suggest the need for a more flexible regularization framework that can separately characterize both intra-frequency low-rankness and inter-frequency sparsity, instead of relying on a uniform scheme like TNN.

These limitations necessitate a new method capable of modeling both levels of sparsity. This raises three interconnected questions:

RQ1 (Modeling): *how to effectively model both intra-frequency and inter-frequency dependencies in tensor data?*

RQ2 (Theory): *can we establish rigorous theoretical guarantees to validate such a framework, given the challenges of analyzing coupled sparsity?*

RQ3 (Algorithm): *can efficient algorithms be designed to tackle the optimization challenges introduced by the coupled sparsity structure?*

To address these questions, we propose the *tensor ℓ_p -Schatten- q quasi-norm*, a novel framework introducing *dual spectral sparsity control* to simultaneously model both within-frequency and across-frequency dependencies. Specifically, parameter p governs sparsity among different frequency components (RQ1), while parameter q controls low-rankness within each frequency component. This framework generalizes and extends TNN, unifying existing methods such as the tensor Schatten- p quasi-norm [23] and tensor average rank [53] into a single, versatile framework.

While our framework offers promising modeling capabilities, the *coupled nature of this dual spectral sparsity* introduces significant theoretical and computational challenges. Our main contributions in developing and validating this framework are as follows:

- **Structural Modeling (RQ1):** To the best of our knowledge, this work is the first to rigorously formalize and explicitly model a coupled spectral structure within the t-SVD framework, where inter-frequency sparsity coexists with intra-frequency low-rankness (Section 3). The proposed ℓ_p -Schatten- q quasi-norm jointly models both inter-frequency sparsity and intra-frequency low-rankness, while allowing separate control over each via parameters p and q . This formulation captures hierarchical spectral structure beyond uniform regularizers such as TNN.
- **Theoretical Guarantees (RQ2):** We establish sharp minimax lower and upper bounds for tensor estimation under dual spectral sparsity, covering both hard and soft regimes (Section 4). The analysis introduces new techniques to characterize the complexity of coupled parameter spaces, extending classical tools such as covering numbers and metric entropy to the tensor spectral setting.
- **Optimization and Empirical Validation (RQ3):** We develop a scalable proximal algorithm tailored to the proposed quasi-norm (Section 5). It employs a reweighted $\ell_{1/2}$ approximation and frequency-wise singular value updates in the transform domain, effectively handling the nonconvexity and structural coupling induced by dual spectral sparsity. Experiments on real-world tensor recovery tasks demonstrate the potential applicability of our method (Section 6).

The remainder of the paper is organized as follows. Section 2 reviews basic preliminaries. Section 3 introduces the proposed quasi-norm. Sections 4 and 5 present the theoretical analysis and optimization algorithm, respectively. Experimental results are reported in Section 6, followed by the conclusion in Section 7. Details on related work, proofs, algorithms, and experiments are provided in the appendix.

2 Notations and Preliminaries

Notations. For any positive integer d , let $[d] := \{1, \dots, d\}$. We denote vectors by lowercase bold letters (e.g., \mathbf{a}), matrices by uppercase bold letters (e.g., \mathbf{A}), and 3-way tensors by underlined uppercase letters (e.g., $\underline{\mathbf{A}}$). Constants, represented as c and its variants (e.g., c_1 , C), may vary in value across contexts. For a 3-way tensor of size $d_1 \times d_2 \times m$, we assume $d_1 \geq d_2$ without loss of generality.

For a matrix $\mathbf{A} \in \mathbb{R}^{d_1 \times d_2}$, we define $\sigma(\mathbf{A})$ as the vector of its singular values, arranged in descending order. The spectral norm $\|\mathbf{A}\|_{\text{spec}}$ and nuclear norm $\|\mathbf{A}\|_*$ of \mathbf{A} are defined as the largest and the sum of its singular values, respectively. For any tensor $\underline{\mathbf{A}}$, we define its ℓ_p -norm as $\|\underline{\mathbf{A}}\|_p := \|\text{vec}(\underline{\mathbf{A}})\|_p$ and its Frobenius norm as $\|\underline{\mathbf{A}}\|_F := \|\text{vec}(\underline{\mathbf{A}})\|_2$, where $\text{vec}(\cdot)$ denotes the vectorization operation [22]. The inner product of two tensors $\underline{\mathbf{A}}$ and $\underline{\mathbf{B}}$ is given by $\langle \underline{\mathbf{A}}, \underline{\mathbf{B}} \rangle := \text{vec}(\underline{\mathbf{A}})^\top \text{vec}(\underline{\mathbf{B}})$. For a tensor $\underline{\mathbf{A}} \in \mathbb{R}^{d_1 \times d_2 \times m}$, we denote its i -th frontal slice as $\underline{\mathbf{A}}_{:, :, i}$ or simply $\underline{\mathbf{A}}_i$ when clear from context.

The t-SVD Framework. The t-SVD framework is based on the t-product operation, a generalization of matrix multiplication to tensors, which operates under an invertible linear transform M [19]. By enhancing low-rank properties through specific linear transformations, this approach effectively exploits intrinsic correlations within the data [61, 51]. This paper adopts the convention of using orthogonal matrices for M due to their stability and computational advantages [29, 50]. Specifically, for an orthogonal matrix $\mathbf{M} \in \mathbb{R}^{m \times m}$, we define the M -linear transform and its inverse on a tensor $\underline{\mathbf{T}} \in \mathbb{R}^{d_1 \times d_2 \times m}$ as:

$$M(\underline{\mathbf{T}}) := \underline{\mathbf{T}} \times_3 \mathbf{M}, \quad \text{and} \quad M^{-1}(\underline{\mathbf{T}}) := \underline{\mathbf{T}} \times_3 \mathbf{M}^{-1}, \quad (1)$$

where \times_3 denotes the mode-3 tensor-matrix product [19]. Using this transform, we introduce the basic notions in the t-SVD framework.

Definition 2.1 (t-product [19]). The t-product of two tensors $\underline{\mathbf{A}} \in \mathbb{R}^{d_1 \times d_2 \times m}$ and $\underline{\mathbf{B}} \in \mathbb{R}^{d_2 \times d_3 \times m}$ under the transform M in (1) is denoted by $\underline{\mathbf{A}} *_M \underline{\mathbf{B}} = \underline{\mathbf{C}} \in \mathbb{R}^{d_1 \times d_3 \times m}$, where $M(\underline{\mathbf{C}}) = M(\underline{\mathbf{A}}) \odot M(\underline{\mathbf{B}})$ in the transformed domain, and \odot denotes the frontal-slice-wise product of the tensors.

Definition 2.2 (M -block-diagonal matrix [50]). For a tensor $\underline{\mathbf{T}} \in \mathbb{R}^{d_1 \times d_2 \times m}$, its M -block-diagonal matrix $\bar{\mathbf{T}} \in \mathbb{R}^{d_1 m \times d_2 m}$ is defined as

$$\bar{\mathbf{T}} := \text{bdiag}(M(\underline{\mathbf{T}})) = \text{diag}(M(\underline{\mathbf{T}})_{:, :, 1}, \dots, M(\underline{\mathbf{T}})_{:, :, m}),$$

where $M(\underline{\mathbf{T}})$ is the mode-3 transform of $\underline{\mathbf{T}}$, and the operator $\text{bdiag}(\cdot)$ stacks the frontal slices as diagonal blocks.

115 We now formally introduce the t-SVD, as illustrated in Fig. 1-(A).

116 **Definition 2.3** (t-SVD and tensor tubal rank [19]). The tensor Singular Value Decomposition (t-SVD)
117 of a tensor $\underline{\mathbf{T}} \in \mathbb{R}^{d_1 \times d_2 \times m}$ under the invertible linear transform M in (1) is:

$$\underline{\mathbf{T}} = \underline{\mathbf{U}} *_{\mathbf{M}} \underline{\mathbf{S}} *_{\mathbf{M}} \underline{\mathbf{V}}^{\top}, \quad (2)$$

118 where $\underline{\mathbf{U}} \in \mathbb{R}^{d_1 \times d_1 \times m}$ and $\underline{\mathbf{V}} \in \mathbb{R}^{d_2 \times d_2 \times m}$ are t-orthogonal tensors, and $\underline{\mathbf{S}} \in \mathbb{R}^{d_1 \times d_2 \times m}$ is an
119 f-diagonal tensor. The tubal rank of $\underline{\mathbf{T}}$ is defined as the number of non-zero tubes in $\underline{\mathbf{S}}$ in the t-SVD,
120 i.e., $r_{\text{tb}}(\underline{\mathbf{T}}) := \#\{i \mid \underline{\mathbf{S}}_{:,i,:} \neq \mathbf{0}, i \leq \min\{d_1, d_2\}\}$.

121 To further model the low-rank structure of tensors in the transformed domain, the tensor nuclear norm
122 (TNN) is proposed as a key regularizer in low-rank tensor learning:

123 **Definition 2.4** (Tensor nuclear norm [31]). The tensor nuclear norm (TNN) of a tensor $\underline{\mathbf{T}} \in$
124 $\mathbb{R}^{d_1 \times d_2 \times m}$ under the transform M are defined as $\|\underline{\mathbf{T}}\|_* := \|\bar{\mathbf{T}}\|_* = \|\sigma(\bar{\mathbf{T}})\|_1$.

125 In this definition, TNN captures *the element-wise sparsity of the transformed spectrum* $\sigma(\bar{\mathbf{T}}) \in$
126 $\mathbb{R}^{m \times \min\{d_1, d_2\}}$, allowing it to promote low-rank characteristics in the spectral domain. This property
127 has made TNN a foundational tool in tensor analysis, particularly for low-rank tensor recovery in
128 various applications such as image inpainting [30].

129 3 Dual Spectral Sparsity in the t-SVD Framework

130 Effectively capturing both intra-frequency low-rankness and inter-frequency sparsity (**RQ1**) is es-
131 sential for modeling structured tensor data. While methods like TNN emphasize within-frequency
132 low-rankness, they overlook sparsity across frequencies, limiting their ability to represent hierar-
133 chical dependencies. To overcome this, we introduce the ℓ_p -Schatten- q quasi-norm, a dual-sparsity
134 regularization framework designed to capture both levels of structure in a unified way.

135 **Limitations of TNN from a Group Sparsity Perspective.** According to Definition 2.4, the tensor
136 nuclear norm (TNN) promotes low-rankness by enforcing element-wise sparsity on singular values in
137 the transformed domain, effectively capturing intra-frequency low-rank structures. However, it applies
138 uniform regularization across all frequency components, regardless of their spectral importance. This
139 design fails to exploit the potential sparsity across frequency slices that is often present in real-world
140 tensors. Fig. 1 presents empirical evidence from three representative datasets—*Salinas A*¹, *Brain MRI*
141 [57], and *Incisix* [10]—demonstrating that only a small portion of frequency components accounts
142 for the majority of spectral energy. Specifically, more than 80% of the energy is concentrated in the
143 top 15%–30% of frequency bands. Meanwhile, the singular value heatmap (Fig. 1(B)-Left) reveals
144 pronounced horizontal sparsity, indicating that many frequency slices contribute minimally. Within
145 each active frequency slice, singular values decay rapidly, confirming low-rankness.

146 These observations suggest a dual-level structure comprising inter-frequency sparsity and intra-
147 frequency low-rankness. From a group sparsity perspective, the spectrum $\sigma(\bar{\mathbf{T}})$ can be partitioned
148 into groups, where each group corresponds to the singular values $\sigma(M(\underline{\mathbf{T}})_{:,i,:})$ of a specific frequency
149 slice. TNN enforces uniform regularization across these groups, overlooking their heterogeneous
150 importance. As a result, it may underperform when modeling data with hierarchical spectral structures.
151 These limitations motivate a more expressive framework that separately accounts for both levels of
152 structure.

153 **Hard Dual Spectral Sparsity.** To address the limitations of TNN, we first define a hard dual spectral
154 sparsity structure, where the tensor is assumed to satisfy exact sparsity constraints across and within
155 frequency components. This serves as an idealized formulation that captures the extreme case of dual
156 spectral sparsity and provides a clean theoretical foundation for later analysis.

157 **Definition 3.1** (Hard Dual Spectral Sparsity). A tensor $\underline{\mathbf{T}} \in \mathbb{R}^{d_1 \times d_2 \times m}$ is said to exhibit (s, r) -dual
158 sparsity under a linear transform M if it satisfies two constraints:

159 **I. Inter-frequency sparsity:** The number of active frequency components is limited to at most s .
160 Specifically, only s out of the m frequency components can have non-zero singular value vectors:
161 $\sum_{i=1}^m \mathbb{I}(\sigma(M(\underline{\mathbf{T}})_{:,i,:}) \neq \mathbf{0}) \leq s$, where $\sigma(M(\underline{\mathbf{T}})_{:,i,:})$ denotes the singular value vector of the i -th
162 frontal slice in the transformed domain.

¹https://www.ehu.eus/ccwintco/index.php?title=Hyperspectral_Remote_Sensing_Scenes

163 **II. Intra-frequency low-rankness:** Within each active frequency component, the number of non-zero
 164 singular values is constrained to at most r . This condition ensures a low-rank structure for
 165 each frequency slice ($\forall i \in [m]$): $\sum_{j=1}^{\min\{d_1, d_2\}} \mathbb{I}(\sigma_j(M(\mathbf{T})_{::,i}) \neq 0) \leq r$, where $\sigma_j(M(\mathbf{T})_{::,i})$
 166 denotes the j -th singular value of the i -th frontal slice of $M(\mathbf{T})$.

167 This definition captures a strict form of dual-level structure by simultaneously enforcing sparsity
 168 across frequencies and low-rankness within each active frequency slice. While such hard constraints
 169 may be too restrictive in practical scenarios, especially where spectral contributions decay grad-
 170 ually, they provide a clear conceptual framework to motivate and analyze the more flexible soft
 171 regularization.

172 **Soft Dual Spectral Sparsity.** While the hard dual spectral sparsity model provides a clean conceptual
 173 foundation, its strict assumption of exact sparsity and fixed-rank constraints is often impractical in
 174 real-world scenarios. In many cases, singular values decay gradually rather than drop abruptly to
 175 zero, and the true number of active frequency components may be ambiguous or noise-sensitive. To
 176 overcome these limitations, we introduce a soft relaxation that allows for approximate sparsity and
 177 low-rankness in a continuous manner. Specifically, we propose the ℓ_p -Schatten- q quasi-norm, which
 178 relaxes the hard dual-sparsity constraints into a soft dual spectral sparsity framework.

179 **Definition 3.2** (Tensor ℓ_p -Schatten- q quasi-norm). For a tensor $\mathbf{T} \in \mathbb{R}^{d_1 \times d_2 \times m}$, we define its tensor
 180 ℓ_p -Schatten- q quasi-norm (abbreviated as $\ell_p(S_q)$ -norm) as:

$$\|\mathbf{T}\|_{\ell_p(S_q)} := \left(\sum_{i=1}^m \left(\sum_{j=1}^{d_1 \wedge d_2} \sigma_j(M(\mathbf{T})_{::,i})^q \right)^{\frac{p}{q}} \right)^{\frac{1}{p}}, \quad (3)$$

181 where the exponents $(p, q) \in (0, 1]^2$.

182 In this quasi-norm, p governs the inter-frequency sparsity by promoting a group-wise regularization
 183 across frequency components, effectively highlighting significant groups while suppressing others.
 184 Simultaneously, q controls the intra-frequency low-rankness by encouraging sparsity in the singular
 185 values within each frequency slice, thereby modeling the intrinsic low-rank structure of the data.
 186 This soft dual spectral sparsity framework provides a unified yet versatile approach to address the
 187 hierarchical complexity of tensor data.

188 The ℓ_p -Schatten- q quasi-norm encompasses several existing regularization methods: it recovers TNN
 189 when $(p, q) = (1, 1)$ [31], approximates the average rank as $(p, q) \rightarrow (1, 0)$ [53], and reduces to the
 190 tensor Schatten- q norm when $p = q$ [23], thereby offering greater modeling flexibility. *Despite*
 191 *generalizing these regularizers, it fundamentally differs by jointly enforcing global frequency sparsity*
 192 *and local spectral low-rankness.*

193 While TNN applies uniform regularization across all singular values, the ℓ_p -Schatten- q quasi-norm
 194 introduces dual spectral sparsity control, modeling both inter-frequency sparsity through the ℓ_p -quasi-
 195 norm and intra-frequency low-rankness via the Schatten- q quasi-norm. This dual-level flexibility
 196 makes the proposed framework particularly well-suited for hierarchical and multi-scale data, where
 197 dependencies and sparsity exhibit layered structures. By bridging the gap between element-wise
 198 sparsity (as in TNN) and structured group sparsity, the ℓ_p -Schatten- q quasi-norm offers a more
 199 expressive and adaptable approach, enabling precise control over structural patterns in modern
 200 tensor-based analysis and recovery tasks.

201 4 Theory of Dual Spectral Sparse Tensor Estimation

202 This section develops the theoretical foundations of tensor estimation with dual spectral sparsity
 203 structures (RQ2).

204 **Challenges.** Dual spectral sparsity, combining inter-frequency sparsity with intra-frequency low-
 205 rankness, leads to a *globally coupled structure* that fundamentally differs from classical decoupled
 206 models like TNN. The ℓ_p -Schatten- q quasi-norm imposes interdependent constraints across frequency
 207 slices, resulting in a highly non-convex parameter space with nested sparsity patterns. This coupling
 208 prohibits slice-wise decomposition and complicates the use of standard tools. Accurately characteriz-
 209 ing the estimation complexity demands novel extensions of covering numbers and metric entropy
 210 that jointly capture discrete sparsity and continuous low-rank structure.

To understand the statistical limits of learning under dual spectral sparsity, we analyze a simplified but representative model: the Gaussian location model, where the observed tensor is corrupted by additive noise. This setting preserves the core structural properties—inter-frequency sparsity and intra-frequency low-rankness—while avoiding complications unrelated to sparsity itself. Within this framework, we define structured parameter spaces that capture hard and soft variants of dual spectral sparsity, and establish sharp minimax lower and upper bounds under each. These results reveal how the joint effects of frequency selection and within-slice spectral decay determine the fundamental estimation limits, and provide theoretical justification for our proposed regularization.

4.1 Gaussian Location Model

Consider the Gaussian location model (GLM) [25], where n independent noisy realizations of the target tensor $\mathbf{L}^* \in \mathbb{R}^{d_1 \times d_2 \times m}$ are observed as:

$$\mathbf{Y}_i = \mathbf{L}^* + \mathbf{E}_i, \quad i \in [n], \quad (4)$$

where $\mathbf{Y}_i \in \mathbb{R}^{d_1 \times d_2 \times m}$ is the observed tensor, \mathbf{L}^* represents the ground truth tensor of interest, and $\mathbf{E}_i \in \mathbb{R}^{d_1 \times d_2 \times m}$ denotes the noise tensor with entries independently drawn from $\mathcal{N}(0, \sigma^2)$. The parameter σ characterizes the noise level. To simplify the analysis, we consider the sample mean of observations $\bar{\mathbf{Y}} = n^{-1} \sum_{i=1}^n \mathbf{Y}_i = \mathbf{L}^* + \bar{\mathbf{E}}$, where $\bar{\mathbf{E}} = n^{-1} \sum_{i=1}^n \mathbf{E}_i$ is the aggregated noise tensor with entries independently distributed as $\mathcal{N}(0, \sigma^2/n)$. The goal is to estimate the ground truth tensor \mathbf{L}^* based on the noisy observations $\{\mathbf{Y}_i\}_{i=1}^n$. In particular, we aim to recover \mathbf{L}^* under dual spectral sparsity assumptions.

Remark 4.1. We adopt the Gaussian location model to isolate the core effects of dual spectral sparsity and the ℓ_p -Schatten- q regularization, avoiding additional complications from design tensors or sampling operators in tensor regression [60, 51, 35]. This simplified setting enables cleaner analysis and yields insights that extend naturally to regression problems under standard conditions such as RIP [60] or RSC [51, 35, 33].

Dual Spectral Sparsity Assumptions. We consider three distinct sparsity models for \mathbf{L}^* :

A1. Hard dual spectral sparsity: Let \mathbf{L}^* belong to the parameter space

$$\mathbf{T}_{0,0}(s, r) = \{\mathbf{L} : \text{at most } s \text{ active frequency slices, each of rank at most } r\}. \quad (5)$$

This model enforces exact inter-frequency sparsity and intra-frequency low-rankness.

A2. Hard frequency sparsity and soft rank constraint (hard-soft sparsity): Let \mathbf{L}^* lie in

$$\mathbf{T}_{0,q}(s, R) = \left\{ \mathbf{L} : |\{i : M(\mathbf{L})_{:, :, i} \neq \mathbf{0}\}| \leq s, \|M(\mathbf{L})_{:, :, i}\|_{S_q}^q \leq R, \forall i \in [m] \right\}. \quad (6)$$

This space imposes hard inter-frequency sparsity and soft Schatten- q constraints within each active slice.

A3. Soft dual spectral sparsity: Let \mathbf{L}^* belong to the parameter space

$$\mathbf{T}_{p,q}(R) = \left\{ \mathbf{L} : \|\mathbf{L}\|_{\ell_p(S_q)}^p \leq R \right\}. \quad (7)$$

Here, p promotes inter-frequency sparsity and q controls intra-frequency low-rankness via spectral decay; R specifies the quasi-norm ball radius.

These parameter spaces offer different views on structured tensor estimation: the *hard sparsity* model enforces strict thresholds, the *hard-soft model* balances structure with adaptability, and the *fully soft model* captures gradual spectral decay. Our goal is to estimate \mathbf{L}^* and derive minimax bounds under these assumptions.

4.2 Minimax Risk over Dual-level Sparse Structures

A key theoretical question in high-dimensional tensor estimation is: *What are the fundamental limits for recovering a tensor with dual spectral sparsity from noisy observations?* To address this, we establish minimax lower and upper bounds that characterize the best possible estimation accuracy achievable by any estimator under dual spectral sparsity assumptions.

$$\mathfrak{M}(\mathbf{T}) = \inf_{\hat{\mathbf{L}}} \sup_{\mathbf{L}^* \in \mathbf{T}} \mathbb{E} \left[\|\hat{\mathbf{L}} - \mathbf{L}^*\|_{\mathbb{F}}^2 \right], \quad (8)$$

where \mathbf{T} is the parameter space. Following [29, 30], we consider $d_1 = d_2 = d$ for simplicity.

Theorem 4.2 (Minimax Bounds). *The minimax risk under dual spectral sparsity satisfies the following bounds under certain conditions²:*

I. *Hard constraints on both frequency sparsity and per-slice low-rankness:*

$$\mathfrak{M}(\mathbf{T}_{0,0}(s, r)) \asymp \frac{\sigma^2}{n} \left(s \log \frac{em}{s} + srd \right).$$

II. *Hard frequency sparsity with soft intra-slice Schatten- q constraints:*

$$\mathfrak{M}(\mathbf{T}_{0,q}(s, R)) \asymp \frac{\sigma^2}{n} s \log \frac{em}{s} + sR \left(\frac{\sigma^2}{n} d \right)^{1-\frac{q}{2}}.$$

III. *Soft $\ell_p(S_q)$ constraints over both frequency and rank dimensions:*

$$\mathfrak{M}(\mathbf{T}_{p,q}(R)) \asymp \begin{cases} R \left(\frac{\sigma^2 n}{d} \right)^{\frac{p-2}{2}} + R \left(\frac{\sigma^2 n}{\log m} \right)^{\frac{p-2}{2}}, & p > q, \\ R^{\frac{q}{p}} \left(\frac{\sigma^2 n}{d} \right)^{\frac{q-2}{2}} + R \left(\frac{\sigma^2 n}{\log m} \right)^{\frac{p-2}{2}}, & p \leq q, m > d^2, \\ R^{\frac{q}{p}} \left(\frac{\sigma^2 n}{d} \right)^{\frac{q-2}{2}}, & p \leq q, m \leq d^2. \end{cases}$$

Theorem 4.2 establishes the fundamental limits of estimation accuracy under different dual spectral sparsity structures. The minimax risk quantifies the worst-case squared Frobenius norm error that any estimator must incur when recovering a structured tensor from noisy observations. The results reveal the intricate balance between inter-frequency sparsity and intra-frequency low-rankness, showing how these factors jointly govern estimation complexity:

I. In the *hard sparsity* case, the estimation error consists of two terms: (i) $s \log(em/s)$, which reflects the difficulty of selecting s active frequency components, and (ii) srd , which characterizes the challenge of estimating rank- r matrices within each component.

II. In the *hard-soft sparsity* setting, the second term adapts to $sR(n^{-1}d)^{1-q/2}$, incorporating a smoother spectral decay controlled by q . Smaller q values impose stronger low-rank constraints, effectively reducing estimation complexity by promoting more aggressive rank sparsity.

III. In the *fully soft sparsity* scenario, where both inter-frequency sparsity and intra-frequency rank constraints are relaxed, the minimax risk follows distinct scaling behaviors across regimes. When $p > q$, the error rate is dominated by ℓ_p sparsity, with S_q low-rankness playing a minor role. For $p \leq q$ and $m \geq d^2$, both the ℓ_p -ball and S_q -ball influence the estimation error, demonstrating an interplay between structured sparsity and low-rank regularization. When $m \leq d^2$, the error rate is dictated by S_q , making it independent of m , emphasizing the fundamental role of rank constraints in this regime.

5 Optimization for Dual Spectral Sparse Tensor Estimation

Efficiently solving tensor estimation problems with dual spectral sparsity (**RQ3**) is key to leveraging the proposed ℓ_p -Schatten- q quasi-norm in practice. However, this task presents substantial challenges due to the non-convexity and coupled structure of this regularization.

Challenges. Even in the vector setting, optimizing dual-level sparse structures is notoriously difficult due to the combination of *non-convexity* and *structural coupling* [15, 26]. In our tensor case, these challenges are further compounded by the need to simultaneously enforce inter-frequency sparsity and intra-frequency low-rankness. Most existing tensor optimization methods either treat frequency components independently or impose low-rank constraints without spectral sparsity considerations, making them ill-suited for the proposed dual-spectral regularization. The ℓ_p -Schatten- q quasi-norm is non-convex whenever $p, q \in (0, 1]$, ruling out standard convex optimization techniques and necessitating a structure-aware, non-convex optimization strategy.

To address these difficulties, our approach is naturally motivated by the structural properties of the problem. We adopt a *proximal update scheme* that takes advantage of the separability of the

²The conditions in each setting are provided in the appendix.

transform-domain representation $M(\underline{\mathbf{L}})$, allowing frequency-wise updates, along with an iterative reweighting strategy that facilitates optimization in the presence of non-convex regularization.

Proximal Operator Formulation. To handle the non-convex ℓ_p -Schatten- q regularization, we adopt a proximal update scheme that enforces dual spectral sparsity while remaining computationally efficient. Specifically, at iteration t , the update is given by solving:

$$\underline{\mathbf{L}}^{t+1} \in \arg \min_{\underline{\mathbf{L}}} \frac{1}{2} \|\underline{\mathbf{L}} - \underline{\mathbf{Z}}\|_{\text{F}}^2 + \lambda \sum_{k=1}^m \|M(\underline{\mathbf{L}})_{::,k}\|_{S_q}^{p/q}, \quad (9)$$

where $\underline{\mathbf{Z}}$ denotes the intermediate variable aggregating previous updates and gradient information. Since the transform $M(\cdot)$ allows slice-wise decomposition [20], Problem (9) reduces to m subproblems over frequency components $k \in [m]$:

$$\min_{\mathbf{A}_k} \frac{1}{2} \|\mathbf{A}_k - M(\underline{\mathbf{Z}})_{::,k}\|_{\text{F}}^2 + \lambda \|\mathbf{A}_k\|_{S_q}^{p/q}, \quad (10)$$

where $\mathbf{A}_k := M(\underline{\mathbf{L}})_{::,k}$ denotes the k -th frontal slice of the transformed tensor $M(\underline{\mathbf{L}})$. Problem (10) is difficult due to the non-convexity and lack of smoothness of the Schatten- q quasi-norm, which admits no closed-form or standard proximal solution in general.

To efficiently approximate Problem (10), we adopt a reweighted $\ell_{1/2}$ -surrogate for $\|\mathbf{A}_k\|_{S_q}^{p/q}$ based on singular values:

$$\sum_{i=1}^d w_{i,k} \cdot \sigma_i(\mathbf{A}_k)^{1/2}, \quad (11)$$

with weights defined as $w_{i,k} = (\sum_{j=1}^d \varsigma_{j,k}^q + \epsilon)^{p/q-1} \cdot (\varsigma_{i,k}^{1/2} + \epsilon)^{2q-1}$, where ϵ is a small regularization constant and $\varsigma_{j,k} := \sigma_j(M(\underline{\mathbf{L}}^t)_{::,k})$ are the singular values from the previous iterate. The update for each singular value then becomes a soft-thresholding step:

$$\sigma_i^{(t+1)}(M(\underline{\mathbf{L}})_{::,k}) = \mathcal{S}_{\lambda w_{i,k}}^{\ell_{1/2}}(\sigma_i(M(\underline{\mathbf{Z}})_{::,k})), \quad (12)$$

where $\mathcal{S}^{\ell_{1/2}}$ is the proximal operator for the $\ell_{1/2}$ -norm (see Appendix for closed-form expression).

After singular value shrinkage, we reconstruct each slice $M(\underline{\mathbf{L}}^{t+1})_{::,k} = \mathbf{U}_k \cdot \text{diag}(\boldsymbol{\sigma}^{(t+1)}) \cdot \mathbf{V}_k^\top$, where \mathbf{U}_k and \mathbf{V}_k are from the SVD of $M(\underline{\mathbf{Z}})_{::,k}$. Finally, applying the inverse transform yields the updated tensor $\underline{\mathbf{L}}^{t+1}$ in the original domain.

6 Experiments

Having established the theoretical foundations and algorithmic framework, we now evaluate the empirical performance of the proposed ℓ_p -Schatten- q quasi-norm in tensor estimation tasks. We conduct extensive experiments on three types of remote sensing data to demonstrate its effectiveness in noisy tensor completion tasks.

Experimental Setup. We consider the noisy tensor completion which involves reconstructing a tensor from noisy incomplete observations. Given a clean tensor $\underline{\mathbf{L}}$ of size $d_1 \times d_2 \times d_3$, we introduce *i.i.d.* Gaussian noise with standard deviation $\sigma = c\sigma_0$, where $c = 0.05$ and $\sigma_0 = \|\underline{\mathbf{L}}\|_{\text{F}} / \sqrt{d_1 d_2 d_3}$. A uniform sampling strategy is applied with sampling ratios $p \in \{0.05, 0.1, 0.15\}$, meaning that 95%, 90%, and 85% of the entries are missing, respectively. Each setting is tested over 10 trials, and the averaged PSNR (dB) and SSIM values are reported. To benchmark our method, we compare the proposed $\ell_p(S_q)$ -quasi-norm against several low-rank regularizers, including matrix nuclear norm (NN) [6], Tucker-based tensor nuclear norm (SNN) [27], TNN-DFT [62], TNN-DCT [31], tensor k -Support norm (k -Supp) ($k = 2$) [51], tensor ℓ_{1-2} -norm (ℓ_{1-2}) [42], tensor Schatten- p -norm ($p = 1/2$) [23]. In our implementation, we set the sparsity parameters³ to $(p, q) = (0.8961, 0.8966)$ and employ the Discrete Cosine Transform (DCT) as the transform operator $M(\cdot)$. Details of the experiments are given in the appendix.

³We first performed a coarse grid search over $p, q \in \{0.1, 0.2, \dots, 1.0\}$ and observed consistent performance peaks near $p = q = 0.9$. We then manually fine-tuned within $[0.88, 0.92]$ based on PSNR, selecting $(p, q) = (0.8961, 0.8966)$ as the best-performing pair.

Table 1: Results for noisy tensor completion on remote sensing datasets are shown below. The best result in each case is highlighted in **bold**, while the second-best is underlined.

Dataset	SR	Metric	NN	SNN	TNN-DFT	TNN-DCT	k-Supp	ℓ_{1-2}	Schatten-1/2	$\ell_p(S_q)$ (proposed)
SalinasA	5%	PSNR	15.21	20.79	22.55	26.52	22.58	22.21	22.45	28.43
		SSIM	0.2594	0.7547	0.5667	0.7384	0.5689	0.5524	0.4474	<u>0.7374</u>
	10%	PSNR	20.62	25.56	25.72	29.61	25.89	26.14	25.86	31.81
		SSIM	0.4775	0.8284	0.7027	<u>0.8403</u>	0.7231	0.7197	0.6058	0.8484
	15%	PSNR	23.09	27.99	28.06	<u>31.32</u>	28.09	28.13	26.98	33.23
		SSIM	0.5643	0.8622	0.7804	<u>0.8798</u>	0.7810	0.7795	0.6505	0.8830
IndianPines	5%	PSNR	20.44	22.01	25.68	<u>26.26</u>	25.70	25.73	24.68	27.05
		SSIM	0.3895	0.6359	0.6293	<u>0.6727</u>	0.6289	0.6316	0.5361	0.6740
	10%	PSNR	22.23	24.94	27.45	<u>28.40</u>	27.48	27.52	25.72	28.92
		SSIM	0.4836	0.7171	0.7226	0.7744	0.7219	0.7249	0.5991	<u>0.7617</u>
	15%	PSNR	23.52	26.61	28.54	<u>29.52</u>	28.53	28.63	26.24	29.89
		SSIM	0.5438	0.7668	0.7713	0.8177	0.7709	0.7741	0.6258	<u>0.7997</u>
Cloth	5%	PSNR	20.10	20.95	25.00	26.09	25.08	25.09	24.96	26.99
		SSIM	0.3762	0.5096	0.6773	<u>0.7283</u>	0.6792	0.6793	0.6305	0.7422
	10%	PSNR	21.14	22.72	28.00	<u>29.24</u>	28.12	28.14	27.98	30.63
		SSIM	0.4341	0.5983	0.8132	<u>0.8540</u>	0.8143	0.8163	0.7668	0.8658
	15%	PSNR	22.05	24.18	30.03	<u>31.36</u>	30.08	30.11	29.50	32.71
		SSIM	0.4889	0.6783	0.8722	<u>0.9054</u>	0.8727	0.8733	0.8153	0.9090
Hair	5%	PSNR	25.33	30.09	33.16	<u>35.31</u>	33.19	33.27	33.43	36.95
		SSIM	0.7147	0.8631	0.8917	0.9248	0.8921	0.8919	0.8240	<u>0.9196</u>
	10%	PSNR	29.52	33.35	36.22	<u>38.18</u>	36.17	36.30	35.69	39.91
		SSIM	0.8008	0.9122	0.9292	0.9535	0.9286	0.9296	0.8640	<u>0.9517</u>
	15%	PSNR	31.12	35.24	38.00	<u>39.88</u>	37.91	38.07	36.46	41.52
		SSIM	0.8364	0.9336	0.9449	0.9650	0.9442	0.9448	0.8735	<u>0.9641</u>
JellyBeans	5%	PSNR	16.33	18.21	25.43	<u>26.47</u>	25.38	25.62	25.39	27.91
		SSIM	0.2397	0.4942	0.6726	0.7223	0.6714	0.6733	0.5504	<u>0.7115</u>
	10%	PSNR	18.12	22.11	28.50	<u>30.14</u>	28.47	28.67	28.41	31.95
		SSIM	0.3169	0.6629	0.7900	0.8518	0.7902	0.7932	0.6905	<u>0.8486</u>
	15%	PSNR	19.92	24.67	30.51	<u>32.33</u>	30.52	30.61	29.96	33.97
		SSIM	0.4053	0.7592	0.8489	0.9030	0.8504	0.8499	0.7516	<u>0.8980</u>
OSU Thermal	5%	PSNR	13.19	15.83	28.06	27.99	28.01	28.19	28.11	30.06
		SSIM	0.1848	0.4759	0.8584	<u>0.8707</u>	0.8579	0.8603	0.7928	0.8759
	10%	PSNR	14.67	19.75	31.30	<u>31.62</u>	31.28	31.60	30.51	33.67
		SSIM	0.2509	0.6594	0.9151	0.9326	0.9147	0.9168	0.8358	<u>0.9272</u>
	15%	PSNR	16.27	22.52	33.02	<u>33.51</u>	33.05	33.11	30.99	35.09
		SSIM	0.3273	0.7621	0.9315	0.9509	0.9321	0.9318	0.8373	<u>0.9404</u>

Datasets. We validate our approach on three categories of remote sensing data. First, for hyperspectral images, we employ the corrected *Indian Pines* and *Salinas A* datasets from the AVIRIS sensor, containing 200 and 204 spectral bands respectively. Due to computational considerations, we utilize the first 30 bands in our experiments. Second, we evaluate on multispectral images from the Columbia MSI Database, including *Cloth*, *Hair*, and *Jelly Beans*, each with dimensions $512 \times 512 \times 31$ and normalized intensity values in $[0,1]$. Finally, for thermal imaging, we use sequences from the *OSU Thermal Database*, specifically the first 30 frames of Sequence 1, forming a tensor of size $320 \times 240 \times 30$.

Results and Analysis. Table 1 summarizes the PSNR and SSIM results across different missing rates. The proposed $\ell_p(S_q)$ -quasi-norm achieves the highest PSNR, demonstrating its effectiveness in preserving spectral information. Its SSIM results rank among the top two, indicating that our approach better retains structural integrity compared to competing methods. These experimental results demonstrate the effectiveness of the proposed ℓ_p -Schatten- q quasi-norm in robust tensor recovery, showing how characterizing dual spectral sparsity structures in transformed domains benefits tensor reconstruction performance.

7 Conclusion

This paper identifies and formalizes a coupled spectral structure within the t-SVD framework, where inter-frequency sparsity coexists with intra-frequency low-rankness. To capture this structure, we propose a unified modeling approach based on the ℓ_p -Schatten- q quasi-norm, which enables separate control over spectral sparsity at different levels and generalizes existing tensor norms. We provide sharp minimax guarantees under both hard and soft sparsity regimes, and develop an efficient proximal algorithm tailored to this setting. Experimental results demonstrate the practical potential of the proposed approach for structured tensor recovery.

Limitation. To highlight the fundamental properties of the proposed ℓ_p -Schatten- q quasi-norm, our analysis employs several simplifications, including Gaussian location model and idealized sparsity patterns. While our optimization algorithm shows promising empirical performance, its theoretical convergence properties remain to be established. These theoretical and algorithmic limitations suggest important directions for future research.

References

- [1] B. Barak and A. Moitra. Noisy tensor completion via the sum-of-squares hierarchy. In *Conference on Learning Theory*, pages 417–445. PMLR, 2016.
- [2] G. Bergqvist and E. G. Larsson. The higher-order singular value decomposition: Theory and an application [lecture notes]. *IEEE signal processing magazine*, 27(3):151–154, 2010.
- [3] P. Breheny and J. Huang. Penalized methods for bi-level variable selection. *Statistics and its interface*, 2(3):369, 2009.
- [4] C. Cai, G. Li, H. V. Poor, and Y. Chen. Nonconvex low-rank tensor completion from noisy data. In *Advances in Neural Information Processing Systems*, volume 32, 2019.
- [5] J.-F. Cai, W. Huang, H. Wang, and K. Wei. Tensor completion via tensor train based low-rank quotient geometry under a preconditioned metric. *arXiv preprint arXiv:2209.04786*, 2022.
- [6] E. J. Candès and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.
- [7] J. D. Carroll and J. Chang. Analysis of individual differences in multidimensional scaling via an n -way generalization of “Eckart-Young” decomposition. *Psychometrika*, 35(3):283–319, 1970.
- [8] A. Cichocki, N. Lee, I. Oseledets, A. H. Phan, Q. Zhao, and D. P. Mandic. Tensor networks for dimensionality reduction and large-scale optimization: Part 1 low-rank tensor decompositions. *Foundations & Trends® in Machine Learning*, 9(4-5):249–429, 2016.
- [9] D. E. Edmunds and Y. Netrusov. Schütt’s theorem for vector-valued sequence spaces. *Journal of Approximation Theory*, 178:13–21, 2014.
- [10] S. Gandy, B. Recht, and I. Yamada. Tensor completion and low-n-rank tensor recovery via convex optimization. *Inverse Problems*, 27(2):025010, 2011.
- [11] S. Gandy, B. Recht, and I. Yamada. Tensor completion and low-n-rank tensor recovery via convex optimization. *Inverse Problems*, 27(2):025010, 2011.
- [12] S. A. Geer, S. van de Geer, and D. Williams. *Empirical Processes in M-estimation*, volume 6. Cambridge university press, 2000.
- [13] A. Hinrichs, J. Prochno, and J. Vybiral. Entropy numbers of embeddings of Schatten classes. *Journal of Functional Analysis*, 273(10):3241–3261, 2017.
- [14] J. Hou, F. Zhang, H. Qiu, J. Wang, Y. Wang, and D. Meng. Robust low-tubal-rank tensor recovery from binary measurements. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.
- [15] Y. Hu, C. Li, K. Meng, J. Qin, and X. Yang. Group sparse optimization via $\ell_{p,q}$ regularization. *Journal of Machine Learning Research*, 18(30):1–52, 2017.
- [16] J. Huang, S. Ma, H. Xie, and C.-H. Zhang. A group bridge approach for variable selection. *Biometrika*, 96(2):339–355, 2009.
- [17] Y. Huo, L. Xin, C. Kang, M. Wang, Q. Ma, and B. Yu. Sgl-svm: a novel method for tumor classification via support vector machine with sparse group lasso. *Journal of Theoretical Biology*, 486:110098, 2020.
- [18] M. Imaizumi, T. Maehara, and K. Hayashi. On tensor train rank minimization: Statistical efficiency and scalable algorithm. In *Advances in Neural Information Processing Systems*, pages 3930–3939, 2017.
- [19] E. Kernfeld, M. Kilmer, and S. Aeron. Tensor–tensor products with invertible linear transforms. *Linear Algebra and its Applications*, 485:545–570, 2015.
- [20] M. E. Kilmer, K. Braman, et al. Third-order tensors as operators on matrices: A theoretical and computational framework with applications in imaging. *SIAM J MATRIX ANAL A*, 34(1):148–172, 2013.
- [21] O. Klopp. Matrix completion by singular value thresholding: sharp bounds. *Electronic Journal of Statistics*, 9(2):2348–2369, 2015.
- [22] T. G. Kolda and B. W. Bader. Tensor decompositions and applications. *SIAM Review*, 51(3):455–500, 2009.

- [23] H. Kong, X. Xie, and Z. Lin. t-Schatten- p norm for low-rank tensor recovery. *IEEE Journal of Selected Topics in Signal Processing*, 12(6):1405–1419, 2018.
- [24] C. Li, W. He, L. Yuan, Z. Sun, and Q. Zhao. Guaranteed matrix completion under multiple linear transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11136–11145, 2019.
- [25] Z. Li, Y. Zhang, and J. Yin. Estimating double sparse structures over $\ell_u(\ell_q)$ -balls: Minimax rates and phase transition. *IEEE Transactions on Information Theory*, 2024.
- [26] R. Lin, S. Chen, H. Feng, and Y. Liu. Computing the proximal operator of the $\ell_{p,q}$ -norm for group sparsity. *arXiv preprint arXiv:2409.14156*, 2024.
- [27] J. Liu, P. Musialski, P. Wonka, and J. Ye. Tensor completion for estimating missing values in visual data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):208–220, 2013.
- [28] X. Liu, S. Aeron, V. Aggarwal, and X. Wang. Low-tubal-rank tensor completion using alternating minimization. *IEEE Transactions on Information Theory*, 66(3):1714–1737, 2020.
- [29] C. Lu. Transforms based tensor robust PCA: Corrupted low-rank tensors recovery via convex optimization. In *ICCV*, pages 1145–1152, 2021.
- [30] C. Lu, J. Feng, W. Liu, Z. Lin, S. Yan, et al. Tensor robust principal component analysis with a new tensor nuclear norm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.
- [31] C. Lu, X. Peng, and Y. Wei. Low-rank tensor completion with a new tensor nuclear norm induced by invertible linear transforms. In *CVPR*, pages 5996–6004, 2019.
- [32] C. D. Martin, R. Shafer, and B. Larue. An order- p tensor factorization with applications in imaging. *SIAM Journal on Scientific Computing*, 35(1), 2013.
- [33] S. Negahban and M. J. Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *The Journal of Machine Learning Research*, 13:1665–1697, 2012.
- [34] I. V. Oseledets. Tensor-train decomposition. *SIAM Journal on Scientific Computing*, 33(5):2295–2317, 2011.
- [35] Y. Qiu, G. Zhou, A. Wang, Q. Zhao, and S. Xie. Balanced unfolding induced tensor nuclear norms for high-order tensor completion. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.
- [36] N. Rao, R. Nowak, C. Cox, and T. Rogers. Classification with the sparse group lasso. *IEEE Transactions on Signal Processing*, 64(2):448–463, 2015.
- [37] G. Raskutti, M. J. Wainwright, and B. Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q -balls. *IEEE Transactions on Information Theory*, 57(10):6976–6994, 2011.
- [38] B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, 2007.
- [39] M. Silver, P. Chen, R. Li, C.-Y. Cheng, T.-Y. Wong, E.-S. Tai, Y.-Y. Teo, and G. Montana. Pathways-driven sparse regression identifies pathways and genes associated with high-density lipoprotein cholesterol in two asian cohorts. *PLoS genetics*, 9(11):e1003939, 2013.
- [40] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani. A sparse-group lasso. *Journal of computational and graphical statistics*, 22(2):231–245, 2013.
- [41] G. Song, M. K. Ng, and X. Zhang. Robust tensor completion using transformed tensor singular value decomposition. *NUMER LINEAR ALGEBRA*, 27(3):e2299, 2020.
- [42] Z. Tan, L. Huang, H. Cai, and Y. Lou. Non-convex approaches for low-rank tensor completion under tubal sampling. In *ICASSP*, pages 1–5. IEEE, 2023.
- [43] M. Thomas and A. T. Joy. *Elements of information theory*. Wiley-Interscience, 2006.
- [44] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.
- [45] T. Tony Cai, A. R. Zhang, and Y. Zhou. Sparse group lasso: Optimal sample complexity, convergence rate, and statistical inference. *IEEE Transactions on Information Theory*, pages 1–1, 2022.

- 448 [46] A. B. Tsybakov and Schatzen. *Introduction to Nonparametric Estimation*. PhD thesis, Springer New York,
449 2011.
- 450 [47] L. R. Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3):279–311,
451 1966.
- 452 [48] J. K. Tugnait. Sparse-group lasso for graph learning from multi-attribute data. *IEEE Transactions on*
453 *Signal Processing*, 69:1771–1786, 2021.
- 454 [49] A. Wang, Z. Jin, and G. Tang. Robust tensor decomposition via t-SVD: Near-optimal statistical guarantee
455 and scalable algorithms. *Signal Processing*, 167:107319, 2020.
- 456 [50] A. Wang, C. Li, M. Bai, Z. Jin, G. Zhou, and Q. Zhao. Transformed low-rank parameterization can help
457 robust generalization for tensor neural networks. *NeurIPS*, 36, 2023.
- 458 [51] A. Wang, G. Zhou, Z. Jin, and Q. Zhao. Tensor recovery via $*_l$ -spectral k -support norm. *IEEE Journal of*
459 *Selected Topics in Signal Processing*, 15(3):522–534, 2021.
- 460 [52] H. Wang, J. Yang, X. Yu, Y. Zhang, J. Qian, and J. Wang. Tensor-flamingo unravels the complexity of
461 single-cell spatial architectures of genomes at high-resolution. *Nature Communications*, 16(1):3435, 2025.
- 462 [53] Z. Wang, J. Dong, X. Liu, and X. Zeng. Low-rank tensor completion by approximating the tensor average
463 rank. In *ICCV*, pages 4612–4620, 2021.
- 464 [54] Y. Wu. Lecture notes on information-theoretic methods for high-dimensional statistics. *Lecture Notes for*
465 *ECE598YW (UIUC)*, 16, 2017.
- 466 [55] D. Xia and M. Yuan. On polynomial time methods for exact low-rank tensor completion. *Foundations of*
467 *Computational Mathematics*, 19(6):1265–1313, 2019.
- 468 [56] Y. Xie, D. Tao, W. Zhang, Y. Liu, L. Zhang, and Y. Qu. On unifying multi-view self-representations for
469 clustering by tensor multi-rank minimization. *International Journal of Computer Vision*, 126(11):1157–
470 1179, 2018.
- 471 [57] Y. Xu, R. Hao, W. Yin, and Z. Su. Parallel matrix factorization for low-rank tensor completion. *Inverse*
472 *Problems and Imaging*, 9(2):601–624, 2015.
- 473 [58] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the*
474 *Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.
- 475 [59] M. Yuan and C. H. Zhang. On tensor completion via nuclear norm minimization. *Foundations of*
476 *Computational Mathematics*, 16(4):1–38, 2016.
- 477 [60] F. Zhang, W. Wang, J. Huang, J. Wang, and Y. Wang. RIP-based performance guarantee for low-tubal-rank
478 tensor recovery. *Journal of Computational and Applied Mathematics*, 374:112767, 2020.
- 479 [61] X. Zhang and M. K.-P. Ng. Low rank tensor completion with poisson observations. *IEEE Transactions on*
480 *Pattern Analysis and Machine Intelligence*, 2021.
- 481 [62] Z. Zhang and S. Aeron. Exact tensor completion using t-svd. *IEEE Transactions on Signal Processing*,
482 65(6):1511–1526, 2017.
- 483 [63] Z. Zhang, G. Ely, S. Aeron, et al. Novel methods for multilinear data completion and de-noising based on
484 tensor-SVD. In *CVPR*, pages 3842–3849, 2014.
- 485 [64] P. Zhou and J. Feng. Outlier-robust tensor PCA. In *Proceedings of the IEEE Conference on Computer*
486 *Vision and Pattern Recognition*, 2017.

Appendix for *Refining Dual Spectral Sparsity in Transformed Tensor Singular Values*

This appendix offers a comprehensive supplement to the main paper, elaborating on the theoretical foundations, algorithmic design, and empirical validation of our proposed framework. The materials are organized to closely follow the development of our contributions, and are structured as follows:

- **Related Work and Problem Context (Section A).** We situate our formulation within the broader landscape of tensor recovery and double sparsity literature. In particular, we highlight the conceptual and technical departures from prior tensor nuclear norm (TNN) models and vector-based $\ell_u(\ell_q)$ regularization, emphasizing the unique challenges introduced by the coupled spectral structure we study.
- **Notations and Technical Preliminaries (Section B).** We formalize key notations under the t-SVD framework and present supporting lemmas that underlie our entropy constructions and minimax risk analysis.
- **Theoretical Results and Proofs (Section C).** We provide detailed derivations of the minimax lower and upper bounds for tensor estimation under dual spectral sparsity assumptions. These results clarify how different regimes of (p, q) give rise to distinct statistical rates and phase transitions.
- **Optimization and Experimental Details (Section D).** We describe the implementation of the ADMM-based optimization algorithm, including the reweighted $\ell_{1/2}$ -surrogate for handling the nonconvex $\ell_p(S_q)$ regularizer. We also report parameter tuning strategies, convergence diagnostics, and additional performance metrics to support the empirical findings in the main text.

509	Appendix Contents	
510	1 Introduction	1
511	2 Notations and Preliminaries	3
512	3 Dual Spectral Sparsity in the t-SVD Framework	4
513	4 Theory of Dual Spectral Sparse Tensor Estimation	5
514	4.1 Gaussian Location Model	6
515	4.2 Minimax Risk over Dual-level Sparse Structures	6
516	5 Optimization for Dual Spectral Sparse Tensor Estimation	7
517	6 Experiments	8
518	7 Conclusion	9
519	A Comparison with Prior Work and Unique Technical Challenges	15
520	A.1 General Context and Related Work	15
521	A.2 Why Our Framework Is Not a Direct Extension of TNN or $\ell_u(\ell_q)$ Norms	15
522	A.3 New Technical Challenges	16
523	B Additional Notations, Preliminaries and Lemmas	17
524	B.1 Preliminaries of t-Singular Value Decomposition	17
525	B.2 Additional Lemmas	18
526	C Theoretical Results for Understanding of Dual-Level Sparsity for GLM	19
527	C.1 Sketch of Minimax Proofs under Dual-Level Sparsity	21
528	C.2 Proof of Theorem C.1	22
529	C.2.1 Proof of case (a)	22
530	C.2.2 Theoretical Tools: Gilbert-Varshamov Theorems	22
531	C.2.3 Proof of case (b)	25
532	C.3 Proofs of Theorem C.2	26
533	C.3.1 Technical Lemmas for upper bounds	26
534	C.3.2 Proof of C.2	31
535	C.4 Proof of Theorem C.4	32
536	C.4.1 Lower bound for $\ell_p(S_q)$	32
537	C.4.2 Covering number of $\mathbb{B}_{\ell_p(S_q)}(R)$	33
538	C.4.3 Proof of upper bounds for $\ell_p(S_q)$	36
539	D Experimental Details	38
540	D.1 Experimental Setup	38
541	D.2 Tensor Completion via an ADMM-Based Algorithm	41
542	D.3 Choice of Weighted $\ell_{1/2}$ Approximation	43
543	D.4 Convergence of the Proposed ADMM-Based Algorithm	43

A Comparison with Prior Work and Unique Technical Challenges

A.1 General Context and Related Work

Our work intersects with two major research areas: double sparse structures and tensor recovery methods. We discuss each in turn and highlight how our work advances these fields.

Double Sparse Structures. Research on double sparse structures has demonstrated their effectiveness in capturing hierarchical sparsity patterns across various domains. In genomics, these structures model pathway-level and SNP-level sparsity in genome-wide association studies [39], while similar hierarchical patterns appear in classification tasks [36, 17] and network analysis [48]. Methodologically, bi-level selection approaches [3] have dominated this field, evolving from the fundamental group bridge method [16] to more sophisticated techniques like sparse group lasso [40], which unified individual [44] and group-level [58] sparsity. Recent theoretical work by [45] has established minimax bounds for double sparse regression, building on earlier investigations of general sparsity structures [37]. [25] developed fundamental theoretical bounds for high-dimensional double sparse structures by establishing novel metric entropy bounds over $\ell_u(\ell_q)$ -balls using Gilbert-Varshamov techniques, providing insights into the simultaneous estimation of group-wise and element-wise sparsity. However, these approaches, while successful for vector and matrix data, cannot address the fundamental challenge in transformed-domain tensor analysis: how to simultaneously model sparsity across different frequency components and low-rank structures within each frequency component—a gap our work bridges through the ℓ_p -Schatten- q framework.

Tensor Recovery Methods. Tensor recovery research has evolved along multiple methodological paths, each addressing different aspects of multi-dimensional data analysis. Traditional approaches utilize various decomposition frameworks: CP decomposition with techniques ranging from sum-of-squares [1] to gradient descent [4], Tucker decomposition employing nuclear norm minimization [11] and manifold optimization [55], and tensor train/ring decompositions using Riemannian methods [5]. Recently, low-tubal-rank recovery has gained attention, with methods spanning both convex approaches like tensor nuclear norm minimization [30] and non-convex alternatives such as Schatten- p norm regularization [23]. However, existing tensor methods, particularly those based on tensor nuclear norm (TNN), apply uniform regularization across all frequencies in the transformed domain. While effective, this single-level sparsity treatment can be further enhanced to fully capture the natural hierarchical patterns in real-world tensor data - for example, in hyperspectral images where some frequency bands carry more information than others while each band itself exhibits low-rank structure. Our ℓ_p -Schatten- q framework extends TNN’s uniform regularization by introducing separate parameters to control sparsity across frequencies (p) and low-rank structure within each frequency (q).

A.2 Why Our Framework Is Not a Direct Extension of TNN or $\ell_u(\ell_q)$ Norms

While our formulation includes the Tensor Nuclear Norm (TNN) [31, 63] as a special case when $(p, q) = (1, 1)$ and shares some similarities with $\ell_u(\ell_q)$ norms [25] used in vector models, it is *structurally different* from both. In particular, our framework captures a new interaction: sparsity across frequency slices and low-rankness within each slice are jointly modeled and cannot be separated. This type of coupling does not appear in previous models and leads to new technical challenges in both theory and algorithm design. We explain the key differences below.

1. **Comparison with TNN.** The TNN penalizes the sum of all singular values across transformed slices uniformly, assuming that each frontal slice contributes equally and independently. This leads to a separable regularization structure, where each slice is processed in isolation. In contrast, our proposed ℓ_p -Schatten- q quasi-norm:

- imposes sparsity across frequency slices via the outer ℓ_p term;
- induces per-slice low-rankness through the inner Schatten- q quasi-norm;
- couples these two effects non-separably, such that frequency selection and rank penalization jointly influence the model.

This interdependence alters both the structure of the parameter space and the behavior of the regularizer, breaking the assumptions underpinning existing theoretical and algorithmic analyses of TNN.

2. **Comparison with $\ell_u(\ell_q)$ norms.** The $\ell_u(\ell_q)$ class has been studied extensively for vector-valued and matrix-valued double sparse models, where group structures are known and fixed [25]. Our setting differs fundamentally in both domain and geometry:

- The group structure (i.e., frequency slices) is not fixed a priori, but emerges from a linear transform applied to the third mode;
- Each group is a matrix with structured spectral decay, rather than a vector block;
- The regularization interacts with both the spectral geometry and the transform, which invalidates direct use of existing entropy results.

These differences imply that established results for $\ell_u(\ell_q)$ -norm regularization do not extend to our setting, and necessitate new techniques tailored to transform-domain tensor analysis.

This distinction in modeling leads to substantial new technical challenges, which we outline next.

A.3 New Technical Challenges

The coupled spectral structure in our framework introduces analytical and algorithmic challenges that go beyond prior work. These arise across three key aspects:

1. **Coupled parameter geometry and entropy analysis.** The induced constraint set is no longer a product of independent low-dimensional balls, but a *hierarchically coupled* space shaped by both inter-group and intra-group constraints. To analyze its minimax complexity:
 - We construct nonstandard packing sets that jointly encode group selection and low-rank structure;
 - We extend Gilbert–Varshamov-based arguments to spectral-domain tensor geometry (see Lemmas C.9–C.15);
 - Our bounds reveal multiple scaling regimes depending on (p, q) and tensor size, which are not captured by standard sparsity or low-rank models.
2. **Coupled structure complicates optimization design.** Unlike models such as TNN, where the regularizer decouples across frequency slices and admits closed-form proximal mappings, our $\ell_p(S_q)$ quasi-norm introduces two intertwined sources of difficulty:
 - First, the nesting of ℓ_p and Schatten- q induces nonconvexity and nonseparability, for which no closed-form proximal operator is available—even when one of the parameters is fixed;
 - Second, the global ℓ_p term couples the optimization across slices, meaning that the regularization strength on one slice implicitly depends on the spectrum of others, breaking slice-wise independence.

To address these challenges, we introduce a reweighted surrogate based on the $\ell_{1/2}$ -norm, enabling efficient frequency-wise updates that approximate the original regularization while preserving its structural intent. These updates are embedded into an ADMM framework that maintains computational tractability and spectral coherence across slices.

These challenges preclude the direct reuse of existing matrix or tensor tools, and motivate the new statistical and optimization techniques developed in this work.

B Additional Notations, Preliminaries and Lemmas

Additional Notations We use several asymptotic notations to describe the relationships between functions. For the sake of clarity, we provide their definitions here:

- The notation $f(n) \lesssim g(n)$ means that there exists a positive constant c and a positive integer n_0 such that for all $n \geq n_0$, we have $f(n) \leq c \cdot g(n)$. This is equivalent to saying $f(n) = O(g(n))$.
- Similarly, $f(n) \gtrsim g(n)$ means that there exists a positive constant c and a positive integer n_0 such that for all $n \geq n_0$, we have $f(n) \geq c \cdot g(n)$. This is equivalent to saying $g(n) = O(f(n))$.
- We write $f(n) \asymp g(n)$ if both $f(n) \lesssim g(n)$ and $f(n) \gtrsim g(n)$ hold. This means that $f(n)$ and $g(n)$ are of the same order.
- The notation $f(n) = o(g(n))$ means that for every positive constant ϵ , there exists a positive integer n_0 such that for all $n \geq n_0$, we have $|f(n)| \leq \epsilon \cdot |g(n)|$.

These notations allow us to express the asymptotic behavior of functions concisely, which is particularly useful in our analysis of algorithmic complexity and error bounds.

B.1 Preliminaries of t-Singular Value Decomposition

Due to space limitations, some concepts related to t-SVD were omitted in the main text. We provide additional notions here.

Definition B.1 (Frontal-slice-wise product [30]). The frontal-slice-wise product of any two tensors $\underline{\mathbf{A}} \in \mathbb{R}^{d_1 \times d_2 \times m}$ and $\underline{\mathbf{B}} \in \mathbb{R}^{d_1 \times d_2 \times m}$, denoted by $\underline{\mathbf{A}} \odot \underline{\mathbf{B}}$, is defined as a tensor $\underline{\mathbf{T}}$ such that

$$\underline{\mathbf{T}}_{::,i} = \underline{\mathbf{A}}_{::,i} \cdot \underline{\mathbf{B}}_{::,i}, \quad i \in [K]$$

where \cdot denotes the standard matrix multiplication. The frontal-slice-wise product performs matrix multiplication on each frontal slice of the tensors, resulting in a new tensor.

Definition B.2 (M -block-diagonal matrix). The M -block-diagonal matrix of any tensor $\underline{\mathbf{T}} \in \mathbb{R}^{d_1 \times d_2 \times m}$, denoted by $\bar{\mathbf{T}}$, is the block diagonal matrix whose diagonal blocks are the frontal slices of $M(\underline{\mathbf{T}})$:

$$\bar{\mathbf{T}} := \text{bdiag}(M(\underline{\mathbf{T}})) := \begin{bmatrix} M(\underline{\mathbf{T}})_{::,1} & & & \\ & M(\underline{\mathbf{T}})_{::,2} & & \\ & & \ddots & \\ & & & M(\underline{\mathbf{T}})^{(K)} \end{bmatrix} \in \mathbb{R}^{d_1 m \times d_2 m}.$$

This concept arranges the slices of a tensor in the frequency domain into a block diagonal matrix, facilitating the theoretical analysis of t-SVD.

We further provide some definitions and properties related to t-SVD:

Definition B.3 ([19]). The t-transpose of a tensor $\underline{\mathbf{T}} \in \mathbb{R}^{d_1 \times d_2 \times m}$ under the M transform (as shown in (1)), denoted by $\underline{\mathbf{T}}^\top$, satisfies

$$M(\underline{\mathbf{T}}^\top)_{::,i} = (M(\underline{\mathbf{T}})_{::,i})^\top, \quad i \in [K].$$

In other words, the t-transpose performs a transpose on each slice in the frequency domain and then transforms back to the time domain. This operation is one of the foundations of t-SVD theory.

Definition B.4 ([19]). The t-identity tensor $\underline{\mathbf{I}} \in \mathbb{R}^{d \times d \times m}$ under the M transform satisfies that each frontal slice of $M(\underline{\mathbf{I}})$ is an $m \times m$ identity matrix, i.e.,

$$M(\underline{\mathbf{I}})_{::,i} = \mathbf{I}, \quad i \in [K].$$

It is easy to verify that $\underline{\mathbf{T}} *_M \underline{\mathbf{I}} = \underline{\mathbf{T}}$ and $\underline{\mathbf{I}} *_M \underline{\mathbf{T}} = \underline{\mathbf{T}}$ hold for appropriate dimensions. The t-identity tensor plays a role similar to the identity matrix in t-SVD.

Definition B.5 ([19]). A tensor $\underline{\mathbf{Q}} \in \mathbb{R}^{d \times d \times m}$ is called t-orthogonal under the M transform if it satisfies

$$\underline{\mathbf{Q}}^\top *_M \underline{\mathbf{Q}} = \underline{\mathbf{Q}} *_M \underline{\mathbf{Q}}^\top = \underline{\mathbf{I}}.$$

T-orthogonality is an important property of tensor transformations, ensuring that the inner product and norm of tensors remain invariant before and after the transformation.

665 **Decomposability of Tensor Nuclear Norm** Consider the reduced t-SVD of $\underline{\mathbf{L}}^*$ given by

$$\underline{\mathbf{L}}^* = \underline{\mathbf{U}} *_M \underline{\mathbf{S}} *_M \underline{\mathbf{V}}^\top$$

666 where $\underline{\mathbf{U}} \in \mathbb{R}^{d_1 \times r \times m}$ and $\underline{\mathbf{V}} \in \mathbb{R}^{d_2 \times r \times m}$ are orthogonal tensors, and $\underline{\mathbf{S}} \in \mathbb{R}^{r \times r \times m}$ is an f-diagonal
667 tensor. We define the projection operators $\mathcal{P}_*(\cdot)$ and $\mathcal{P}_{*\perp}(\cdot)$ as follows:

$$\mathcal{P}_*(\underline{\mathbf{T}}) = \underline{\mathbf{U}} *_M \underline{\mathbf{U}}^\top *_M \underline{\mathbf{T}} + \underline{\mathbf{T}} *_M \underline{\mathbf{V}} *_M \underline{\mathbf{V}}^\top - \underline{\mathbf{U}} *_M \underline{\mathbf{U}}^\top *_M \underline{\mathbf{T}} *_M \underline{\mathbf{V}} *_M \underline{\mathbf{V}}^\top \quad (13)$$

$$\mathcal{P}_{*\perp}(\underline{\mathbf{T}}) = (\underline{\mathbf{I}} - \underline{\mathbf{U}} *_M \underline{\mathbf{U}}^\top) *_M \underline{\mathbf{T}} *_M (\underline{\mathbf{I}} - \underline{\mathbf{V}} *_M \underline{\mathbf{V}}^\top). \quad (14)$$

668 These operators decompose the tensor $\underline{\mathbf{T}}$ into components aligned with the sub-modules t-spanned by
669 $\underline{\mathbf{U}}$ and $\underline{\mathbf{V}}$, and their orthogonal complements, respectively.

670 As shown in the appendix of [49], the following properties hold:

- 671 a). Any tensor $\underline{\mathbf{T}} \in \mathbb{R}^{d_1 \times d_2 \times m}$ can be uniquely decomposed as $\underline{\mathbf{T}} = \mathcal{P}_*(\underline{\mathbf{T}}) + \mathcal{P}_{*\perp}(\underline{\mathbf{T}})$.
- 672 b). The inner product between the projections $\mathcal{P}_*(\underline{\mathbf{X}})$ and $\mathcal{P}_{*\perp}(\underline{\mathbf{Y}})$ is zero, i.e.,
673 $\langle \mathcal{P}_*(\underline{\mathbf{X}}), \mathcal{P}_{*\perp}(\underline{\mathbf{Y}}) \rangle = 0$, for all tensors $\underline{\mathbf{X}}, \underline{\mathbf{Y}} \in \mathbb{R}^{d_1 \times d_2 \times m}$.
- 674 c). The tubal rank of the projected tensor $\mathcal{P}_*(\underline{\mathbf{T}})$ is at most twice the rank of $\underline{\mathbf{L}}^*$, i.e., $r_{\text{tb}}(\mathcal{P}_*(\underline{\mathbf{T}})) \leq$
675 $2 \cdot r_{\text{tb}}(\underline{\mathbf{L}}^*)$, for all $\underline{\mathbf{T}} \in \mathbb{R}^{d_1 \times d_2 \times m}$.

676 Additionally, the following properties related to the tensor nuclear norm (TNN) can be established:

- 677 a). **(Decomposability of TNN)** For any tensors $\underline{\mathbf{X}}, \underline{\mathbf{Y}} \in \mathbb{R}^{d_1 \times d_2 \times m}$ satisfying $\underline{\mathbf{X}} *_M \underline{\mathbf{Y}}^\top = 0$ and
678 $\underline{\mathbf{X}}^\top *_M \underline{\mathbf{Y}} = 0$, the tensor nuclear norm decomposes additively:

$$\|\underline{\mathbf{X}} + \underline{\mathbf{Y}}\|_* = \|\mathcal{P}_*(\underline{\mathbf{X}})\|_* + \|\mathcal{P}_{*\perp}(\underline{\mathbf{Y}})\|_*.$$

- 679 b). **(Norm compatibility inequality)** For any tensor $\underline{\mathbf{T}} \in \mathbb{R}^{d_1 \times d_2 \times m}$, the tensor nuclear norm can
680 be related to the tensor Frobenius norm and the tensor rank as follows:

$$\|\underline{\mathbf{T}}\|_* \leq \sqrt{r_{\text{tb}}(\underline{\mathbf{T}}) \cdot m} \cdot \|\underline{\mathbf{T}}\|_F.$$

681 B.2 Additional Lemmas

682 The concept of covering and packing numbers play an important role in our remaining analysis.

683 **Definition B.6** (Covering and Packing Numbers, [37]). Consider a compact metric space consisting
684 of a set \mathcal{S} and a metric $\varrho : \mathcal{S} \times \mathcal{S} \rightarrow \mathbb{R}^+$

- 685 • An ϵ -covering of \mathcal{S} with respect to the metric ϱ is a collection $\{\underline{\mathbf{L}}^1, \dots, \underline{\mathbf{L}}^N\} \subset \mathcal{S}$ such that
686 for all $\underline{\mathbf{L}} \in \mathcal{S}$, there exists some $\underline{\mathbf{L}}^i, i \in \{1, \dots, N\}$ with $\varrho(\underline{\mathbf{L}}, \underline{\mathbf{L}}^i) \leq \epsilon$. The ϵ -covering number
687 $N(\epsilon; \mathcal{S}, \varrho)$ is the cardinality of the smallest ϵ -covering.
- 688 • A δ -packing of \mathcal{S} with respect to the metric ϱ is a collection $\{\underline{\mathbf{L}}^1, \dots, \underline{\mathbf{L}}^M\} \subset \mathcal{S}$ such that
689 $\varrho(\underline{\mathbf{L}}^i, \underline{\mathbf{L}}^j) > \delta$ for all distinct i, j . The δ -packing number $M(\delta; \mathcal{S}, \varrho)$ is the cardinality of the
690 largest δ -packing.

691 Covering and packing numbers provide essentially the same measure of the massiveness of a
692 set. In particular, the relation between covering number and packing number is described as
693 $M(2\epsilon; \mathcal{S}, \varrho) \leq N(\epsilon; \mathcal{S}, \varrho) \leq M(\epsilon; \mathcal{S}, \varrho)$. These two quantities exhibit the same scaling behav-
694 ior as $\epsilon \rightarrow 0$. Additionally, the logarithm of the covering number $\log N(\epsilon; \mathcal{S}, \varrho)$ is known as the
695 metric entropy of \mathcal{S} with respect to ϱ .

Definition B.7 (entropy number). Consider a quasi-Banach space consisting a compact set \mathcal{S} and a
quasi-metric ϱ . $N(\epsilon; \mathcal{S}, \varrho)$ denotes the covering number with radius ϵ . For $k = 1, 2, \dots$ the dyadic
entropy number is defined as

$$\epsilon_k(\mathcal{S}, \varrho) := \inf\{\epsilon > 0 : N(\epsilon; \mathcal{S}, \varrho) \leq 2^{k-1}\}.$$

Lemma B.8 (Entropy number of Schatten- q ball [13]). Consider a $d \times d$ -dimensional vector space.
Suppose \mathcal{S} is a S_q unit-ball and ϱ is the metric induced by F -norm. Then, we have the following
theorem for $q \leq 2$:

$$\epsilon_k(\mathbb{B}_{S_q}^d(1), \|\cdot\|_F) \asymp_q \begin{cases} 1 & \text{for } 1 \leq k \leq d \\ \left(\frac{d}{k}\right)^{\frac{1}{q}-\frac{1}{2}} & \text{for } d \leq k \leq d^2 \\ 2^{-\frac{k}{d^2}} \cdot d^{\frac{1}{2}-\frac{1}{q}} & \text{for } k \geq d^2. \end{cases} \quad (15a)$$

$$(15b)$$

$$(15c)$$

C Theoretical Results for Understanding of Dual-Level Sparsity for GLM

We now present a unified minimax analysis of our dual-level sparse framework. We begin by establishing lower bounds on the Frobenius-norm loss for estimators operating over dual-level sparse parameter spaces, showing the fundamental difficulty imposed by having to both *identify nonzero frequency components* and *estimate each low-rank slice*. Then, we turn to upper bounds by analyzing constrained least squares estimators that match these lower limits. Finally, we extend our discussion to more general $\ell_p(S_q)$ -type parameter spaces, deriving analogous minimax rates and discussing how the interplay between ℓ_p -sparsity and Schatten- q low-rankness affects the overall error.

(I) Minimax lower bounds over ℓ_0 - S_0 or ℓ_0 - S_q -balls. We first characterizes the *worst-case* error any estimator must incur under two types of dual-level sparse constraints:

(a) ℓ_0 - S_0 case ($q = 0$), meaning we have at most s nonzero frequency components and each slice is rank at most r .

(b) ℓ_0 - S_q case with $q \in (0, 1]$, meaning we again allow at most s active frequencies, but each slice's rank structure is relaxed into a Schatten- q -ball of radius R .

Theorem C.1. Consider the linear observation model $\mathbf{Y}_i = \mathbf{L}^* + \mathbf{E}_i$ under dual-level sparse spectrum, with n i.i.d. observations and noise variance σ^2 . Then the following bounds hold:

(a) If $q = 0$ and $\mathbf{L}^* \in \mathbf{T}_{0,0}(s, r)$, any measurable estimator $\hat{\mathbf{L}}$ satisfies

$$\inf_{\hat{\mathbf{L}}} \sup_{\mathbf{L} \in \mathbf{T}_{0,0}(s, r)} \mathbb{P} \left(\|\hat{\mathbf{L}} - \mathbf{L}\|_F^2 \geq C_\ell \frac{\sigma^2}{n} \left[s \log\left(\frac{em}{s}\right) + s r d \right] \right) \geq \frac{1}{2},$$

implying

$$\mathfrak{M}(\mathbf{T}_{0,0}(s, r)) \geq \frac{1}{2} C_\ell \frac{\sigma^2}{n} \left[s \log\left(\frac{em}{s}\right) + s r d \right].$$

(b) If $q \in (0, 1]$ and $\mathbf{L}^* \in \mathbf{T}_{0,q}(s, R)$, for any estimator $\hat{\mathbf{L}}$,

$$\inf_{\hat{\mathbf{L}}} \sup_{\mathbf{L} \in \mathbf{T}_{0,q}(s, R)} \mathbb{P} \left(\|\hat{\mathbf{L}} - \mathbf{L}\|_F^2 \geq C_\ell \left\{ \frac{\sigma^2}{n} s \log\left(\frac{em}{s}\right) + s R \left(\frac{\sigma^2}{n} d\right)^{1-\frac{q}{2}} \right\} \right) \geq \frac{3}{8},$$

which likewise implies

$$\mathfrak{M}(\mathbf{T}_{0,q}(s, R)) \geq \frac{3}{8} C_\ell \left\{ \frac{\sigma^2}{n} s \log\left(\frac{em}{s}\right) + s R \left(\frac{\sigma^2}{n} d\right)^{1-\frac{q}{2}} \right\}.$$

From the expressions in Theorem C.1, we see that the *estimation complexity* has two components:

(a) A term of order $s \log(\frac{em}{s})$ captures the combinatorial cost of identifying which s out of m frequency slices are nonzero.

(b) A second term quantifies the difficulty of *estimating each slice*, either in the rank- r case ($s r d$) or in the Schatten- q case $\left[s R \left(\frac{\sigma^2}{n} d\right)^{1-\frac{q}{2}} \right]$.

Hence, Theorem C.1 highlights a fundamental trade-off in learning dual-level sparse structures from noisy observations.

(II) Minimax upper bounds over ℓ_0 - S_0 or ℓ_0 - S_q -balls. We next confirm that these lower bounds are sharp by analyzing a *constrained least-squares* (CLS) estimator, defined for $q \in [0, 1]$ via

$$\hat{\mathbf{L}}_q \in \arg \min_{\mathbf{L} \in \mathbf{T}_{0,q}(s, R)} \|\mathbf{L} - \bar{\mathbf{Y}}\|_F^2, \quad (16)$$

where $\bar{\mathbf{Y}} = \frac{1}{n} \sum_{i=1}^n \mathbf{Y}_i$. One shows that $\hat{\mathbf{L}}_q$ attains an ε -accurate solution with high probability as soon as ε^2 is on the same order as the lower-bound terms in Theorem C.1.

Theorem C.2. Under the same dual-level sparse setups and i.i.d. noise model, the following hold:

(a) If $q = 0$ and we form the estimator $\hat{\mathbf{L}}_0$ by minimizing $\|\mathbf{L} - \bar{\mathbf{Y}}\|_F^2$ over $\mathbf{T}_{0,0}(s, r)$, then for any $\varepsilon^2 \geq C_u \frac{\sigma^2}{n} \left[s \log\left(\frac{em}{s}\right) + s r d \right]$,

$$\sup_{\mathbf{L} \in \mathbf{T}_{0,0}(s, r)} \|\hat{\mathbf{L}}_0 - \mathbf{L}\|_F^2 \leq \varepsilon^2, \quad (17)$$

with probability at least $1 - C_1 \exp(-C_2 n \varepsilon^2)$.

731 (b) If $q \in (0, 1]$ and we form $\hat{\mathbf{L}}_q$ similarly over $\mathbf{T}_{0,q}(s, R)$, then for any $\epsilon^2 \geq C_u \left\{ \frac{\sigma^2}{n} s \log\left(\frac{em}{s}\right) + \right.$
 732 $\left. s R \left(\frac{\sigma^2}{n} d\right)^{1-\frac{q}{2}} \right\}$,

$$\sup_{\mathbf{L} \in \mathbf{T}_{0,q}(s, R)} \|\hat{\mathbf{L}}_q - \mathbf{L}\|_F^2 \leq \epsilon^2, \quad (18)$$

733 holds with probability at least $1 - C_1 \exp(-C_2 n \epsilon^2)$.

734 These upper bounds match the lower bounds (Theorem C.1) up to constant factors, establishing the
 735 ℓ_0 - S_0 or ℓ_0 - S_q rates as *optimal*. Consequently, we arrive at the minimax rates:

$$\mathfrak{M}(\mathbf{T}_{0,0}(s, r)) \asymp \frac{\sigma^2}{n} \left[s \log\left(\frac{em}{s}\right) + s r d \right] \quad \text{and} \quad \mathfrak{M}(\mathbf{T}_{0,q}(s, R)) \asymp \frac{\sigma^2}{n} \left[s \log\left(\frac{em}{s}\right) + s R \left(\frac{\sigma^2}{n} d\right)^{1-\frac{q}{2}} \right].$$

736 Thus, Theorem C.2 shows that the above CLS estimators are *rate-optimal*.

737 **Remark C.3.** As a corollary, the joint results of Theorems C.1 and C.2 show that the minimax rates
 738 match up to constant factors in both the ℓ_0 - S_0 and ℓ_0 - S_q scenarios. Specifically, for $q = 0$:

$$\mathfrak{M}(\mathbf{T}_{0,0}(s, r)) \asymp \frac{\sigma^2}{n} \left[s \log\left(\frac{em}{s}\right) + s r d \right],$$

739 and for $q \in (0, 1]$:

$$\mathfrak{M}(\mathbf{T}_{0,q}(s, R)) \asymp \frac{\sigma^2}{n} \left[s \log\left(\frac{em}{s}\right) + s R \left(\frac{\sigma^2}{n} d\right)^{1-\frac{q}{2}} \right].$$

740 **(III) Minimax rates over general $\ell_p(S_q)$ -balls.** Finally, we move beyond the hard-sparsity con-
 741 straints to the more flexible $\ell_p(S_q)$ spaces:

$$\mathbf{T}_{p,q}(R) = \left\{ \mathbf{L} : \|\mathbf{L}\|_{\ell_p(S_q)}^p \leq R \right\}.$$

742 By combining similar lower/upper bound arguments (now requiring more subtle entropy and covering
 743 results) we arrive at Theorem C.4: To avoid over-complicated scenarios, we assume

$$\begin{cases} \log m < R \left(\frac{n}{\sigma^2 d}\right)^{\frac{p}{2}} \leq \frac{m}{\log m} \\ \log m < R \left(\frac{n}{\sigma^2 \log m}\right)^{\frac{p}{2}} \leq \frac{m}{\log m} \\ c_1 d < R^{\frac{q}{\sigma^2 p}} \left(\frac{n}{d}\right)^{\frac{q}{2}} < C_1 d. \end{cases} \quad (19)$$

744 **Theorem C.4** (Minimax Rates for $\ell_p(S_q)$ -balls). Suppose $p, q \in (0, 1]$ and condition (19) holds to
 745 avoid degenerate parameter ranges. Then the minimax risk over the $\ell_p(S_q)$ -ball is:

$$\begin{aligned} \mathfrak{M}(\mathbf{T}_{p,q}(R)) &= \inf_{\mathbf{L}} \sup_{\mathbf{L}^* \in \mathbf{T}_{p,q}(R)} \mathbb{E} \left[\|\hat{\mathbf{L}} - \mathbf{L}^*\|_F^2 \right] \\ &\asymp \begin{cases} R \left(\frac{\sigma^2 n}{d}\right)^{\frac{p-2}{2}} + R \left(\frac{\sigma^2 n}{\log m}\right)^{\frac{p-2}{2}}, & p > q, \\ R^{\frac{q}{p}} \left(\frac{\sigma^2 n}{d}\right)^{\frac{q-2}{2}} + R \left(\frac{\sigma^2 n}{\log m}\right)^{\frac{p-2}{2}}, & p \leq q, m > d^2, \\ R^{\frac{q}{p}} \left(\frac{\sigma^2 n}{d}\right)^{\frac{q-2}{2}}, & p \leq q, m \leq d^2. \end{cases} \end{aligned} \quad (20)$$

746 Examining (20) reveals three distinct regimes:

- 747 • $p > q$. The ℓ_p -type group sparsity dominates, rendering the Schatten- q penalties secondary.
- 748 • $p \leq q, m > d^2$. Both frequency-domain ℓ_p -grouping and intra-slice Schatten- q -norm jointly
 749 control the risk, leading to a sum of two terms.
- 750 • $p \leq q, m \leq d^2$. The S_q -ball effectively saturates the error, making the risk independent of m .

751 Hence, $\ell_p(S_q)$ quasi-norms admit richer, *soft* sparsity structures beyond the hard ℓ_0 - S_0 or ℓ_0 - S_q
 752 constraints, but yield analogous minimax phenomena once careful covering-entropy or packing
 753 arguments are applied.

754 In conclusion, these results unify the lower and upper bounds for dual-level sparse tensor estimation
 755 under both hard-sparsity (ℓ_0 - S_0 or ℓ_0 - S_q) and soft-sparsity ($\ell_p(S_q)$) constraints. The key takeaway
 756 is that the *minimax error* always balances identifying relevant frequencies with estimating each
 757 slice's rank structure. Hard- and soft-sparsity assumptions shift how these two aspects interact, but

the big-picture story remains consistent: multi-frequency, low-rank modeling carries a fundamental combinatorial cost (for frequency selection) plus a continuous cost (for matrix parameter estimation). These findings rigorously justify the dual-level sparsity approach and characterize the fundamental limits of any estimator hoping to learn such structured tensors from noisy observations.

Remark C.5 (On the Practical Validation and Use of Theorem 4.2). While Theorem 4.2 establishes minimax lower bounds for dual-sparse tensor recovery under the ℓ_p -Schatten- q prior, directly validating these rates through simulation is inherently difficult. This is because minimax guarantees are defined over worst-case estimators, which are often inaccessible—especially under nonconvex regularization. Even for vector-valued estimators with $\ell_p(\ell_q)$ sparsity, constructing such minimax-optimal procedures remains an open challenge when (p, q) are general.

C.1 Sketch of Minimax Proofs under Dual-Level Sparsity

Our proofs of the minimax bounds proceed by considering three successively more general forms of dual-level sparsity, each requiring distinct technical arguments due to their underlying structural assumptions.

(1) Hard–Hard Dual-Level Sparsity $\mathbf{T}_{0,0}(s, r)$. When both the number of active frequency components and the rank of each active slice are bounded by s and r , respectively, we construct a *packing set* that simultaneously encodes frequency sparsity and low-rankness:

1. *Support selection:* First, choose which s frequency slices (out of m) can be nonzero, ensuring a combinatorial factor $\binom{m}{s}$.
2. *Within-slice low-rank matrices:* Next, for each chosen frequency slice, define a family of low rank $\leq r$ matrices whose entries are set to a suitable scale δ , ensuring the separation properties

$$\|\underline{\mathbf{L}}^i - \underline{\mathbf{L}}^j\|_F^2 \geq c_0 s r d \delta^2 \quad \text{and} \quad \|\underline{\mathbf{L}}^i - \underline{\mathbf{L}}^j\|_F^2 \leq 2 s r d \delta^2,$$

while also achieving large enough cardinality for the packing set.

Applying Fano’s inequality to this well-separated set establishes a lower bound by appropriately choosing $\delta \propto \sqrt{\frac{\sigma^2}{n} [s \log(\frac{em}{s}) + s r d]}$. For the matching upper bound, we employ a constrained least squares estimator and use covering number arguments (based on matrix rank $\leq r$ covers and frequency-support combinatorics) to show it achieves the same rate.

(2) Hard–Soft Dual-Level Sparsity $\mathbf{T}_{0,q}(s, R)$. When the frequency sparsity is still hard-constrained by s , but each slice now belongs to a *soft low-rank* Schatten- q ball of radius R , the proof becomes more intricate:

- We decompose into (i) tensors with *different* frequency supports, incurring a cost of $\frac{\sigma^2}{n} s \log(\frac{em}{s})$, and (ii) tensors with the *same* support but *different* within-slice Schatten- q structures, contributing $s R (\frac{\sigma^2}{n} d)^{1-\frac{q}{2}}$.
- By carefully combining these two cases (through union bounds and entropy number estimates for Schatten- q sets), we derive matching lower and upper bounds, again establishing minimax optimality.

(3) Soft–Soft Dual-Level Sparsity $\mathbf{T}_{p,q}(R)$. Finally, in the most general setting, both frequency sparsity and low-rankness are relaxed into an $\ell_p(S_q)$ quasi-norm. Here, the analysis must carefully track how p and q govern *group-level* versus *within-group* regularity:

- When $p > q$, the ℓ_p penalty dominates, so the rates follow primarily from frequency-sparsity arguments.
- If $p \leq q$, we observe a phase transition depending on whether m exceeds d^2 . For $m > d^2$, the two penalties are both active, summing their respective complexities; for $m \leq d^2$, the Schatten- q term saturates the error, rendering it independent of m .

Technically, this requires sophisticated entropy tools, in particular Schütt’s theorem on entropy numbers for vector-valued sequence spaces, and a careful *chaining* analysis that tracks interactions across frequencies and singular values in each slice.

Unified Upper Bound Approach. Although the detailed proofs vary, the estimation upper bounds all follow a similar template via empirical process theory:

- For $\mathbf{T}_{0,0}(s, r)$, we rely on discrete packing/covering of finite-support rank $\leq r$ matrices.

- For $\mathbf{T}_{0,q}(s, R)$, we combine discrete frequency selection with known entropy results for Schatten- q balls.
- For $\mathbf{T}_{p,q}(R)$, we perform a chaining argument on $\ell_p(S_q)$ quasi-norms, applying Schütt-type entropy estimates.

In each case, bounding the least squares estimator's risk at the matching lower-bound rate confirms it is optimal up to constants.

Across these three regimes of *hard-hard*, *hard-soft*, and *soft-soft* dual-level sparsity, the main technical challenge is to *precisely quantify the geometric complexity* imposed by the interplay of frequency sparsity and low-rankness (whether hard or soft). Once we derive suitable packing/covering or entropy bounds, standard empirical process arguments transform that complexity into matching lower and upper rates. Consequently, the minimax results show a consistent story: learning multi-frequency structures involves a combinatorial cost from selecting active frequencies plus a continuous cost from estimating each low-rank (or Schatten- q) slice, culminating in the optimal rates detailed in our main theorems.

C.2 Proof of Theorem C.1

C.2.1 Proof of case (a)

Proof. For case (a), consider the $\frac{sr d}{32}$ -packing set $\tilde{\mathbf{T}}_{0,0}(s, r) = \{\underline{\mathbf{L}}^1, \dots, \underline{\mathbf{L}}^N\}$ constructed in Lemma C.9, where N is its cardinality. We set all nonzero entries of each $\underline{\mathbf{L}}^i \in \tilde{\mathbf{T}}_{0,0}(s, r)$ to be 1, and define $\vartheta^i = \underline{\mathbf{L}}^i \delta$, with δ a parameter to be determined. Since each $M(\underline{\mathbf{L}}^i)$ has at most $s r d$ nonzero elements, for any $\vartheta^i \neq \vartheta^j$, it follows that

$$\|\vartheta^i - \vartheta^j\|_{\mathbb{F}}^2 = \|M(\vartheta^i) - M(\vartheta^j)\|_{\mathbb{F}}^2 = \delta^2 \|M(\underline{\mathbf{L}}^i) - M(\underline{\mathbf{L}}^j)\|_{\mathbb{F}}^2 \leq 2 s r d \delta^2, \quad \forall i, j \in [N]. \quad (21)$$

On the other hand, from the construction of $\tilde{\mathbf{T}}_{0,0}(s, r)$, we also have

$$\|\vartheta^i - \vartheta^j\|_{\mathbb{F}}^2 \geq \frac{1}{32} s r d \delta^2, \quad \forall i, j \in [N]. \quad (22)$$

Using standard mutual information arguments [54] gives

$$\begin{aligned} I(\bar{\mathbf{Y}}; \psi) &\leq \frac{1}{\binom{N}{2}} \sum_{i \neq j} \text{KL}(\vartheta^i \parallel \vartheta^j) \\ &= \frac{1}{\binom{N}{2}} \sum_{i \neq j} \frac{n}{2 \sigma^2} \|\vartheta^i - \vartheta^j\|_{\mathbb{F}}^2 \\ &\leq \frac{n}{\sigma^2} s r d \delta^2, \end{aligned} \quad (23)$$

where $\text{KL}(\cdot \parallel \cdot)$ is the Kullback–Leibler divergence, and the last step uses (21). By applying Fano's inequality [43], we obtain

$$\mathbb{P}(\hat{\psi} \neq \psi) \geq 1 - \frac{\frac{n}{\sigma^2} s r d \delta^2 + \log 2}{\log N},$$

where ψ is uniformly distributed over the packing set $\tilde{\mathbf{T}}_{0,0}(s, r)$. To ensure $\mathbb{P}(\hat{\psi} \neq \psi) \geq \frac{1}{2}$, it suffices to choose

$$\delta = \frac{1}{2} \sqrt{\left[(c_1 r s d) + (c_2 s \log(\frac{em}{s})) \right] \frac{\sigma^2}{s r d n}}.$$

Substituting this choice into (22) and invoking Lemma C.9 completes the proof, showing that

$$\inf_{\underline{\mathbf{L}}} \sup_{\underline{\mathbf{L}} \in \mathbf{T}_{0,0}(s, r)} \mathbb{P} \left[\|\hat{\underline{\mathbf{L}}} - \underline{\mathbf{L}}\|_{\mathbb{F}}^2 \geq \frac{c \sigma^2}{n} (r s d + s \log \frac{em}{s}) \right] \geq \frac{1}{2}.$$

Finally, a Markov's inequality argument yields the same lower bound in expectation. \square

C.2.2 Theoretical Tools: Gilbert-Varshamov Theorems

The proof of case (a) requires a $\frac{sr d}{32}$ -packing set $\tilde{\mathbf{T}}_{0,0}(s, r)$. Before proceeding with the construction of this packing set, we first introduce several versions of the Gilbert-Varshamov theorem that will be crucial for our analysis. These results provide guarantees on the existence of well-separated binary and K -ary codes.

The first version deals with binary codes containing the zero vector:

842 **Lemma C.6** (Gilbert-Varshamov Theorem for Binary Codes [46]). *Consider a length- m code with*
 843 *binary symbols (2-ary coding). There exists a subset $\Omega = \{\omega_0, \dots, \omega_N\}$ of the code book where*
 844 *$\omega_i \in \{0, 1\}^m$ that satisfies:*

- 845 • *The zero vector is included: $\omega_0 = (0, \dots, 0)$.*
- 846 • *The minimum Hamming distance is bounded: $d_H(\omega_i, \omega_j) \geq m/8$ for all $0 \leq i < j \leq N$.*
- 847 • *The set size is exponential: $N \geq 2^{m/8}$.*

848 This lemma guarantees the existence of a large set of well-separated binary vectors, which will be
 849 used to construct the low-rank matrices $M(\mathbf{L})_{:, :, k}$ within the k -th ($k \in [m]$) frequency component of
 850 $\mathbf{L} \in \mathbb{R}^{d_1 \times d_2 \times m}$ in the transformed domain defined by M in (1).

851 Then, for more general alphabets beyond $\{0, 1\}$, we have:

852 **Lemma C.7** (Gilbert-Varshamov Theorem for K -ary Codes). *For a length- m code with K -ary*
 853 *symbols, there exists a ϱ -separated set whose cardinality is at least:*

$$N_K(m, \varrho) \geq \frac{K^m}{\sum_{i=0}^{\varrho-1} \binom{m}{i} (K-1)^i}$$

854 where ϱ denotes the minimum Hamming distance between any two codewords.

855 Further, for codes restricted to a Hamming sphere, we have:

856 **Lemma C.8** (Gilbert-Varshamov Theorem for Bounded-Weight Codes, [54]). *Consider the Hamming*
 857 *sphere of radius s for a length- m code with K -ary symbols. There exists a ϱ -separated set within*
 858 *this sphere with cardinality at least:*

$$N_K(m, s, \varrho) \geq \frac{\binom{m}{s} (K-1)^s}{\sum_{i=0}^{\varrho-1} \binom{m}{i} (K-1)^i} \quad (24)$$

859 A particularly useful special case occurs when $\varrho = c_1 s$ and we consider binary coding ($K = 2$):

$$N_2(m, s, \varrho) \geq (em/s)^{c_2 s} \quad (25)$$

860 where c_1 and c_2 are absolute constants.

861 **Lemma C.9** (Existence of a Packing Set). *There exists a packing set $\tilde{\mathbf{T}}_{0,0}(s, r) \subset \mathbf{T}_{0,0}(s, r)$*
 862 *satisfying the following properties:*

- 863 1. **Frequency sparsity constraint:** *Each tensor in $\tilde{\mathbf{T}}_{0,0}(s, r)$ has at most s nonzero frequency*
 864 *components.*
- 865 2. **Low-rank structure:** *Each frequency component is represented by a matrix of rank at most r .*
- 866 3. **Separation property:** *Any two distinct tensors $\mathbf{L}^1, \mathbf{L}^2 \in \tilde{\mathbf{T}}_{0,0}(s, r)$ satisfy:*

$$\|\mathbf{L}^1 - \mathbf{L}^2\|_F^2 \geq \frac{sr d_1}{32}. \quad (26)$$

867 4. **Cardinality:** *The packing set has size at least:*

$$|\tilde{\mathbf{T}}_{0,0}(s, r)| \geq (em/s)^{cs} \cdot 2^{rs d_1/32} \geq \exp(c_1 r s d_1 + c_2 s \log(em/s)). \quad (27)$$

868 *Proof.* We now construct our packing set $\tilde{\mathbf{T}}_{0,0}(s, r) \subset \mathbf{T}_{0,0}(s, r)$ through a sequence of carefully
 869 designed steps. This packing set must be sufficiently rich to capture the essential complexity of
 870 $\mathbf{T}_{0,0}(s, r)$. The construction must simultaneously enforce *frequency sparsity bounded by s* , maintain
 871 the *low-rank structure within each frequency component characterized by r* , and ensure *good*
 872 *separation properties* across all constructed sub-sets at each step.

873 **Step 1: Selection of Non-zero Frequencies** First, we determine a subset $\tilde{\Gamma} \subset \Gamma$ of the support of the
 874 m frequency components. This step establishes the *frequency sparsity pattern* of our tensors \mathbf{L} in the
 875 transformed domain defined by $M(\cdot)$. We proceed as follows:

- 876 1) *Ensure frequency sparsity:* First, we use a 2-ary code of length m on the Hamming sphere of radius
 877 s to represent possible support patterns.
- 878 2) *Ensure sufficient separation:* Second, by requiring the code to be $s/4$ -separated, we can guarantee
 879 the existence of a set $\tilde{\Gamma}$ with a minimum number of frequency patterns by Lemma C.8:

$$N_2(m, s, s/4) \geq (em/s)^{cs} \geq \exp(cs \log(em/s)) \quad (28)$$

880 where c is an absolute constant.

881 **Step 2: Construction of Low-rank Matrices** We then consider constructing appropriate low-rank
 882 matrices. Without loss of generality, assume $d_1 \geq d_2$. Motivated by [21], we first construct the set of
 883 matrices $\mathbf{A}_{\text{low-rank}}$ as follows:

884 For positions (i, j) of a matrix $\mathbf{A} \in \mathbf{A}_{\text{low-rank}}$ with $i \leq r$ and $j \in [d_2]$, we set $\mathbf{A}_{i,j} \in \{0, 1\}$, and for
 885 all other positions (i.e., $i > r$), we set $\mathbf{A}_{i,j} = 0$.

886 Then, we can ensure that $\text{rank}(\mathbf{A}) \leq r, \forall \mathbf{A} \in \mathbf{A}_{\text{low-rank}}$. Further by Lemma C.6, this construction
 887 yields a $\{0, 1\}$ -code of length rd_1 , for which we can find a subset $\{\mathbf{0}, \mathbf{A}^1, \dots, \mathbf{A}^{N_0}\}$ satisfying:

888 i) *Sufficiently many low-rank patterns*: $N_0 \geq 2^{rd_1/8}$.

889 ii) *Sufficient separation*: The Hamming distance $d_H(\mathbf{A}^i, \mathbf{A}^j) \geq rd_1/8$ for all $0 \leq i < j \leq N_0$.

890 In sumamry, by construction, we can find the matrix set $\tilde{\mathbf{A}}_{\text{low-rank}} := \{\mathbf{A}^1, \dots, \mathbf{A}^{N_0}\}$ that satisfy:

$$|\tilde{\mathbf{A}}_{\text{low-rank}}| \geq 2^{rd_1/8}, \quad \text{rank}(\mathbf{A}^i) \leq r, \quad \|\mathbf{A}^i\|_F^2 \leq rd_1, \quad \|\mathbf{A}^i - \mathbf{A}^j\|_F^2 \geq rd_1/8, \quad \text{and} \quad \|\mathbf{A}^i\|_F^2 \geq rd_1/8. \quad (29)$$

891 **Step 3: Assign Low-rank Patterns for a Fixed Frequency Pattern** Now have founded the s -sparsity
 892 frequency support patterns in $\tilde{\Gamma}$ and the r -low-rank frequency matrices in $\tilde{\mathbf{A}}_{\text{low-rank}}$. We then need to
 893 assign appropriate low-rank patterns to each selected frequency $\gamma \in \tilde{\Gamma}$. This step requires careful
 894 consideration of both the separation properties and the cardinality of our construction.

895 Motivated by the proof of Lemma 3 in [25], we consider the following analysis:

896 1) From **Step 2**, we know that each frequency component can take $|\tilde{\mathbf{A}}_{\text{low-rank}}| \geq 2^{rd_1/8}$ different
 897 low-rank patterns. This gives us a large alphabet size for coding each frequency component.

898 2) We can view this as a $|\tilde{\mathbf{A}}_{\text{low-rank}}|$ -ary coding problem for each selected frequency, where we also
 899 need to ensure the resulting codes are $s/2$ -separated in Hamming distance.

900 3) For a fixed frequency sparsity pattern $\gamma \in \tilde{\Gamma}$, we can lower bound the **cardinality of the resulting**
 901 **set $\tilde{\mathbf{T}}_\gamma$** according to Lemma C.7 as follows:

$$\begin{aligned} N_{|\tilde{\mathbf{A}}_{\text{low-rank}}|}(s, s/2) &\geq \frac{(2^{rd_1/8})^s}{\sum_{i=0}^{s/2-1} \binom{s}{i} (2^{rd_1/8} - 1)^i} \\ &\geq \frac{(2^{rd_1/8})^s}{\sum_{i=0}^{s/2-1} \binom{s}{i} (2^{rd_1/8})^{s/2-1}} \\ &\geq \frac{(2^{rd_1/8})^s}{2^s \cdot (2^{rd_1/8})^{s/2-1}} \\ &= \frac{(2^{rd_1/8})^{s/2+1}}{2^s} \\ &= 2^{rsd_1/16-s} \geq 2^{rsd_1/32} \end{aligned} \quad (30)$$

902 The inequalities above are derived through the following steps. The first inequality follows from the
 903 application of the Gilbert-Varshamov theorem for K -ary codes. The second inequality is obtained by
 904 upper bounding $(2^{rd_1/8})^i - 1$ with $(2^{rd_1/8})^{s/2-1}$. The third inequality results from bounding the
 905 sum of binomial coefficients by 2^s . Finally, the last inequality holds under the assumption that rd_1 is
 906 sufficiently large, specifically $rd_1 \geq 32$.

907 **Step 4: Integration of Frequency Sparsity and Low-rankness** Now, for each $\gamma \in \tilde{\Gamma}$, we have
 908 found a set of tensors $\tilde{\mathbf{T}}_\gamma$ specified by γ . Then, we totally found a set $\tilde{\mathbf{T}}$ of at least $|\tilde{\Gamma}| \cdot 2^{rsd_1/32}$
 909 tensors. Next, we will show that $\tilde{\mathbf{T}}$ the ideal packing set $\tilde{\mathbf{T}}_{0,0}(s, r)$ of $\mathbf{T}_{0,0}(s, r)$ we are looking for.
 910 We consider two cases:

Case 4.1: Different Frequency Sparsity Patterns When any two tensors $\mathbf{L}^1, \mathbf{L}^2 \in \tilde{\mathbf{T}}$ in our construction
 have different frequency supports $\gamma^1, \gamma^2 \in \tilde{\Gamma}$, then according to the construction of $\tilde{\Gamma}$, the frequency
 supports γ^1 and γ^2 are $s/2$ -separated. That means the tensors $\mathbf{L}^1, \mathbf{L}^2$ have at least $s/4$ different
 frequency positions, thus the distance

$$\|\mathbf{L}^1 - \mathbf{L}^2\|_F^2 = \|M(\mathbf{L}^1) - M(\mathbf{L}^2)\|_F^2 \geq \frac{s}{4} \cdot \frac{rd_1}{8} = \frac{sr d_1}{32}$$

911 *Case 4.2: Same Frequency Sparsity Pattern* When any two tensors $\mathbf{L}^1, \mathbf{L}^2 \in \tilde{\mathbf{T}}$ in our construction
 have the same frequency support $\gamma \in \tilde{\Gamma}$, then according to the construction of $\tilde{\mathbf{T}}_\gamma$, there are at least

$s/2$ different low-rank patterns of $\underline{\mathbf{L}}^1, \underline{\mathbf{L}}^2$. That means the tensors $\underline{\mathbf{L}}^1, \underline{\mathbf{L}}^2$ have at least $s/2$ frequency positions that are $rd_1/8$ separated, thus the distance

$$\|\underline{\mathbf{L}}^1 - \underline{\mathbf{L}}^2\|_F^2 = \frac{s}{2} \cdot \frac{rd_1}{8} = \frac{sr d_1}{16}$$

Combining both cases, we can conclude that the cardinality of our constructed set is at least:

$$|\tilde{\mathbf{T}}| = \prod_{\gamma \in \tilde{\mathbf{F}}} |\tilde{\mathbf{T}}_\gamma| \geq (em/s)^{cs} \cdot 2^{rsd_1/32} \geq \exp(c_1 rsd_1 + c_2 s \log(em/s)), \quad (31)$$

which completes the proof. \square

C.2.3 Proof of case (b)

Proof. Here, we assume the tensor $\underline{\mathbf{L}}$ has *hard* frequency sparsity (at most s nonzero frequency components) but a *soft* low-rank constraint on each active slice, enforced by a Schatten- q ball of radius R . Proving a minimax lower bound under this mixed constraint is more subtle than in case (a). Specifically, we split the argument into two complementary subcases:

- *Subcase 1:* The *location* of the s nonzero frequency slices varies among tensors, but *within each active slice* we use the same low-rank matrix (scaled by some δ). This isolates the combinatorial complexity of deciding which frequency slices are nonzero.
- *Subcase 2:* The *set of active frequency slices is fixed*, but *the matrix in each slice* can differ (still subject to the Schatten- q ball). This reveals the complexity arising from “soft” rank variations within each frequency component.

By analyzing each subcase and then applying a union bound, we show that any estimator must fail on at least one scenario with nontrivial probability, yielding the desired lower bound of order

$$\frac{\sigma^2}{n} \left[s \log\left(\frac{em}{s}\right) + s R \left(\frac{\sigma^2}{n} d\right)^{1-\frac{q}{2}} \right].$$

Step 1: Constructing a suitable matrix family \mathbf{A} . Before detailing subcases, we define a set of rank- $\leq r$ matrices \mathbf{A} that, when properly scaled by δ , remain inside the Schatten- q ball of radius R . Concretely,

$$\mathbf{A} := \{ \mathbf{A} : \mathbf{A} \in \delta \times \{0, 1\}^{d \times d}, \|\mathbf{A}\|_0 \leq r d, \text{rank}(\mathbf{A}) \leq r \},$$

where δ satisfies:

$$\sum_{i=1}^r \sigma_i^2(\mathbf{A}) \leq \delta^2 r d \quad \text{and} \quad \sum_{i=1}^r \sigma_i^q(\mathbf{A}) \leq R.$$

From a simple calculation, letting

$$\delta = (R/r)^{\frac{1}{q}} d^{-\frac{1}{2}}$$

ensures each $\mathbf{A} \in \mathbf{A}$ actually lies in the Schatten- q ball of radius R . This “reference family” \mathbf{A} will be used in subcase 1 to examine which frequency slices are activated.

Subcase 1: Different frequency supports but the same slice matrix. In this subcase, we focus on “which s slices are nonzero?” while fixing the same rank- $\leq r$ matrix across those slices.

- (i) *Fix a baseline matrix $\mathbf{A} \in \mathbf{A}$.* Since $\|\mathbf{A}\|_{S_q} \leq R$ under our choice of δ , we can replicate \mathbf{A} across active slices without violating the soft rank constraint.
- (ii) *Choose different frequency supports.* By a combinatorial argument similar to Lemma C.9, we know there exists a family $\tilde{\mathbf{F}}$ of cardinality at least

$$c_1 s \log\left(\frac{em}{s}\right)$$

that enumerates possible subsets $\gamma \subseteq [m]$ of size s . Define

$$\tilde{\mathbf{T}}_{0,q}(s, R, \mathbf{A}) := \{ \gamma \otimes \mathbf{A} : \gamma \in \tilde{\mathbf{F}} \}.$$

Each tensor in this set places the same \mathbf{A} (scaled by δ) in different positions γ , capturing up to s activated slices.

943 (iii) *Packing separation.* If two different supports γ^i and γ^j differ in at least $s/4$ coordinates,
 944 the corresponding tensors $\underline{\mathbf{L}}^i, \underline{\mathbf{L}}^j$ differ by at least $c_0 s r d \delta^2$ in Frobenius norm (for some
 945 constant $c_0 > 0$), while they differ by at most $2 s r d \delta^2$ if the supports overlap substantially.
 946 Thus, $\tilde{\mathbf{T}}_{0,q}(s, R, \mathbf{A})$ forms a $(c_0 s r d \delta^2)$ -packing set.
 947 (iv) *Mutual information & Fano.* As in case (a), using $\text{KL}(\underline{\mathbf{L}}^i \| \underline{\mathbf{L}}^j) \approx \frac{n}{2\sigma^2} \|\underline{\mathbf{L}}^i - \underline{\mathbf{L}}^j\|_F^2$ and ap-
 948 plying Fano's inequality, we show no estimator can distinguish all elements of $\tilde{\mathbf{T}}_{0,q}(s, R, \mathbf{A})$
 949 with probability above $1/2$ if $\delta \sim \sqrt{\frac{\sigma^2}{n r d} \log(\frac{em}{s})}$. This step isolates the combinatorial cost
 950 $s \log(\frac{em}{s})$ in the final bound.

951 **Subcase 2: A fixed frequency set but different slice entries.** Now we fix which s slices are
 952 nonzero (i.e., fix some $\gamma \in \tilde{\Gamma}$), but allow *different* matrices in each of these slices. Since the slices are
 953 constrained by $\|\mathbf{B}_i\|_{S_q} \leq R$, we consider a large packing of s -tuples $(\mathbf{B}_1, \dots, \mathbf{B}_s)$, each \mathbf{B}_i rank- $\leq r$,
 954 lying in the Schatten- q ball. Formally,

$$\tilde{\mathbf{T}}_{0,q}(s, R, \gamma) := \{\underline{\mathbf{L}} : M(\underline{\mathbf{L}})_{:,i} = \mathbf{B}_i \text{ for } i \in [s], M(\underline{\mathbf{L}})_{:,i} = \mathbf{0} \text{ otherwise; } \mathbf{B} \in (\mathbb{B}_{S_q}^d(R))^s\}.$$

955 Repeating the packing-based argument (analogous to steps 2–4 in Lemma C.9), we build a family
 956 in which every pair $\underline{\mathbf{L}}^i, \underline{\mathbf{L}}^j$ differs by at least $\frac{1}{16} s r d \delta^2$ in $\|\cdot\|_F$. An information-theoretic (Fano)
 957 calculation then shows no estimator can identify all such slice-tuples simultaneously with high
 958 probability (e.g. above $7/8$). This yields a second component in the lower bound tied to $s R (\frac{\sigma^2}{n} d)^{1-\frac{q}{2}}$,
 959 reflecting the *intra-slice* degrees of freedom from the Schatten- q ball.

960 **Combining the two subcases via a union bound.** An intuitive way to see the final bound is to
 961 note that any estimator $\hat{\underline{\mathbf{L}}}$ must work *both* when the s nonzero slices vary in location (subcase 1) *and*
 962 when they are fixed but each slice can vary in a Schatten- q manner (subcase 2). By a union bound,
 963 the probability of success in both subcases is at most the sum of success probabilities, ensuring that
 964 *some* scenario fails with positive probability. Thus we obtain a probability statement of the form

$$\inf_{\hat{\underline{\mathbf{L}}}} \sup_{\underline{\mathbf{L}} \in \tilde{\mathbf{T}}_{0,q}(s, R)} P[\|\hat{\underline{\mathbf{L}}} - \underline{\mathbf{L}}\|_F^2 \geq c (s \log(\frac{em}{s}) + s R (\frac{\sigma^2}{n} d)^{1-\frac{q}{2}})] > 0$$

965 for some constant $c > 0$. A final Markov or Chebyshev step then converts this probability bound to
 966 an expectation form, concluding the lower bound in the minimax sense.

967 **Detailed union-bound argument.** More explicitly, one shows that subcase 1 alone forces a risk at
 968 least

$$\frac{\sigma^2}{n} [s \log(\frac{em}{s})], \quad \text{and subcase 2 alone forces a risk of at least } s R (\frac{\sigma^2}{n} d)^{1-\frac{q}{2}}.$$

969 So the event of “ $\hat{\underline{\mathbf{L}}}$ having small error on *both* subcases” is bounded by the sum of their separate
 970 probabilities (i.e. a union bound), leading to the conclusion that *any* estimator must fail on at least
 971 one scenario with probability at least a fixed positive constant. From this, we deduce

$$\inf_{\hat{\underline{\mathbf{L}}}} \sup_{\underline{\mathbf{L}} \in \tilde{\mathbf{T}}_{0,q}(s, R)} \mathbb{E}[\|\hat{\underline{\mathbf{L}}} - \underline{\mathbf{L}}\|_F^2] \gtrsim \frac{\sigma^2}{n} [s \log(\frac{em}{s}) + s R (\frac{\sigma^2}{n} d)^{1-\frac{q}{2}}],$$

972 which completes the proof for case (b). □

973 C.3 Proofs of Theorem C.2

974 Before formally providing the proof of upper bounds, we provide some useful technical lemmas.

975 C.3.1 Technical Lemmas for upper bounds

976 We first prove the upper bounds for the covering number of the parameter spaces.

977 **Lemma C.10** (Upper bounds for the covering number). *Denote $N(\mathbf{T}, \|\cdot\|_F, \varepsilon)$ as the ε -covering*
 978 *number of parameter set \mathbf{T} .*

979 (a) For $q = 0$ and $\varepsilon \in (0, 1]$, let $\mathbb{S}^{md^2-1} := \{\underline{\mathbf{L}} \in \mathbb{R}^{d \times d \times m} : \|\underline{\mathbf{L}}\|_F = 1\}$, then we have

$$\log N(\varepsilon; \mathbf{T}_{0,0}(s, r) \cap \mathbb{S}^{md^2-1}, \|\cdot\|_F) \leq s \log \frac{em}{s} + 2srd \log \frac{1}{\varepsilon/\sqrt{s}}.$$

980 (b) For $q \in (0, 1]$ and for all $\varepsilon \in [c_q \sqrt{s} R^{\frac{1}{q}} d^{\frac{1}{2} - \frac{1}{q}}, c_q \sqrt{s} R^{\frac{1}{q}}]$,

$$\log N(\varepsilon; \mathbf{T}_{0,q}(s, R), \|\cdot\|_F) \lesssim_q s \log \frac{em}{s} + s(C_q \frac{s R^{\frac{2}{q}}}{\varepsilon^2})^{\frac{q}{2-q}} d.$$

981 *Proof. Case (a):* Any tensor $\mathbf{L} \in \mathbf{T}_{0,0}(s, r)$ has at most s nonzero frontal slices in the transformed
982 domain $M(\cdot)$. Denote the nonzero slice indices by

$$\Gamma(\mathbf{L}) = \{i \in [m] : M(\mathbf{L})_{:,i} \neq \mathbf{0}\}.$$

983 Since $|\Gamma(\mathbf{L})| \leq s$, the total number of ways to pick these supports is bounded by $\binom{m}{s}$. Therefore,

$$\log \binom{m}{s} \leq s \log \left(\frac{em}{s} \right).$$

984 This accounts for selecting *which* frequency slices are potentially nonzero.

985 For each fixed support $\gamma \subseteq [m]$ with $|\gamma| \leq s$, we need to cover the set of matrices $\{\mathbf{A} \in \mathbb{R}^{d \times d} :$
986 $\|\mathbf{A}\|_F \leq 1, \text{rank}(\mathbf{A}) \leq r\}$ in the $\|\cdot\|_F$ metric by balls of radius ε/\sqrt{s} . It is a standard fact
987 (see, e.g., [38]) that the δ -covering number for rank- $\leq r$ matrices of Frobenius norm at most 1 is
988 upper-bounded by

$$N\left(\delta; \{\mathbf{A} : \|\mathbf{A}\|_F \leq 1, \text{rank}(\mathbf{A}) \leq r\}, \|\cdot\|_F\right) \leq \exp\left(C r d \log\left(\frac{1}{\delta}\right)\right),$$

989 for some constant $C > 0$. Substituting $\delta = \varepsilon/\sqrt{s} \leq 1$ gives a covering number of order

$$\exp\left(C r d \log\left(\frac{\sqrt{s}}{\varepsilon}\right)\right).$$

990 Taking logarithms yields

$$\log N\left(\frac{\varepsilon}{\sqrt{s}}; \{\mathbf{A} : \|\mathbf{A}\|_F \leq 1, \text{rank}(\mathbf{A}) \leq r\}, \|\cdot\|_F\right) \lesssim r d \log\left(\frac{\sqrt{s}}{\varepsilon}\right).$$

991 Given s nonzero frequency slices, each can be approximated by a matrix in the covering set with
992 radius ε/\sqrt{s} . Since there are at most s active slices, the total squared error in Frobenius norm sums
993 up to at most $s \cdot (\varepsilon/\sqrt{s})^2 = \varepsilon^2$. Hence,

$$(\text{Total covering number}) \leq \binom{m}{s} \times \left[\exp\left(C r d \log\left(\frac{\sqrt{s}}{\varepsilon}\right)\right) \right]^s.$$

994 Taking logarithms and using $\log \binom{m}{s} \leq s \log \left(\frac{em}{s} \right)$ completes the argument:

$$\log N\left(\varepsilon; \mathbf{T}_{0,0}(s, r) \cap \mathbb{S}^{md^2-1}, \|\cdot\|_F\right) \leq s \log\left(\frac{em}{s}\right) + s [r d \log\left(\frac{\sqrt{s}}{\varepsilon}\right)] = s \log\left(\frac{em}{s}\right) + 2 s r d \log\left(\frac{1}{\varepsilon/\sqrt{s}}\right),$$

995 possibly absorbing constants into the notation. This completes the proof of Case (a).

996 **Case (b):** We now analyze the covering number of $\mathbf{T}_{0,q}(s, R)$ by leveraging the entropy properties
997 of S_q -balls.

998 To construct an ε -covering of $\mathbf{T}_{0,q}(s, R)$, we first define a covering set $\tilde{\mathbb{B}}_{S_q}^d(R)$ of $\mathbb{B}_{S_q}^d(R)$ with
999 respect to the Frobenius norm, where each covering element approximates a matrix within $\mathbb{B}_{S_q}^d(R)$
1000 with an error at most $\frac{\varepsilon}{\sqrt{s}}$.

1001 For any $\mathbf{L} \in \mathbf{T}_{0,q}(s, R)$, its j -th frequency component satisfies $M(\mathbf{L})_{:,j} \in \mathbb{B}_{S_q}^d(R)$. By the definition
1002 of a covering set, there exists a matrix $\mathbf{A}^j \in \tilde{\mathbb{B}}_{S_q}^d(R)$ such that: $\|\mathbf{A}^j - M(\mathbf{L})_{:,j}\|_F^2 \leq \frac{\varepsilon^2}{s}$.

1003 This guarantees that any element in $\mathbf{T}_{0,q}(s, R)$ can be approximated using a combination of s selected
1004 frequency components from m total frequencies, each of which is covered by $\tilde{\mathbb{B}}_{S_q}^d(R)$. Using this,
1005 the covering number of $\mathbf{T}_{0,q}(s, R)$ satisfies:

$$N(\varepsilon; \mathbf{T}_{0,q}(s, R), \|\cdot\|_F) \leq \binom{m}{s} \left(N\left(\frac{\varepsilon}{\sqrt{s}}; \tilde{\mathbb{B}}_{S_q}^d(R), \|\cdot\|_F\right) \right)^s.$$

1006 Here, the first term accounts for the number of ways to select s active frequency components, while
1007 the second term reflects the covering number for each selected frequency component.

1008 Next, we employ entropy estimates for S_q -balls. From (15b), the entropy number of $\mathbb{B}_{S_q}^d(1)$ satisfies:

$$\epsilon_k(\mathbb{B}_{S_q}^d(1)) = \frac{\varepsilon}{\sqrt{s}} \lesssim_q \left(\frac{d}{k}\right)^{\frac{1}{q}-\frac{1}{2}}.$$

1009 Inverting this relation to express k in terms of ε , we set:

$$k = \log N\left(\frac{\varepsilon}{\sqrt{s}}; \mathbb{B}_{S_q}^d(1), \|\cdot\|_F\right),$$

1010 which, after allowing for a ball radius $R^{1/q}$, leads to:

$$\log N\left(\frac{\varepsilon}{\sqrt{s}}; \mathbb{B}_{S_q}^d(R), \|\cdot\|_F\right) \asymp \left(C_q \frac{s R^{\frac{2}{q}}}{\varepsilon^2}\right)^{\frac{q}{2-q}} d.$$

1011 The constraint on ε ensures that $k \in [d, d^2]$, keeping the bounds valid.

1012 Combining these results, we obtain:

$$\begin{aligned} \log N(\varepsilon; \mathbf{T}_{0,q}(s, R), \|\cdot\|_F) &\lesssim_q s \log \frac{em}{s} + s \log N\left(\frac{\varepsilon}{\sqrt{s}}; \mathbb{B}_{S_q}^d(R), \|\cdot\|_F\right) \\ &\lesssim_q s \log \frac{em}{s} + s \left(C_q \frac{s R^{\frac{2}{q}}}{\varepsilon^2}\right)^{\frac{q}{2-q}} d. \end{aligned}$$

1013 This bound quantifies how the covering number of $\mathbf{T}_{0,q}(s, R)$ depends on the sparsity level s , the
1014 dimensionality d , and the spectral decay parameter q . \square

1015 **Refined Statement and Explanation.** We begin by defining a function $f(\underline{\mathbf{T}}; \mathcal{X})$ for a tensor
1016 $\underline{\mathbf{T}} \in \mathbb{R}^{d \times d \times m}$ and some data structure (or random tensor) \mathcal{X} . We consider a constrained supremum

$$\sup_{\varrho(\underline{\mathbf{T}}) \leq \nu, \underline{\mathbf{T}} \in \mathbf{T}} f(\underline{\mathbf{T}}; \mathcal{X}),$$

1017 where $\varrho: \mathbb{R}^{d \times d \times m} \rightarrow \mathbb{R}^+$ is an *increasing* constraint function and \mathbf{T} is any nonempty collection of
1018 tensors. Let $\nu > 0$ be a fixed threshold, and define the event

$$\mathcal{E} := \{\mathcal{X} : \exists \underline{\mathbf{T}} \in \mathbf{T} \text{ such that } f(\underline{\mathbf{T}}; \mathcal{X}) \geq 2g(\varrho(\underline{\mathbf{T}}))\},$$

1019 where $g: \mathbb{R} \rightarrow \mathbb{R}^+$ is strictly increasing. Our aim is to bound $\mathbb{P}(\mathcal{E})$, i.e. the probability that there is
1020 some tensor $\underline{\mathbf{T}}$ with constraint $\varrho(\underline{\mathbf{T}}) \leq \nu$ for which $f(\underline{\mathbf{T}}; \mathcal{X})$ exceeds $2g(\varrho(\underline{\mathbf{T}}))$. This setup is quite
1021 general: for example, f might be a residual or cost function in an empirical process framework, $\varrho(\underline{\mathbf{T}})$
1022 might measure the size or norm of $\underline{\mathbf{T}}$, and g could be a nondecreasing penalty or bound we wish to
1023 enforce.

1024 **Lemma C.11 (Peeling Bound).** Lemma 9 of [37], reproduced below, gives a powerful “peeling”-type
1025 argument. It says that if we can control

$$\mathbb{P}\left[\sup_{\underline{\mathbf{T}}: \varrho(\underline{\mathbf{T}}) \leq \nu} f(\underline{\mathbf{T}}; \mathcal{X}) \geq g(\nu)\right] \leq 2 \exp(-c a_n g(\nu))$$

1026 for some constants $c > 0$ and $a_n > 0$, then one can derive a stronger tail bound for $\mathbb{P}(\mathcal{E})$. Formally:

1027 **Lemma C.11 (Peeling, Lemma 9 of [37]).** *Suppose that for all $\nu \geq 0$, we have $g(\nu) \geq \mu$. Then*
1028 *there exists a constant $c > 0$ such that for all $\nu > 0$,*

$$\mathbb{P}\left[\sup_{\underline{\mathbf{T}} \in \mathbf{T}, \varrho(\underline{\mathbf{T}}) \leq \nu} f(\underline{\mathbf{T}}; \mathcal{X}) \geq g(\nu)\right] \leq 2 \exp(-c a_n g(\nu)).$$

1029 Hence,

$$\mathbb{P}(\mathcal{E}) = \mathbb{P}(\exists \underline{\mathbf{T}} \in \mathbf{T} : f(\underline{\mathbf{T}}; \mathcal{X}) \geq 2g(\varrho(\underline{\mathbf{T}}))) \leq \frac{2 \exp(-4 c a_n \mu)}{1 - \exp(-4 c a_n \mu)}.$$

1030 **Why “peeling” is useful.** This lemma effectively “peels” off the largest values of $\varrho(\mathbf{T})$ in layers
 1031 and bounds the supremum in each layer by $g(\nu)$. One then aggregates or unions over these layers to
 1032 control the probability that $f(\mathbf{T}; \mathcal{X})$ can exceed $2g(\varrho(\mathbf{T}))$ for *any* \mathbf{T} . Such arguments often appear in
 1033 minimax or empirical process proofs, where one partitions the parameter space according to $\varrho(\mathbf{T})$.

1034 **Additional Lemmas for Dual-Level Sparse Spectra.** In our dual-level sparse setting, we require
 1035 more specialized versions of standard covering or chaining arguments. For instance, Lemma C.12
 1036 below extends Lemma 6 of [37] to handle an ℓ_0 -type frequency-sparsity set with $\max(s)$ active slices
 1037 and $\max(r)$ rank constraints:

$$\mathbf{T}_{0,0}(s, r)(2s, 2r) = \{\mathbf{L} : \#\{\text{nonzero freq. slices}\} \leq 2s, \text{rank}(M(\mathbf{L})_{:, :, i}) \leq 2r\}.$$

1038 We also define

$$\tilde{\mathbf{S}}(\mathbf{T}_{0,0}(s, r)(2s, 2r), \rho) = [\mathbf{T}_{0,0}(s, r)(2s, 2r)] \cap \{\mathbf{L} : \|\mathbf{L}\|_F \leq \rho\}.$$

1039 **Lemma C.12.** *There exist positive constants $C_1, C_2 > 0$ such that for any $\rho > 0$,*

$$\sup_{\mathbf{L} \in \tilde{\mathbf{S}}(\mathbf{T}_{0,0}(s, r)(2s, 2r), \rho)} |\langle \bar{\mathbf{E}}, \mathbf{L} \rangle| \leq C_u \sigma \rho \sqrt{\frac{1}{n} \left[s \log\left(\frac{em}{s}\right) + s r d \right]},$$

1040 *with probability at least $1 - C_1 \exp(-C_2 [s \log(\frac{em}{s}) + s r d])$, where $\bar{\mathbf{E}}$ is the (scaled) noise tensor.*

1041 *Sketch of proof.* This follows from Lemma 6 of [37] if we replace the covering number for a naive
 1042 ℓ_0 -ball by the more specific covering number of $\mathbf{T}_{0,0}(s, r)(2s, 2r)$. The detailed estimate of that
 1043 covering number is provided by part (a) of Lemma C.10. Essentially, one shows that among all
 1044 frequency-sparse and rank-limited tensors of Frobenius norm up to ρ , the uniform covering can
 1045 be done with cardinality roughly $\exp[s \log(\frac{em}{s}) + s r d \log(\frac{1}{\epsilon})]$, and then translates that into a tail
 1046 bound on $\sup_{\mathbf{L}} \langle \bar{\mathbf{E}}, \mathbf{L} \rangle$.

1047 **A chaining argument (Lemma C.13) for the hard–soft setting.** When frequency sparsity is still
 1048 “hard” but the rank constraint is “soft” via a Schatten- q ball, we adapt the standard chaining approach
 1049 to $\mathbf{T}_{0,q}(s, R)(2s, 2R)$, i.e. the set of tensors with *up to* $2s$ active frequency slices and *each* slice
 1050 having Schatten- q norm up to $2R$. Let

$$\tilde{\mathbf{S}}(\mathbf{T}_{0,q}(s, R)(2s, 2R), \rho) = \mathbf{T}_{0,q}(s, R)(2s, 2R) \cap \{\mathbf{L} : \|\mathbf{L}\|_F \leq \rho\}.$$

1051 We would like to show a tail bound of the form

$$\sup_{\mathbf{L} \in \tilde{\mathbf{S}}(\mathbf{T}_{0,q}(s, R)(2s, 2R), \rho)} |\langle \bar{\mathbf{E}}, \mathbf{L} \rangle| \leq \left(\sqrt{\frac{s \log(\frac{em}{s})}{n}} + \sqrt{s R} \left(\frac{d}{n}\right)^{\frac{1}{2} - \frac{q}{4}} \right) \rho$$

1052 with high probability. The chaining lemma from [12] is invoked, requiring a delicate construction of
 1053 small δ and an integral bound on the covering number $N(t; \tilde{\mathbf{S}}(\mathbf{T}_{0,q}(s, R)(2s, 2R), \rho), \|\cdot\|_F)$.

1054 **Lemma C.13 (Chaining bound).** *Assume*

$$\log\left(\frac{em}{s}\right) \leq C_2 n R^{\frac{2}{q}}. \quad (32)$$

1055 *Then there exist constants $C_3, C_4 > 0$ such that for any*

$$\rho \geq c \left(\sqrt{\frac{s \log(\frac{em}{s})}{n}} + \sqrt{s R} \left(\frac{d}{n}\right)^{\frac{1}{2} - \frac{q}{4}} \right),$$

1056 *we have*

$$\sup_{\mathbf{L} \in \tilde{\mathbf{S}}(\mathbf{T}_{0,q}(s, R)(2s, 2R), \rho)} |\langle \bar{\mathbf{E}}, \mathbf{L} \rangle| \leq \left(\sqrt{\frac{s \log(\frac{em}{s})}{n}} + \sqrt{s R} \left(\frac{d}{n}\right)^{\frac{1}{2} - \frac{q}{4}} \right) \rho$$

1057 *with probability at least*

$$1 - C_3 \exp\left\{-C_4 n \left(\frac{s \log(\frac{em}{s})}{n} + s R \left(\frac{d}{n}\right)^{1 - \frac{q}{2}}\right)\right\}.$$

1058 *Proof.* Following the approach of Lemma 7 in [37], we aim to construct a constant δ that satisfies the
 1059 conditions

$$\sqrt{n}\delta \geq C_1\rho, \quad (33)$$

1060 and

$$C_2\sqrt{n}\delta \geq \int_{\frac{\delta}{16}}^{\rho} \sqrt{\log N\left(t; \tilde{\mathbf{S}}(\mathbf{T}_{0,q}(2s, 2R), \rho), \|\cdot\|_F\right)} dt := J(\rho, \delta), \quad (34)$$

1061 where $N\left(t; \tilde{\mathbf{S}}(\mathbf{T}_{0,q}(2s, 2R), \rho), \|\cdot\|_F\right)$ denotes the t -covering number of $\tilde{\mathbf{S}}(\mathbf{T}_{0,q}(2s, 2R), \rho)$.

1062 Applying Lemma 3.2 in [12], we obtain that for $\|\bar{\mathbf{E}}\|_F^2 \leq 16$,

$$\mathbb{P}\left[\sup_{\mathbf{L} \in \tilde{\mathbf{S}}(\mathbf{T}_{0,q}(2s, 2R), \rho)} |\langle \bar{\mathbf{E}}, \mathbf{L} \rangle| \geq \delta, \quad \|\bar{\mathbf{E}}\|_F^2 \leq 16\right] \leq C_3 \exp(-C_4 \frac{n\delta^2}{\rho^2}).$$

1063 Since each entry of $\bar{\mathbf{E}}$ is drawn from $N(0, \frac{\sigma^2}{n})$, applying standard tail bounds for χ^2 random variables
 1064 [37] yields

$$\mathbb{P}[\|\bar{\mathbf{E}}\|_F^2 \geq 16] \leq C_5 \exp(-C_6 n).$$

1065 Consequently, we obtain the bound

$$\mathbb{P}\left[\sup_{\mathbf{L} \in \tilde{\mathbf{S}}(\mathbf{T}_{0,q}(2s, 2R), \rho)} |\langle \bar{\mathbf{E}}, \mathbf{L} \rangle| \geq \delta\right] \leq C_3 \exp(-C_4 \frac{n\delta^2}{\rho^2}) + C_5 \exp(-C_6 n).$$

1066 Next, we construct δ to satisfy conditions (33) and (34). Define

$$\delta = \rho \left(\sqrt{\frac{s \log \frac{em}{s}}{n}} + \omega \right),$$

1067 where $\omega > 0$ is a constant to be determined later. This choice immediately satisfies (33). For (34),
 1068 using condition (32), we set

$$\rho = \Omega \left(\sqrt{\frac{s \log \frac{em}{s}}{n}} + \sqrt{sR} \left(\frac{d}{n} \right)^{\frac{1}{2} - \frac{q}{4}} \right) \wedge \sqrt{sR}^{\frac{1}{q}}.$$

1069 It follows that $(\frac{\delta}{16}, \rho)$ lies within the valid range of ε in Lemma C.10. By applying part (b) of Lemma
 1070 C.10, we obtain

$$\begin{aligned} J(\rho, \delta) &= \int_{\frac{\delta}{16}}^{\rho} \sqrt{\log N\left(t; \tilde{\mathbf{S}}(\mathbf{T}_{0,q}(2s, 2R), \rho)\right)} dt \\ &\leq \int_0^{\rho} \sqrt{2s \log \frac{em}{s} + 2s \left(\frac{s}{t^2} R^{\frac{2}{q}} \right)^{\frac{q}{2-q}}} d t \\ &\leq \sqrt{2s \log \frac{em}{s}} \rho + \sqrt{2} (sR)^{\frac{1}{2-q}} \sqrt{d} \rho^{1 - \frac{q}{2-q}}. \end{aligned}$$

1071 Dividing both sides by $\sqrt{n}\delta$ gives

$$\frac{J(\rho, \delta)}{\sqrt{n}\delta} \leq \frac{\sqrt{2s \log \frac{em}{s}} \rho + \sqrt{2} (sR)^{\frac{1}{2-q}} \sqrt{d} \rho^{1 - \frac{q}{2-q}}}{\rho \sqrt{s \log \frac{em}{s}} + \rho \sqrt{n}\omega}.$$

1072 Setting $\omega = \sqrt{2} (sR)^{\frac{1}{2-q}} \sqrt{\frac{d}{n}} \rho^{1 - \frac{q}{2-q}}$ ensures that

$$\frac{J(\rho, \delta)}{\sqrt{n}\delta} \leq \sqrt{2}.$$

1073 Thus, condition (34) holds.

1074 Finally, we conclude that

$$\delta = \rho \left(\sqrt{\frac{s \log \frac{em}{s}}{n}} + \sqrt{sR} \left(\frac{d}{n} \right)^{\frac{1}{2} - \frac{q}{4}} \right).$$

1075 Substituting into our probability bound, we obtain

$$\begin{aligned} \mathbb{P} \left[\sup_{\mathbf{L} \in \tilde{\mathbf{T}}(\mathbf{T}_{0,q}(2s, 2R), \rho)} |\langle \bar{\mathbf{E}}, \mathbf{L} \rangle| \geq \left(\sqrt{\frac{s \log \frac{em}{s}}{n}} + \sqrt{sR} \left(\frac{d}{n} \right)^{\frac{1}{2} - \frac{q}{4}} \right) \rho \right] \\ \leq C_3 \exp \left(-C_4 n \left(\frac{s \log \frac{em}{s}}{n} + sR \left(\frac{d}{n} \right)^{1 - \frac{q}{2}} \right) \right), \end{aligned}$$

1076 which completes the proof of Lemma C.13. \square

1077 C.3.2 Proof of C.2

1078 We analyze the constrained MLE estimator in (16). Since for any $q \in [0, 1]$, the estimator satisfies

$$\|\bar{\mathbf{Y}} - \hat{\mathbf{L}}_q\|_F^2 \leq \|\bar{\mathbf{Y}} - \mathbf{L}^*\|_F^2,$$

1079 rearranging terms gives

$$\|\hat{\mathbf{L}}_q - \mathbf{L}^*\|_F^2 \leq 2|\langle \bar{\mathbf{E}}, \hat{\mathbf{L}}_q - \mathbf{L}^* \rangle|. \quad (35)$$

1080 Proof of Case (a)

1081 *Proof.* For case (a), since both $\hat{\mathbf{L}}_0$ and \mathbf{L}^* belong to $\mathbf{T}_{0,0}(s, r)$, their difference satisfies

$$\hat{\mathbf{L}}_0 - \mathbf{L}^* \in \mathbf{T}_{0,0}(2s, 2r).$$

1082 Applying Lemma C.12, for any $\rho > 0$, we obtain

$$\sup_{\mathbf{L} \in \tilde{\mathbf{T}}(\mathbf{T}_{0,0}(2s, 2r), \rho)} |\langle \bar{\mathbf{E}}, \mathbf{L} \rangle| \leq C_u \sigma \rho \sqrt{\frac{1}{n} (s \log \frac{em}{s} + srd)}$$

1083 with probability at least $1 - C_1 \exp\{-C_2(s \log \frac{em}{s} + srd)\}$.

1084 Next, consider the event \mathcal{E} where there exists some $\mathbf{L} \in \mathbf{T}_{0,0}(2s, 2r)$ such that

$$|\langle \bar{\mathbf{E}}, \mathbf{L} \rangle| \geq C_u \sigma \|\mathbf{L}\|_F \sqrt{\frac{1}{n} (s \log \frac{em}{s} + srd)}. \quad (36)$$

1085 By Lemma C.11, the probability of this event satisfies

$$\mathbb{P}[\mathcal{E}] \leq \frac{2 \exp(-C_3(s \log \frac{em}{s} + srd))}{1 - \exp(-C_3(s \log \frac{em}{s} + srd))}.$$

1086 This follows by applying Lemma C.11 with function $f(\mathbf{T}; \mathcal{X}) = \langle \bar{\mathbf{E}}, \mathbf{L} \rangle$, set $\mathbf{T} =$
1087 $\mathbf{T}_{0,0}(2s, 2r)$, sequence $a_n = n/\sigma^2$, function $\varrho(\mathbf{T}) = \|\mathbf{T}\|_F$, and threshold function $g(\nu) =$

1088 $C_u \sigma \nu \sqrt{\frac{1}{n} (s \log \frac{em}{s} + srd)}$. For any $\nu \geq \sigma \sqrt{\frac{1}{n} (s \log \frac{em}{s} + srd)}$, we ensure that $g(\nu) \geq$

1089 $\frac{\sigma^2}{n} (s \log \frac{em}{s} + srd)$, allowing us to apply the lemma.

1090 Combining (35) and (36) yields

$$\|\hat{\mathbf{L}}_0 - \mathbf{L}^*\|_F^2 \leq C_u \frac{\sigma^2}{n} (s \log \frac{em}{s} + srd).$$

1091 This bound holds with probability at least $1 - C_1 \exp\{-C_2(s \log \frac{em}{s} + srd)\}$, completing the proof
1092 of (17). \square

1093 Proof of case(b)

1094 *Proof.* For case (b), since $\hat{\mathbf{L}}_q, \mathbf{L}^* \in \mathbf{T}_{0,q}(s, R)$, it follows that their difference satisfies $\hat{\mathbf{L}}_q - \mathbf{L}^* \in$
1095 $\mathbf{T}_{0,q}(2s, 2R)$.

1096 We define the event \mathcal{E} as the existence of some $\mathbf{L} \in \mathbf{T}_{0,q}(s, R)$ such that, by Lemma C.13, the
1097 following holds with probability at least $1 - C_3 \exp\{-C_4 n (\frac{s \log \frac{em}{s}}{n} + sR (\frac{d}{n})^{1 - \frac{q}{2}})\}$:

$$|\langle \bar{\mathbf{E}}, \mathbf{L} \rangle| \geq C_u \|\mathbf{L}\|_F \left(\sqrt{\frac{s \log \frac{em}{s}}{n}} + \sqrt{sR} \left(\frac{d}{n} \right)^{\frac{1}{2} - \frac{q}{4}} \right).$$

1098 Applying Lemma C.11, we further obtain the probability bound:

$$\mathbb{P}[\mathcal{E}] \leq \frac{2 \exp(-C_3 n (\frac{s \log \frac{em}{s}}{n} + s R (\frac{d}{n})^{1-\frac{q}{2}}))}{1 - \exp(-C_3 n (\frac{s \log \frac{em}{s}}{n} + s R (\frac{d}{n})^{1-\frac{q}{2}}))}. \quad (37)$$

1099 This follows from Lemma C.11 by setting:

$$f(\mathbf{T}; \mathcal{X}) = \langle \bar{\mathbf{E}}, \mathbf{L} \rangle, \quad \mathbf{T} = \mathbf{T}_{0,q}(2s, 2R), \quad a_n = n, \quad \varrho(\mathbf{T}) = \|\mathbf{T}\|_F, \quad g(\nu) = \nu \left(\sqrt{\frac{s \log \frac{em}{s}}{n}} + \sqrt{s R} (\frac{d}{n})^{\frac{1}{2}-\frac{q}{4}} \right).$$

1100 Combining (35) with these results, we conclude that:

$$\|\hat{\mathbf{L}} - \mathbf{L}^*\|_F^2 \leq C_u \left(\sqrt{\frac{s \log \frac{em}{s}}{n}} + \sqrt{s R} (\frac{d}{n})^{\frac{1}{2}-\frac{q}{4}} \right),$$

1101 which holds with probability at least

$$1 - C_3 \exp\{-C_4 n (\frac{s \log \frac{em}{s}}{n} + s R (\frac{\sigma^2(1-v)}{4n} d)^{1-\frac{q}{2}})\},$$

1102 thus completing the proof of (18). \square

1103 C.4 Proof of Theorem C.4

1104 C.4.1 Lower bound for $\ell_p(S_q)$

1105 *Proof.* We aim to prove a minimax lower bound under the dual-level sparse structure imposed by an
1106 $\ell_p(S_q)$ quasi-norm. Our proof strategy considers two key parameters that govern dual-level sparsity:

- 1107 • *Frequency sparsity* $1 \leq s \leq m$, controlling how many frequency components can be nonzero.
- 1108 • *Within-frequency low-rankness* $1 \leq r \leq d$, limiting the rank of each active frequency component.

1109 *Subspace construction and parameter setting.* We begin with the subspace $\mathbf{T}_{0,0}(s, r)$ of tensors, as
1110 introduced in earlier sections, where both frequency indices and within-frequency ranks are hard-
1111 constrained. For any tensor in this subspace, suppose the absolute value of each nonzero entry is set
1112 to $\delta > 0$. This δ is chosen so that the (p, q) -quasi-norm constraint is satisfied, i.e.,

$$s \cdot \left(r \cdot (\delta \sqrt{d})^q \right)^{\frac{p}{q}} = R.$$

1113 Solving for δ yields

$$\delta = R^{\frac{1}{p}} s^{-\frac{1}{p}} r^{-\frac{1}{q}} d^{-\frac{1}{2}}.$$

1114 Thus, δ encodes how large each nonzero entry must be so that the tensor simultaneously satisfies
1115 frequency-sparsity and low-rankness in a consistent manner for the $\ell_p(S_q)$ -norm. Notice that δ scales
1116 inversely with s , r , and \sqrt{d} , reflecting the interplay among frequency selection, rank constraints, and
1117 the Frobenius norm.

1118 *Applying generalized Fano's inequality.* Let us express the probability that any estimator $\hat{\mathbf{L}}$ incurs
1119 significant estimation error. Using a generalized Fano argument, we obtain

$$\inf_{\hat{\mathbf{L}}} \sup_{\mathbf{L}^* \in \mathbf{T}_{p,q}(R)} \mathbb{P} \left[\|\hat{\mathbf{L}} - \mathbf{L}^*\|_F^2 \geq \frac{1}{16} R^{\frac{2}{p}} s^{1-\frac{2}{p}} r^{1-\frac{2}{q}} \right] \geq 1 - \frac{\frac{n}{\sigma^2} R^{\frac{2}{p}} s^{1-\frac{2}{p}} r^{1-\frac{2}{q}} + \log 2}{\frac{1}{4} (s r d + s \log(\frac{m}{s}))}. \quad (38)$$

1120 The denominator $s r d + s \log(\frac{m}{s})$ captures (1) the cost of identifying s nonzero frequency slices
1121 each of rank at most r , plus (2) the combinatorial complexity $s \log(\frac{m}{s})$ for subset selection among
1122 m frequencies.

1123 *Reformulating via linear programming.* To analyze (38) precisely, we adopt a linear programming
1124 approach similar to [25] for the $\ell_u(\ell_q)$ -ball. Let $y = \log s$ and $x = \log r$. Then

$$R^{\frac{2}{p}} s^{1-\frac{2}{p}} r^{1-\frac{2}{q}} = \exp \left[\log R^{\frac{2}{p}} + \left(1 - \frac{2}{p}\right) \log s + \left(1 - \frac{2}{q}\right) \log r \right].$$

1125 Hence, maximizing $R^{\frac{2}{p}} s^{1-\frac{2}{p}} r^{1-\frac{2}{q}}$ is equivalent to maximizing

$$z = \left(1 - \frac{2}{p}\right) y + \left(1 - \frac{2}{q}\right) x,$$

1126 subject to constraints bounding x and y (i.e. $0 \leq x \leq \log d$ and $0 \leq y \leq \log m$), plus an additional
 1127 constraint from balancing numerator and denominator in (38). Concretely:

$$\begin{cases} 0 \leq x \leq \log d, & 0 \leq y \leq \log m, \\ y \geq \min\left\{-\frac{p}{q}x + \log R + \frac{p}{2}(\log n - \log(\sigma^2) - \log d), -\left(\frac{p}{q} - \frac{p}{2}\right)x + \log R + \frac{p}{2}(\log n - \log(\sigma^2) - \log(\log m))\right\}. \end{cases}$$

1128 The first two lines capture $s \leq m$, $r \leq d$, while the last line encodes how $\frac{n}{\sigma^2} R^{\frac{2}{p}} s^{1-\frac{2}{p}} r^{1-\frac{2}{q}}$
 1129 compares with $s r d + s \log(\frac{m}{s})$ to ensure a valid Fano-type bound.

1130 *Analyzing slopes and boundary points.* For convenience, define:

$$\begin{aligned} x_1 &= \log R + \frac{p}{2} \log\left(\frac{n}{\sigma^2 \log m}\right), & x_2 &= \log R + \frac{p}{2} \log\left(\frac{n}{\sigma^2 d}\right), \\ y_1 &= \frac{q}{p} \log R + \frac{q}{2} \log\left(\frac{n}{\sigma^2 \log m}\right), & y_2 &= \frac{q}{p} \log R + \frac{q}{2} \log\left(\frac{n}{\sigma^2 d}\right). \end{aligned}$$

1132 We then consider the lines:

- 1133 • *Line A:* $y = -\frac{p}{q}x + \log R + \frac{p}{2}(\log n - \log(\sigma^2) - \log d)$, with slope $-\frac{p}{q}$ in the (x, y) -plane.
- 1134 • *Line B:* $y = -\left(\frac{p}{q} - \frac{p}{2}\right)x + \log R + \frac{p}{2}(\log n - \log(\sigma^2) - \log(\log m))$, whose slope is $\frac{p}{2} - \frac{p}{q}$.
- 1135 • *Objective slope:* The slope of $z = (1 - \frac{2}{p})y + (1 - \frac{2}{q})x$ is $\frac{p(q-2)}{q(p-2)}$ in the (x, y) -plane.

1136 A standard slope comparison yields these observations:

- 1137 1. If $p > q$, the slope of z is larger than slopes of Lines A and B, so the maximum of z is
 1138 attained at boundary points like $(0, y_2)$, $(0, y_1)$, or $(x_1, 0)$.
- 1139 2. If $p \leq q$, the slope of z is smaller than the slope of A but larger than that of B, so maxima
 1140 can occur at $(x_1, 0)$, $(x_2, 0)$, or intersections of A and B, depending on m vs. d .

1141 Evaluating z at these boundary points, one then obtains the resulting minimax lower bounds:

$$\begin{cases} R\left(\frac{\sigma^2 n}{d}\right)^{\frac{p-2}{2}} + R\left(\frac{\sigma^2 n}{\log m}\right)^{\frac{p-2}{2}}, & \text{if } p > q, \\ R^{\frac{\sigma^2 q}{p}} \left(\frac{\sigma^2 n}{d}\right)^{\frac{q-2}{2}} + R\left(\frac{n}{\log m}\right)^{\frac{p-2}{2}}, & \text{if } p \leq q, m \geq d^2, \\ R^{\frac{q}{p}} \left(\frac{\sigma^2 n}{d}\right)^{\frac{q-2}{2}}, & \text{if } p \leq q, m \leq d^2. \end{cases}$$

1142 These match precisely the piecewise expressions for the lower bound in the $\ell_p(S_q)$ setting. Hence,
 1143 combining with the initial Fano-based argument (38) concludes the minimax lower bound proof. \square

1145 C.4.2 Covering number of $\mathbb{B}_{\ell_p(S_q)}(R)$

1146 Before deriving the upper bounds for $\ell_p(S_q)$, we first need to derive the covering number of
 1147 $\mathbb{B}_{\ell_p(S_q)}(R)$ equipped with the $\ell_p(S_q)$ -norm. To this end, we generalize Schütt's theorem for vector-
 1148 valued sequence spaces [9]. The analysis relies on entropy numbers and their relationships under
 1149 different parameter ranges. We introduce several key lemmas and derive the upper bound for e_k .

Lemma C.14 (Schütt's Theorem for Vector-valued Sequence Spaces [9]). *Let X and Y be r -normed quasi-Banach spaces, and let $0 < q < r \leq \infty$. The unit ball $\mathbb{B}_{\ell_q^m(X)}$ is defined as:*

$$\mathbb{B}_{\ell_q^m(X)} = v_1 \mathbb{B}_X \times v_2 \mathbb{B}_X \times \cdots \times v_m \mathbb{B}_X,$$

1150 where \mathbb{B}_X is the unit ball with X -norm, and $v \in \mathbb{B}_q$. For $k, k_0 \in \mathbb{N}$ such that $k_0 \leq k$, let:

$$\begin{aligned} D(k_0, k) &= \max_{l \in \mathbb{N}, k_0 \leq l \leq k} \left(\frac{l}{k}\right)^{\frac{1}{q} - \frac{1}{r}} e_l(\text{id} : X \rightarrow Y), \\ A(k, m) &= \max \left\{ \|\text{id} : X \rightarrow Y\| \left(\frac{\log(em/k)}{k}\right)^{\frac{1}{q} - \frac{1}{r}}, D(1, k) \right\}, \end{aligned}$$

1151 where $\|\text{id} : X \rightarrow Y\|$ denotes the operator norm, and $e_l(\text{id} : X \rightarrow Y)$ denotes the l -th entropy
 1152 number. For $k \geq \log_2(m)$, the entropy numbers satisfy:

- If $k \leq m$, then

$$e_k(id : \ell_q^m(X) \rightarrow \ell_r^m(Y)) \simeq A(k, m).$$

- If $k \geq m$, then there exist constants $C_1, C_2 > 0$ such that:

$$D(C_1 k/m, k) \leq e_k(id : \ell_q^m(X) \rightarrow \ell_r^m(Y)) \leq D(C_2 k/m, k).$$

Let $q = p$, $r = 2$, $X = S_q^d$, or $Y = \ell_2^d$ for our problem, so that $\|id : X \rightarrow Y\| = 1$ and $e_l(id : X \rightarrow Y)$ is given by (15a), (15b) and (15c). Using the results of this lemma, we define the function $\phi(l)$ to analyze the behavior of entropy numbers for $\ell_p(S_q)$ -balls. Specifically, we let:

$$\phi(l) := \left(\frac{l}{k}\right)^{\frac{1}{q} - \frac{1}{r}} e_l(id : X \rightarrow Y) = \begin{cases} \left(\frac{l}{k}\right)^{\frac{1}{p} - \frac{1}{2}} & 1 \leq l \leq d, \\ \left(\frac{l}{k}\right)^{\frac{1}{p} - \frac{1}{2}} \left(\frac{d}{l}\right)^{\frac{1}{q} - \frac{1}{2}} & d \leq l \leq d^2, \\ \left(\frac{l}{k}\right)^{\frac{1}{p} - \frac{1}{2}} 2^{-\frac{l}{d^2}} d^{\frac{1}{2} - \frac{1}{q}} & l \geq d^2. \end{cases}$$

1153 The monotonicity behavior of $\phi(l)$ across different ranges of l is summarized in Table 2.

Table 2: Monotonicity of $\phi(l)$ for $p \leq q$ and $p > q$ in different ranges of l when $p \leq 1$

Range of l	Expression for $\phi(l)$	Monotonicity (if $p \leq q$)	Monotonicity (if $p > q$)	Critical Point
$1 \leq l \leq d$	$\left(\frac{l}{k}\right)^{\frac{1}{p} - \frac{1}{2}}$	Increasing		None
$d \leq l \leq d^2$	$\left(\frac{l}{k}\right)^{\frac{1}{p} - \frac{1}{2}} \left(\frac{d}{l}\right)^{\frac{1}{q} - \frac{1}{2}}$	Increasing	Decreasing	None
$l \geq d^2$	$\left(\frac{l}{k}\right)^{\frac{1}{p} - \frac{1}{2}} 2^{-\frac{l}{d^2}} d^{\frac{1}{2} - \frac{1}{q}}$	Increasing then Decreasing with maximum at $l^* = \frac{(\frac{1}{p} - \frac{1}{2})d^2}{\ln 2}$		$l^* > d^2$
		Decreasing		$l^* \leq d^2$

1154 **Lemma C.15** (Entropy Number for $\ell_p(S_q) \hookrightarrow \ell_2(S_2)$). For $k \geq \max\{\log m, d\}$, the entropy
1155 numbers e_k for $\mathbb{B}_{\ell_p(S_q)}(R)$ satisfy:

$$e_k \simeq_{p,q} \begin{cases} \left(\frac{d}{k}\right)^{\frac{1}{q} - \frac{1}{2}} & p \leq q, m \leq d^2 \\ \max \left\{ \left(\frac{d}{k}\right)^{\frac{1}{q} - \frac{1}{2}}, \left(\frac{\log(em)}{k}\right)^{\frac{1}{p} - \frac{1}{2}} \right\} & p \leq q, m \geq d^2 \\ \left(\frac{\max\{d, \log(em)\}}{k}\right)^{\frac{1}{p} - \frac{1}{2}} & q \leq p. \end{cases} \quad (39)$$

1156 *Proof of Lemma C.15.* Let us prove this lemma by carefully analyzing different cases based on the
1157 relationships between p, q, m , and d . Our analysis will heavily rely on the behavior of $\phi(l)$ as shown
1158 in Table 2 and the application of Lemma C.14.

1159 **Case (a):** For $p \leq q, m \leq d^2$, we consider the range $d \leq k \leq d^2$ and divide our analysis into two
1160 subcases based on the relationship between k and m .

1161 (i) First, consider $k \leq m$: According to Lemma C.14, we have:

$$e_k \simeq A(k, m) = \max \left\{ \left(\frac{\log(em/k)}{k}\right)^{\frac{1}{p} - \frac{1}{2}}, D(1, k) \right\}.$$

1162 To evaluate $D(1, k)$, we need to analyze $\phi(l)$ in different ranges:

- For $1 \leq l \leq d$:

$$\phi(l) = \left(\frac{l}{k}\right)^{\frac{1}{p} - \frac{1}{2}}.$$

1164 This function is strictly increasing as $\frac{1}{p} - \frac{1}{2} > 0$ for $p \leq 1$. The maximum in this range occurs at
1165 $l = d$.

- For $d \leq l \leq d^2$:

$$\phi(l) = \left(\frac{l}{k}\right)^{\frac{1}{p} - \frac{1}{2}} \left(\frac{d}{l}\right)^{\frac{1}{q} - \frac{1}{2}}.$$

1167 Since $p \leq q$, we have $\frac{1}{p} - \frac{1}{q} \geq 0$, making this function increasing. The maximum in this range
1168 occurs at $l = k$ since $k \leq d^2$.

1169 Combining these results, we find:

$$D(1, k) = \max_{1 \leq l \leq k} \phi(l) = \left(\frac{d}{k}\right)^{\frac{1}{q} - \frac{1}{2}}.$$

1170 Now, since $m \leq d^2$, we can show:

$$\left(\frac{\log(em/k)}{k}\right)^{\frac{1}{p} - \frac{1}{2}} \lesssim \left(\frac{d}{k}\right)^{\frac{1}{q} - \frac{1}{2}}.$$

1171 Therefore:

$$e_k \simeq \left(\frac{d}{k}\right)^{\frac{1}{q} - \frac{1}{2}}.$$

1172 (ii) Next, consider $m \leq k \leq d^2$: By Lemma C.14, when $k \geq m$, we have:

$$D(C_1 k/m, k) \leq e_k \leq D(C_2 k/m, k).$$

1173 For each $C \in \{C_1, C_2\}$, we need to evaluate:

$$D(Ck/m, k) = \max_{Ck/m \leq l \leq k} \phi(l).$$

1174 Let's analyze $\phi(l)$ in the relevant ranges:

1175 • When $Ck/m \leq l \leq d$:

$$\phi(l) = \left(\frac{l}{k}\right)^{\frac{1}{p} - \frac{1}{2}}.$$

1176 This is increasing, reaching its maximum at $l = d$ if d is in this range.

1177 • When $d \leq l \leq k \leq d^2$:

$$\phi(l) = \left(\frac{l}{k}\right)^{\frac{1}{p} - \frac{1}{2}} \left(\frac{d}{l}\right)^{\frac{1}{q} - \frac{1}{2}}.$$

1178 Since $p \leq q$, this is increasing and reaches its maximum at $l = k$.

1179 The monotonicity of $\phi(l)$ implies that both bounds achieve their maximum at $l = k$:

$$D(Ck/m, k) = \left(\frac{k}{k}\right)^{\frac{1}{p} - \frac{1}{2}} \left(\frac{d}{k}\right)^{\frac{1}{q} - \frac{1}{2}} = \left(\frac{d}{k}\right)^{\frac{1}{q} - \frac{1}{2}}.$$

1180 Therefore:

$$D(C_1 k/m, k) = D(C_2 k/m, k) = \left(\frac{d}{k}\right)^{\frac{1}{q} - \frac{1}{2}}.$$

1181 This gives us:

$$e_k \simeq \left(\frac{d}{k}\right)^{\frac{1}{q} - \frac{1}{2}}.$$

1182 **Case (b):** For $p \leq q, m \geq d^2$, when $d \leq k \leq m$, Lemma C.14 gives:

$$e_k \simeq \max\left\{\left(\frac{\log(em/k)}{k}\right)^{\frac{1}{p} - \frac{1}{2}}, D(1, k)\right\}.$$

1183 We analyze this in two subcases:

1184 (i) For $d \leq k \leq d^2$: Similar to Case (a), analyzing $\phi(l)$ in three ranges:

1185 • $1 \leq l \leq d$: $\phi(l) = \left(\frac{l}{k}\right)^{\frac{1}{p} - \frac{1}{2}}$ is increasing;

1186 • $d \leq l \leq d^2$: $\phi(l) = \left(\frac{l}{k}\right)^{\frac{1}{p} - \frac{1}{2}} \left(\frac{d}{l}\right)^{\frac{1}{q} - \frac{1}{2}}$ is increasing since $p \leq q$;

1187 • $l \geq d^2$: $\phi(l)$ is decreasing from $l = d^2$.

1188 Therefore:

$$D(1, k) = \left(\frac{d}{k}\right)^{\frac{1}{q} - \frac{1}{2}}.$$

1189 (ii) For $d^2 \leq k \leq m$: In this range:

$$D(1, k) = \max_{1 \leq l \leq k} \phi(l) = \left(\frac{d^2}{k}\right)^{\frac{1}{p} - \frac{1}{2}} d^{\frac{1}{2} - \frac{1}{q}}.$$

1190 When $k \geq d^2$, we can show:

$$\left(\frac{d}{k}\right)^{\frac{1}{q} - \frac{1}{2}} \geq \left(\frac{d^2}{k}\right)^{\frac{1}{p} - \frac{1}{2}} d^{\frac{1}{2} - \frac{1}{q}} = \left(\frac{d}{k}\right)^{\frac{1}{p} - \frac{1}{2}} d^{\frac{1}{p} - \frac{1}{q}}.$$

1191 Therefore, for Case (b):

$$e_k \simeq \max\left\{\left(\frac{\log(em/k)}{k}\right)^{\frac{1}{p} - \frac{1}{2}}, \left(\frac{d}{k}\right)^{\frac{1}{q} - \frac{1}{2}}\right\}.$$

1192 **Case (c):** For $p > q$, we divide this case into two subcases based on the range of k .

1193 (i) For $\max\{d, \log m\} \leq k \leq md$: When $p > q$, we know $\phi(l)$ is decreasing in $[d, d^2]$. Analyzing
1194 $\phi(l)$:

- 1195 • For $1 \leq l \leq d$: Maximum occurs at $l = \max\{d, \log(em)\}$;
- 1196 • For $d \leq l \leq d^2$: Monotonically decreasing;
- 1197 • For $l \geq d^2$: Strictly decreasing;

1198 Therefore:

$$D(1, k) = \left(\frac{\max\{d, \log(em)\}}{k}\right)^{\frac{1}{p} - \frac{1}{2}}.$$

1199 (ii) For $md \leq k \leq md^2$: By similar analysis and considering $p > q$:

$$e_k \simeq m^{\frac{1}{q} - \frac{1}{p}} \left(\frac{d}{k}\right)^{\frac{1}{q} - \frac{1}{2}}.$$

1200 Under the assumption $k \geq m \cdot \max\{d, \log(em)\}$:

$$\left(\frac{\max\{d, \log(em)\}}{k}\right)^{\frac{1}{p} - \frac{1}{2}} \geq m^{\frac{1}{q} - \frac{1}{p}} \left(\frac{d}{k}\right)^{\frac{1}{q} - \frac{1}{2}} \iff \left(\frac{\max\{d, \log(em)\}}{d}\right)^{\frac{1}{p} - \frac{1}{2}} \geq m^{\frac{1}{q} - \frac{1}{p}}.$$

1201 This inequality holds under our assumptions.

1202 Combining all three cases yields the result in (39). □

1203 C.4.3 Proof of upper bounds for $\ell_p(S_q)$

1204 Recall that we define

$$\tilde{\mathbf{S}}(\mathbf{T}_{p,q}(R), \rho) = \{\mathbf{L} \in \mathbb{R}^{d \times m} : \|\mathbf{L}\|_F \leq \rho\} \cap \mathbf{T}_{p,q}(R).$$

1205 *Proof.* We prove the theorem by separating into three distinct cases based on the relationships among
1206 p , q , and m . In each case, we construct appropriate constants and verify the necessary conditions.

1207 **Case 1:** $p \leq q$ and $m \geq d^2$. We begin by adding a radius factor $R^{\frac{1}{p}}$ to the entropy number bounds
1208 in (39), yielding

$$\epsilon \leq R^{\frac{1}{p}} \left(\frac{d}{k}\right)^{\frac{1}{q} - \frac{1}{2}} + R^{\frac{1}{p}} \left(\frac{\log(em)}{k}\right)^{\frac{1}{p} - \frac{1}{2}}.$$

1209 Solving for k provides an upper bound on the covering number, giving

$$\log N(\epsilon; \tilde{\mathbf{S}}(\mathbf{T}_{p,q}(R), r)) \leq d(\epsilon^{-1} R^{\frac{1}{p}})^{\frac{2q}{2-q}} + \log(em) (\epsilon^{-1} R^{\frac{1}{p}})^{\frac{2p}{2-p}}. \quad (40)$$

1210 To apply Lemma 3.2 from [12], we need to construct constants (δ, ρ) satisfying two key conditions:

$$\sqrt{n} \delta \geq C_1 \rho \quad (\text{Condition 1}) \quad (41)$$

$$C_2 \sqrt{n} \delta \geq J(\rho, \delta) \quad (\text{Condition 2}) \quad (42)$$

1211 where

$$J(\rho, \delta) = \int_{\frac{\delta}{16}}^{\rho} \sqrt{\log N(t; \tilde{\mathbf{S}}(\mathbf{T}_{p,q}(R), \rho))} dt \leq \int_0^{\rho} \sqrt{d (t^{-1} R^{\frac{1}{p}})^{\frac{2q}{2-q}} + \log(em) (t^{-1} R^{\frac{1}{p}})^{\frac{2p}{2-p}}} dt.$$

1212 A direct calculation yields:

$$J(\rho, \delta) \leq \sqrt{d R^{\frac{q}{p(2-q)}}} \rho^{1-\frac{q}{2-q}} + \sqrt{\log(em)} R^{\frac{1}{2-p}} \rho^{1-\frac{p}{2-p}}.$$

1213 **Choice of constants ρ, δ .** Let

$$\rho = \Omega \left(R^{\frac{q}{2p}} \left(\frac{n}{d} \right)^{\frac{q-2}{4}} + R^{\frac{1}{2}} \left(\frac{n}{\log m} \right)^{\frac{p-2}{4}} \right) \wedge R^{\frac{1}{p}}, \quad (\rho \text{ definition})$$

$$\delta = C \rho \left(R^{\frac{q}{2p}} \left(\frac{n}{d} \right)^{\frac{q-2}{4}} + R^{\frac{1}{2}} \left(\frac{n}{\log m} \right)^{\frac{p-2}{4}} \right). \quad (\delta \text{ definition})$$

1214 *Verifying Condition (41).*

$$\frac{\sqrt{n} \delta}{\rho} \geq C \left[R^{\frac{q}{p}} n^{\frac{q}{2}} \left(\frac{1}{d} \right)^{\frac{q-2}{2}} + R n^{\frac{p}{2}} \left(\frac{1}{\log m} \right)^{\frac{p-2}{2}} \right] \geq C_1,$$

1215 where the second inequality follows from (19).

1216 *Verifying Condition (42).* Given that ρ is chosen as in ([ρ definition](#)), an analogous ratio bound shows

$$\frac{J(\rho, \delta)}{\sqrt{n} \delta} = \frac{\sqrt{d R^{\frac{q}{p(2-q)}}} \rho^{-\frac{q}{2-q}} + \sqrt{\log(m)} R^{\frac{1}{2-p}} \rho^{-\frac{p}{2-p}}}{C \sqrt{n} \left(R^{\frac{q}{2p}} \left(\frac{n}{d} \right)^{\frac{q-2}{4}} + R^{\frac{1}{2}} \left(\frac{n}{\log m} \right)^{\frac{p-2}{4}} \right)} = \frac{\sqrt{2}}{C},$$

1217 implying $C_2 \sqrt{n} \delta \geq J(\rho, \delta)$. Additionally, we must check that $(\frac{\delta}{16}, \rho)$ is a valid interval for covering;

1218 an argument similar to (40) and $\rho < R^{\frac{1}{p}}$ ensures

$$\log N(\delta; \tilde{\mathbf{S}}(\mathbf{T}_{p,q}(R), \rho)) \geq \max\{d, \log m\},$$

1219 so the condition for Lemma C.11 is met.

1220 Hence, applying Lemma C.11 gives that, with probability at least $1 - C_5 \exp\left\{-C_6 n \left[R^{\frac{q}{p}} \left(\frac{n}{d} \right)^{\frac{q-2}{2}} + \right. \right.$
 1221 $\left. \left. R \left(\frac{n}{\log m} \right)^{\frac{p-2}{2}} \right] \right\}$, we have

$$\sup_{\mathbf{L} \in \tilde{\mathbf{S}}(\mathbf{T}_{p,q}(R), \rho)} |\langle \bar{\mathbf{E}}, \mathbf{L} \rangle| \leq \left[R^{\frac{q}{2p}} \left(\frac{\sigma^2 n}{d} \right)^{\frac{q-2}{4}} + R^{\frac{1}{2}} \left(\frac{\sigma^2 n}{\log m} \right)^{\frac{p-2}{4}} \right] \rho. \quad (43)$$

1222 (The other two cases for $p \leq q, m \leq d^2$ and $p \geq q$ follow the same procedure, so we omit full detail
 1223 here.)

1224 Combining the arguments for all three regimes and applying Lemma C.11 completes the proof of the
 1225 upper bounds for $\ell_p(S_q)$. \square

D Experimental Details

This appendix provides a comprehensive description of the experimental setup and the ADMM-based optimization algorithm used for $\ell_p(S_q)$ -norm-based tensor completion.

D.1 Experimental Setup

Noisy Tensor Completion Task Formulation The noisy tensor completion problem aims to recover a structured tensor $\underline{\mathbf{L}}^*$ from a set of noisy and incomplete observations. This problem is particularly relevant in applications such as hyperspectral image restoration, video inpainting, and remote sensing data reconstruction, where missing and corrupted data are common due to sensor limitations or transmission errors.

We consider a third-order tensor $\underline{\mathbf{L}}^* \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ that represents a clean, fully observed data source. However, due to data corruption and missing values, we only have access to a partially observed noisy tensor $\underline{\mathbf{Y}}$, which is generated as:

$$\underline{\mathbf{Y}} = \underline{\mathbf{B}} \odot (\underline{\mathbf{L}}^* + \underline{\mathbf{E}}),$$

where:

- **$\underline{\mathbf{B}}$ (Binary Mask):** A binary tensor of the same size as $\underline{\mathbf{L}}^*$, where each entry $\underline{\mathbf{B}}_{i,j,k} \in \{0, 1\}$ indicates whether the corresponding entry in $\underline{\mathbf{L}}^*$ is observed ($\underline{\mathbf{B}}_{i,j,k} = 1$) or missing ($\underline{\mathbf{B}}_{i,j,k} = 0$).
- **\odot (Hadamard Product):** The element-wise product operator ensures that only the observed entries are retained, while unobserved entries are set to zero.
- **$\underline{\mathbf{E}}$ (Noise Tensor):** Represents random additive noise introduced in the observed entries. Each entry of $\underline{\mathbf{E}}$ is sampled independently from a Gaussian distribution:

$$\underline{\mathbf{E}}_{i,j,k} \sim \mathcal{N}(0, \sigma^2),$$

where the noise level σ is set as:

$$\sigma = c\sigma_0, \quad \text{with } c = 0.05, \quad \sigma_0 = \frac{\|\underline{\mathbf{L}}^*\|_F}{\sqrt{d_1 d_2 d_3}}.$$

Here, σ_0 represents a normalized noise scale based on the Frobenius norm of the clean tensor.

Sampling Strategy and Experimental Settings We apply a uniform random sampling strategy, where each entry of $\underline{\mathbf{L}}^*$ is independently observed with probability p , meaning that a fraction $1 - p$ of the entries is missing. We consider three different missing ratios: $p \in \{0.05, 0.1, 0.15\}$, which correspond to scenarios where 95%, 90%, and 85% of the entries are missing, respectively. Each experiment is conducted over 10 independent trials to ensure statistical reliability, and the averaged Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) are reported to evaluate reconstruction performance.

Evaluation Metrics To assess the quality of tensor reconstruction, we use the following two widely adopted metrics:

- **Peak Signal-to-Noise Ratio (PSNR):**

$$\text{PSNR} = 10 \log_{10} \left(\frac{\max(\underline{\mathbf{L}}^*)^2}{\frac{1}{d_1 d_2 d_3} \|\hat{\underline{\mathbf{L}}} - \underline{\mathbf{L}}^*\|_F^2} \right).$$

A higher PSNR value indicates better reconstruction quality.

- **Structural Similarity Index (SSIM):**

$$\text{SSIM}(\hat{\underline{\mathbf{L}}}, \underline{\mathbf{L}}^*) = \frac{(2\mu_{\hat{\underline{\mathbf{L}}}}\mu_{\underline{\mathbf{L}}^*} + c_1)(2\sigma_{\hat{\underline{\mathbf{L}}}\underline{\mathbf{L}}^*} + c_2)}{(\mu_{\hat{\underline{\mathbf{L}}}}^2 + \mu_{\underline{\mathbf{L}}^*}^2 + c_1)(\sigma_{\hat{\underline{\mathbf{L}}}}^2 + \sigma_{\underline{\mathbf{L}}^*}^2 + c_2)}.$$

This metric measures perceptual similarity between the recovered tensor $\hat{\underline{\mathbf{L}}}$ and the ground truth $\underline{\mathbf{L}}^*$, where $\mu_{\hat{\underline{\mathbf{L}}}}, \mu_{\underline{\mathbf{L}}^*}$ denote mean values, $\sigma_{\hat{\underline{\mathbf{L}}}}, \sigma_{\underline{\mathbf{L}}^*}$ denote standard deviations, and $\sigma_{\hat{\underline{\mathbf{L}}}\underline{\mathbf{L}}^*}$ represents cross-covariance. Parameters c_1 and c_2 are small constants to stabilize the division.

These metrics together provide a comprehensive evaluation of the reconstruction performance, ensuring that both numerical fidelity and structural integrity are preserved.

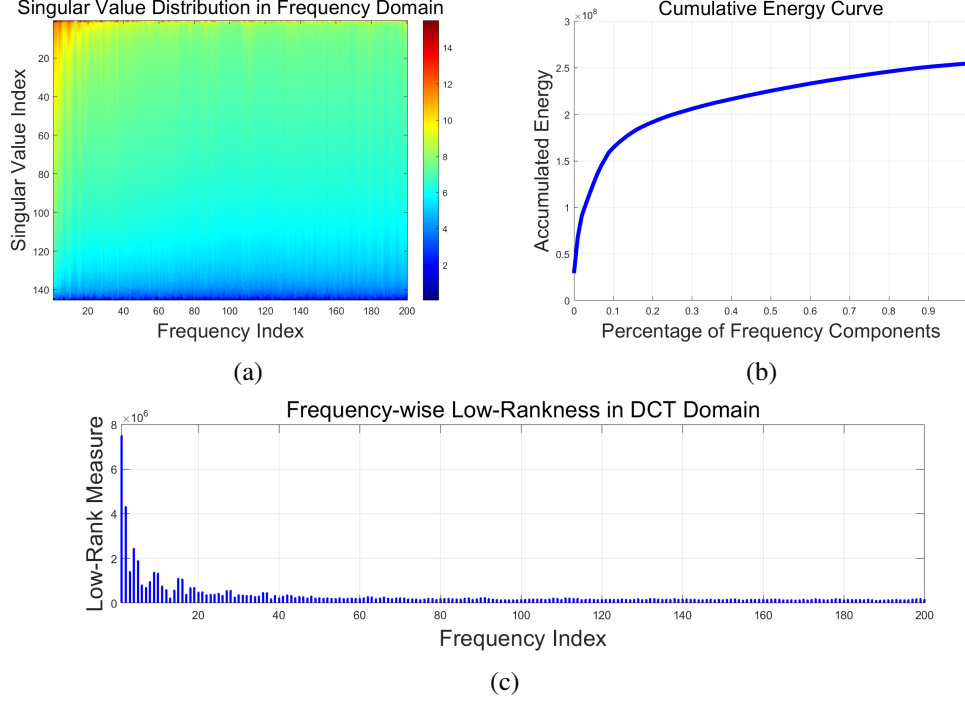


Figure 2: Visualization of dual-level sparsity structure using Indian Pines dataset. (a) The singular value heatmap exhibits both inter-frequency sparsity (horizontal variation) and intra-frequency low-rankness (vertical variation). (b) The cumulative energy curve reveals a majority of energy concentration in first 20% frequencies. (c) The frequency-wise low-rank measure $\|\sigma(M(\mathbf{T})^{(i)})\|_1$ shows significant peaks in low frequencies and rapid decay afterwards.

Benchmark Methods We compare the proposed $\ell_p(S_q)$ -quasi-norm against several existing low-rank tensor regularization techniques:

- *NN*: Matrix nuclear norm [6]
- *SNN*: Tucker-based tensor nuclear norm [27]
- *TNN-DFT*: Tensor nuclear norm with Discrete Fourier Transform [62]
- *TNN-DCT*: Tensor nuclear norm with Discrete Cosine Transform [31]
- *k-Supp*: Tensor k -Support norm ($k = 2$) [51]
- ℓ_{1-2} -norm: Tensor ℓ_{1-2} -norm [42]
- *Schatten- p -norm*: Tensor Schatten- p -norm ($p = 1/2$) [23]
- *LpSq (Proposed)*: The proposed $\ell_p(S_q)$ -norm with optimal parameters $(p, q) = (0.8961, 0.8966)$. We employ DCT as the transform operator $M(\cdot)$.

Parameter Tuning We employ a two-stage procedure to select the hyperparameters (p, q) , balancing systematic exploration with data-efficiency considerations.

- *Coarse grid search*: We begin by sweeping (p, q) over the grid $\{0.1, 0.2, \dots, 1.0\}^2$. PSNR is evaluated on the full set of observed entries (e.g., 20% of the tensor), ensuring that the tuning objective is directly aligned with the final completion goal without introducing artificial splits.
- *Manual refinement*: The coarse results suggest that optimal values lie near the diagonal $p = q \approx 0.9$. We then perform manual fine-tuning in the range $[0.88, 0.92]^2$, using PSNR on the same set of observed entries as the selection criterion. This refinement yields the best-performing pair $(p, q) = (0.8961, 0.8966)$.

We avoid splitting the already sparse observed data into separate training and validation subsets (e.g., tuning on only 10%), as this can reduce signal strength and lead to suboptimal parameter choices. Direct tuning on the full observation mask provides a more faithful estimate of recovery quality and is widely adopted in tensor completion practice.

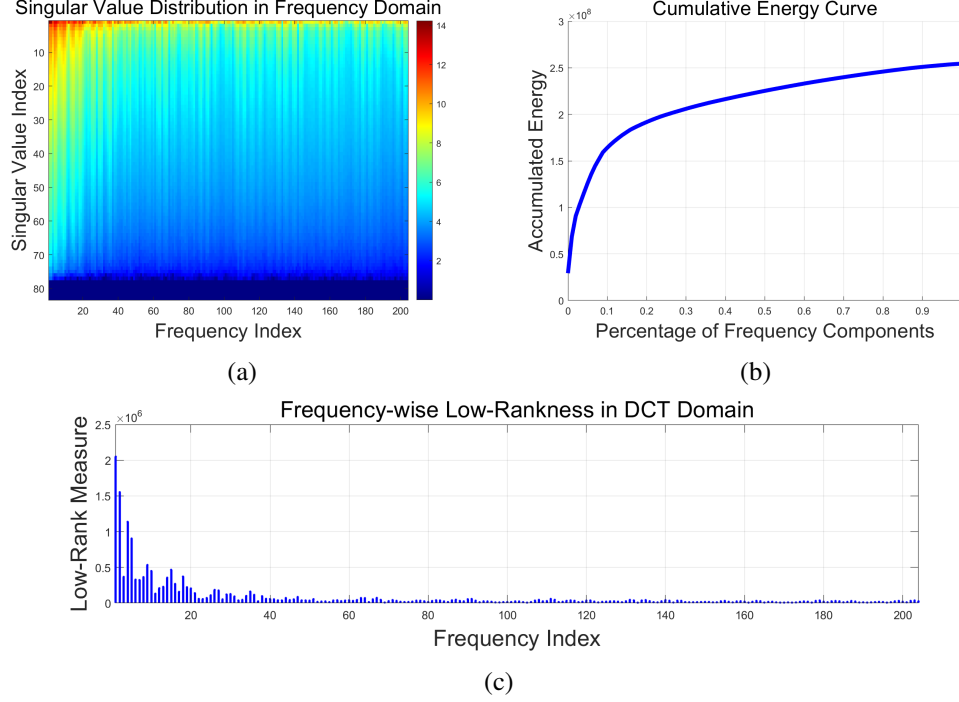


Figure 3: Visualization of dual-level sparsity structure using SalinasA dataset. (a) The singular value heatmap exhibits both inter-frequency sparsity (horizontal variation) and intra-frequency low-rankness (vertical variation). (b) The cumulative energy curve reveals a majority of energy concentration in first 20% frequencies. (c) The frequency-wise low-rank measure $\|\sigma(M(\mathbf{I})^{(i)})\|_1$ shows significant peaks in low frequencies and rapid decay afterwards.

1288 **Datasets** The evaluation is conducted across multiple remote sensing datasets, encompassing
 1289 hyperspectral, multispectral, and thermal imaging data.

1290 1. *Hyperspectral Data.* We conduct noisy tensor completion on subsets of two representative
 1291 hyperspectral datasets:

- 1292 • *Indian Pines:* This dataset was collected by the AVIRIS sensor in 1992 over the Indian
 1293 Pines⁴ test site in North-western Indiana and consists of 145×145 pixels and 200
 1294 corrected spectral reflectance bands. We use the first 30 bands in the experiments due to
 1295 computational constraints and parameter tuning.
- 1296 • *Salinas A:* Acquired by the AVIRIS sensor over the Salinas Valley, California, in 1998,
 1297 this dataset consists of 224 bands over a spectral range of 400–2500 nm, with a spatial
 1298 extent of 86×83 pixels at a resolution of 3.7m. We use the first 30 bands in the
 1299 experiments⁵.

1300 2. *Multispectral Images.* Multispectral imaging captures image data within specific wavelength
 1301 ranges across the electromagnetic spectrum and is one of the most widely used modalities in
 1302 remote sensing. This section presents simulated experiments on multispectral images. The
 1303 original data consists of three multispectral images: *Cloth*, *Hair*, and *Jelly Beans*, from the
 1304 Columbia MSI Database⁶. These images contain scenes of a variety of real-world objects,
 1305 each with a resolution of $512 \times 512 \times 31$, with intensity values scaled to $[0,1]$.

1306 3. *Thermal Imaging Data.* Thermal infrared data provide crucial measurements of surface
 1307 energy fluxes and temperatures for various remote sensing applications. We conduct ex-

⁴The dataset is publicly available at https://www.ehu.es/ccwintco/index.php?title=Hyperspectral_Remote_Sensing_Scenes.

⁵The dataset is available at https://www.ehu.es/ccwintco/index.php?title=Hyperspectral_Remote_Sensing_Scenes.

⁶Available at <http://www1.cs.columbia.edu/CAVE/databases/multispectral>

periments on an infrared dataset: *OSU Thermal Database*⁷. The sequences were recorded on the Ohio State University campus during February and March 2005, capturing multiple individuals moving in groups through the scene. We use the first 30 frames of Sequence 1, forming a tensor of size $320 \times 240 \times 30$.

1312 D.2 Tensor Completion via an ADMM-Based Algorithm

1313 We aim to estimate a structured tensor $\underline{\mathbf{L}}$ from noisy and incomplete observations $\underline{\mathbf{Y}}$. The optimization
1314 problem is formulated as:

$$\min_{\underline{\mathbf{L}}} \frac{1}{2} \|\underline{\mathbf{Y}} - \underline{\mathbf{B}} \odot \underline{\mathbf{X}}\|_{\text{F}}^2 + \lambda \|\underline{\mathbf{L}}\|_{\ell_p(S_q)}^p,$$

1315 We introduce an auxiliary variable $\underline{\mathbf{K}}$ and reformulate the problem as:

$$\min_{\underline{\mathbf{L}}, \underline{\mathbf{K}}} \frac{1}{2} \|\underline{\mathbf{Y}} - \underline{\mathbf{B}} \odot \underline{\mathbf{K}}\|_{\text{F}}^2 + \lambda \|\underline{\mathbf{L}}\|_{\ell_p(S_q)}^p, \quad \text{s.t.} \quad \underline{\mathbf{K}} = \underline{\mathbf{L}}.$$

1316 This reformulation decouples the data fidelity term from the regularization, making it amenable to
1317 optimization via the Alternating Direction Method of Multipliers (ADMM).

1318 We solve this problem using ADMM by iteratively updating $\underline{\mathbf{L}}$, $\underline{\mathbf{K}}$, and the dual variable $\underline{\mathbf{W}}$.

1319 **Update of $\underline{\mathbf{L}}$:** The $\underline{\mathbf{L}}$ -subproblem is:

$$\underline{\mathbf{L}}^{t+1} = \arg \min_{\underline{\mathbf{L}}} \lambda \|\underline{\mathbf{L}}\|_{\ell_p(S_q)}^p + \frac{\mu^t}{2} \|\underline{\mathbf{L}} - \underline{\mathbf{K}}^t - \frac{\underline{\mathbf{W}}^t}{\mu^t}\|_{\text{F}}^2.$$

1320 Let $\underline{\mathbf{Z}} = \underline{\mathbf{K}}^t - \frac{\underline{\mathbf{W}}^t}{\mu^t}$, reducing the above problem to:

$$\underline{\mathbf{L}}^{t+1} = \arg \min_{\underline{\mathbf{L}}} \lambda \|\underline{\mathbf{L}}\|_{\ell_p(S_q)}^p + \frac{\mu^t}{2} \|\underline{\mathbf{L}} - \underline{\mathbf{Z}}\|_{\text{F}}^2.$$

1321 Since $M(\underline{\mathbf{L}})$ allows separable updates across frequency components, this problem decomposes into
1322 m independent subproblems:

$$\min \frac{1}{2} \|M(\underline{\mathbf{L}})_k - M(\underline{\mathbf{Z}})_k\|_{\text{F}}^2 + \frac{\lambda}{\mu^t} \|M(\underline{\mathbf{L}})_k\|_{S_q}^{p/q}, \quad \forall k \in [m].$$

1323 To efficiently handle the Schatten- q term, we employ a *weighted $\ell_{1/2}$ -norm approximation* (see
1324 Section D.3 for comparison with ℓ_1 -norm approximation):

$$\sum_{i=1}^d w_{i,k} \cdot \sigma_i(M(\underline{\mathbf{L}})_k)^{1/2},$$

1325 where the weight

$$w_{i,k} = \left(\sum_{j=1}^d \varsigma_{j,k}^q + \epsilon \right)^{\frac{p}{q}-1} \left(\varsigma_{j,k}^{1/2} + \epsilon \right)^{2q-1},$$

1326 with $\varsigma_{j,k} = \sigma_j(M(\underline{\mathbf{L}}^t)_k)$, where $\underline{\mathbf{L}}^t$ denotes the tensor $\underline{\mathbf{L}}$ at the t -th iteration.

1327 This formulation leads to a closed-form $\ell_{1/2}$ -soft-thresholding update for each singular value:

$$\sigma_i^{(t+1)}(M(\underline{\mathbf{L}})_k) = \mathcal{S}_{\theta}^{\ell_{1/2}}(\sigma_i(M(\underline{\mathbf{Z}})_k)),$$

1328 where $\theta = \frac{\lambda}{\mu^t} w_{i,k}$ and the $\ell_{1/2}$ -soft-thresholding operator is defined as:

$$\mathcal{S}_{\theta}^{\ell_{1/2}}(\sigma) = \begin{cases} \phi_{\theta}(\sigma), & |\sigma| > \frac{3\sqrt[3]{54}}{4}\theta^{2/3}, \\ \{\phi_{\theta}(\sigma), 0\}, & |\sigma| = \frac{3\sqrt[3]{54}}{4}\theta^{2/3}, \\ 0, & |\sigma| < \frac{3\sqrt[3]{54}}{4}\theta^{2/3}. \end{cases} \quad (44)$$

1329 Here, the function $\phi_{\theta}(\sigma)$ is given by:

$$\phi_{\theta}(\sigma) = \frac{2}{3}\sigma \left(1 + \cos \left(\frac{2\pi}{3} - \frac{2}{3} \arccos \left(\frac{\theta}{8} |\sigma|^{-3/2} \right) \right) \right).$$

⁷The dataset is available at <http://vcip1-okstate.org/pbvs/bench/Data/03/download.html>.

1330 After updating singular values, the frequency component is reconstructed as:

$$M(\underline{\mathbf{L}}^{t+1})_k = \mathbf{U}_k \cdot \text{diag}(\boldsymbol{\sigma}^{(t+1)}(M(\underline{\mathbf{Z}})_k)) \cdot \mathbf{V}_k^\top, \quad \forall k \in [m],$$

1331 where \mathbf{U}_k and \mathbf{V}_k are the left and right singular matrices of $M(\underline{\mathbf{Z}})_k$. Finally, applying the inverse
1332 M -transform yields the updated tensor $\underline{\mathbf{L}}^{t+1}$.

1333 **Update of $\underline{\mathbf{K}}$:** The auxiliary variable $\underline{\mathbf{K}}$ is updated by solving:

$$\underline{\mathbf{K}}^{t+1} = \arg \min_{\underline{\mathbf{K}}} \frac{1}{2} \|\underline{\mathbf{Y}} - \underline{\mathbf{B}} \odot \underline{\mathbf{K}}\|_F^2 + \frac{\mu^t}{2} \left\| \underline{\mathbf{K}} - \underline{\mathbf{L}}^{t+1} + \frac{\underline{\mathbf{W}}^t}{\mu^t} \right\|_F^2.$$

1334 This step ensures that the solution remains within the feasible constraint region.

1335 **Dual Variable Update:** The Lagrange multiplier is updated as:

$$\underline{\mathbf{W}}^{t+1} = \underline{\mathbf{W}}^t + \mu^t (\underline{\mathbf{K}}^{t+1} - \underline{\mathbf{L}}^{t+1}).$$

1336 The penalty parameter μ^t is dynamically updated to ensure convergence:

$$\mu^{t+1} = \min\{\gamma\mu^t, \mu_{\max}\}.$$

1337 where $\gamma > 1$ is a predefined scaling factor.

1338 This ADMM-based algorithm (summarized in Algorithm 1) efficiently solves the dual-level sparse
1339 tensor completion problem by iteratively enforcing structured sparsity through proximal updates while
1340 maintaining computational efficiency. The proposed weighted $\ell_{1/2}$ -soft-thresholding mechanism
1341 ensures that the non-convex Schatten- q regularization is effectively handled in each iteration.

1342 **Complexity Analysis:** Each iteration of our algorithm involves (i) a linear transform on $d_1 d_2$ tubes
1343 of length m (reducible to $O(d_1 d_2 m \log m)$ using DCT or FFT), and (ii) m SVDs of $d_1 \times d_2$ matrices,
1344 yielding $O(md_1 d_2 \min(d_1, d_2))$ complexity. Despite the nonconvexity, our algorithm converges
efficiently in practice.

Algorithm 1 ADMM for dual-sparse tensor completion (ℓ_p -Schatten- q)

Input: Observed tensor $\underline{\mathbf{Y}}$; binary mask $\underline{\mathbf{B}}$; transform M (with inverse M^{-1}); regulariser λ ; parameters p, q ; penalty $\mu_0 > 0$, growth factor $\gamma > 1$, maximum μ_{\max} ; tolerance ε .

Output: Completed tensor $\underline{\mathbf{L}}$.

```

1: Initialise:  $\underline{\mathbf{L}}^0 \leftarrow 0, \underline{\mathbf{K}}^0 \leftarrow \underline{\mathbf{Y}}, \underline{\mathbf{W}}^0 \leftarrow 0, \mu \leftarrow \mu_0, t \leftarrow 0$ .
2: repeat
  1.  $\underline{\mathbf{L}}$ -update
  3:    $\underline{\mathbf{Z}} \leftarrow \underline{\mathbf{K}}^t - \underline{\mathbf{W}}^t / \mu$  ▷ dual correction
  4:   for each frequency slice  $k = 1, \dots, m$  do
  5:      $U_k \text{diag}(\boldsymbol{\sigma}_k) V_k^\top \leftarrow \text{SVD}(M(\underline{\mathbf{Z}})_k)$ 
  6:     Compute weights  $w_{i,k} = (\sum_j \varsigma_{j,k}^q + \epsilon)^{\frac{p}{q}-1} (\varsigma_{i,k}^{1/2} + \epsilon)^{2q-1}$ , where  $\varsigma_{i,k} = \sigma_i(M(\underline{\mathbf{L}}^t)_k)$ 
  7:      $\theta_{i,k} \leftarrow (\lambda/\mu) w_{i,k}$ 
  8:     Half-threshold:  $\sigma_{i,k}^{\text{new}} = \mathcal{S}_{\theta_{i,k}}^{\ell_{1/2}}(\sigma_{i,k})$  ▷ Eq. (56)–(57)
  9:      $M(\underline{\mathbf{L}}^{t+1})_k \leftarrow U_k \text{diag}(\boldsymbol{\sigma}_k^{\text{new}}) V_k^\top$ 
  10:   end for
  11:    $\underline{\mathbf{L}}^{t+1} \leftarrow M^{-1}(\{M(\underline{\mathbf{L}}^{t+1})_k\}_{k=1}^m)$ 
  2.  $\underline{\mathbf{K}}$ -update
  12:    $\underline{\mathbf{K}}^{t+1} \leftarrow (\underline{\mathbf{B}} \odot \underline{\mathbf{Y}} + \mu(\underline{\mathbf{L}}^{t+1} - \underline{\mathbf{W}}^t / \mu)) \odot (\underline{\mathbf{B}} + \mu \mathbf{1})$  ▷ element-wise  $\odot$  division
  3. Dual update
  13:    $\underline{\mathbf{W}}^{t+1} \leftarrow \underline{\mathbf{W}}^t + \mu(\underline{\mathbf{K}}^{t+1} - \underline{\mathbf{L}}^{t+1})$ 
  4. Penalty update
  14:    $\mu \leftarrow \min(\gamma\mu, \mu_{\max})$ 
  15:    $t \leftarrow t + 1$ 
  16: until  $\|\underline{\mathbf{K}}^t - \underline{\mathbf{L}}^t\|_F / \|\underline{\mathbf{Y}}\|_F < \varepsilon$ 
  17: return  $\underline{\mathbf{L}} \leftarrow \underline{\mathbf{L}}^t$ 

```

1345

1346 D.3 Choice of Weighted $\ell_{1/2}$ Approximation

1347 To approximate the nonconvex Schatten- q regularizer in a computationally efficient manner, we
 1348 considered both weighted ℓ_1 and weighted $\ell_{1/2}$ surrogates during model development. While the
 1349 weighted ℓ_1 formulation is convex and widely used, it proved empirically less effective in preserving
 1350 the underlying structure of transformed tensor slices. As shown in Table 3, the weighted $\ell_{1/2}$
 1351 surrogate consistently achieves higher PSNR and SSIM scores across different datasets and sampling
 1352 rates, including *Salinas A* and *Indian Pines*. This improvement suggests that the nonconvex $\ell_{1/2}$
 1353 penalty better captures the spectral decay and low-rank characteristics inherent in each transformed
 1354 slice. Based on these results, we adopt the weighted $\ell_{1/2}$ approximation throughout our method.

Table 3: Preliminary results of weighted ℓ_1 -based approximation

Dataset	SR	Metric	Weighted ℓ_1	Weighted $\ell_{1/2}$
<i>Salinas A</i>	5%	PSNR	28.03	28.43
		SSIM	0.7297	0.7374
	10%	PSNR	30.38	31.81
		SSIM	0.8126	0.8484
	15%	PSNR	32.38	33.23
		SSIM	0.8588	0.8830
<i>Indian Pines</i>	5%	PSNR	25.73	27.05
		SSIM	0.6600	0.6740
	10%	PSNR	26.88	28.92
		SSIM	0.6680	0.7617
	15%	PSNR	27.04	29.89
		SSIM	0.6731	0.7997

1355 D.4 Convergence of the Proposed ADMM-Based Algorithm

1356 Our ADMM-based algorithm exhibits stable empirical convergence across all experiments. As
 1357 illustrated in Figure 4, the objective value typically stabilizes within 200 iterations. This behavior is
 1358 consistent across different sampling ratios and datasets.

1359 From a theoretical perspective, establishing convergence guarantees is challenging due to the non-
 1360 convex and nonsmooth nature of the objective, which involves transform-domain operations, SVD,
 1361 and a weighted $\ell_{1/2}$ -quasi-norm. These components render standard ADMM theory inapplicable.
 1362 While recent advances in nonconvex ADMM provide promising tools, they require problem-specific
 1363 adaptations to accommodate the spectral structure and inter-frequency coupling present in our model.
 1364 We consider this a promising direction for future analysis and ongoing work.

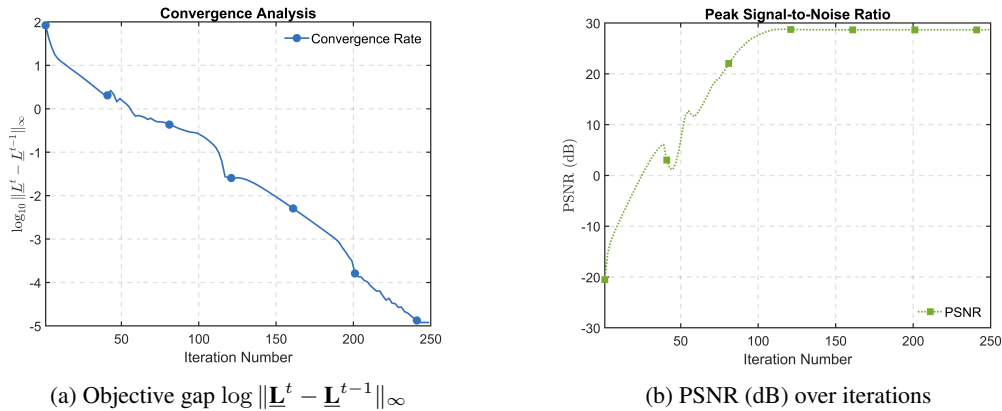


Figure 4: Empirical convergence of the proposed ADMM algorithm on the *Salinas A* dataset. Left: the variable gap between consecutive iterates. Right: the evolution of PSNR across iterations.