Convolutional Analysis Operator Learning: Dependence on Training Data

Il Yong Chun, *Member, IEEE*, David Hong, *Student Member, IEEE*, Ben Adcock, and Jeffrey A. Fessler, *Fellow, IEEE*

***Abstract*—Convolutional analysis operator learning (CAOL) enables the unsupervised training of (hierarchical) convolutional sparsifying operators or autoencoders from large datasets. One can use many training images for CAOL, but a precise understanding of the impact of doing so has remained an open question. This letter presents a series of results that lend insight into the impact of dataset size on the filter update in CAOL. The first result is a general deterministic bound on errors in the estimated filters, and is followed by a bound on the *expected* errors as the number of training samples increases. The second result provides a *high probability* analogue. The bounds depend on properties of the training data, and we investigate their empirical values with real data. Taken together, these results provide evidence for the potential benefit of using more training data in CAOL.**

***Index Terms*—Convolutional neural network, unsupervised learning, dependence on training sample size.**

I. Introduction

LEARNING convolutional operators from large datasets is a growing trend in signal/image processing, computer vision, machine learning, and artificial intelligence. The *convolutional* approach resolves the large memory demands of patch-based operator learning and enables unsupervised operator learning from "big data," i.e., many high-dimensional signals. See [1], [2] and references therein. Examples include *convolutional dictionary learning* [2], [3] and *convolutional analysis operator learning* (CAOL) [1], [4]. CAOL trains an autoencoding CNN in an unsupervised manner, and is useful for training multi-layer CNNs from many training images [1]. In particular, the block proximal gradient method using a majorizer [1], [2] leads to rapidly converging and memory-efficient CAOL [1]. However, a theoretical understanding of the impact of using many training images in CAOL has remained an open question.

This letter presents new insights on this topic. Our first main result provides a *deterministic* bound on filter estimation error, and is followed by a bound on the *expected* error when "model mismatch" has zero mean. (See Theorem 1 and Corollary 2, respectively.) The expected error bound depends on the training data, and we provide empirical evidence of its decrease with an increase in training samples. Our second main result provides a *high probability* bound that explicitly decreases with increasingly many i.i.d. training samples. The bound improves when model mismatch and samples are uncorrelated. (See Theorem 3.) Additional empirical findings provide evidence that the correlation can indeed be small in practice. Put together, our findings provide new insight into how using many samples can improve CAOL, underscoring the benefits of the low memory usage of CAOL.

Manuscript received February 21, 2019; revised April 25, 2019; accepted May 28, 2019. Date of publication June 7, 2019; date of current version June 24, 2019. This work was supported in part by the Keck Foundation and in part by NIH under Grant U01 EB018753. The work of B. Adcock was supported by NSERC under Grant 611675. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Qing Ling. *(Il Yong Chun and David Hong contributed equally to this work.) (Corresponding author: Il Yong Chun.)*

I. Y. Chun, D. Hong, and J. A. Fessler are with the Department of Electrical Engineering and Computer Science, The University of Michigan, Ann Arbor, MI 48019 USA (e-mail: iychun@umich.edu; dahong@umich.edu; fessler@umich.edu).

B. Adcock is with the Department of Mathematics, Simon Fraser University, Burnaby BC V5A 1S6, Canada (e-mail: ben_adcock@sfu.ca).

Digital Object Identifier 10.1109/LSP.2019.2921446

II. Backgrounds and Preliminaries

A. CAOL With Orthogonality Constraints

CAOL seeks a set of filters that "best" sparsify a set of training images $\{x_l \in \mathbb{C}^N : l = 1, \dots, L\}$ by solving the optimization problem [1, §II-A] (see Appendix for notation):

$$\underset{D=[d_1,\dots,d_K]}{\operatorname{argmin}} \min_{\{z_{l,k}\}} F(D, \{z_{l,k}\}), \text{ subj. to } DD^H = \frac{1}{R} \cdot I, \quad \text{(P0)}$$

$$F(D, \{z_{l,k}\}) := \sum_{l=1}^{L} \sum_{k=1}^{K} \|d_k \circledast x_l - z_{l,k}\|_2^2 + \alpha \|z_{l,k}\|_0,$$

where $\circledast$ denotes convolution, $\{d_k \in \mathbb{C}^R : k = 1, \dots, K\}$ is a set of $K \geq R$ convolutional kernels, $\{z_{l,k} \in \mathbb{C}^N : l = 1, \dots, L, k = 1, \dots, K\}$ is a set of sparse codes, $\alpha > 0$ is a regularization parameter controlling the sparsity of features $\{z_{l,k}\}$, and $\|\cdot\|_0$ denotes the ℓ^0-quasi-norm. We group the K filters into a matrix:

$$D := [\, d_1 \ \cdots \ d_K \,] \in \mathbb{C}^{R \times K}. \quad (1)$$

The orthogonality condition $DD^H = \frac{1}{R} I$ in (P0) enforces *1)* a tight-frame condition on the filters, i.e., $\sum_{k=1}^{K} \|d_k \circledast x\|_2^2 = \|x\|_2^2$, $\forall x$ [1, Prop. 2.1]; and *2)* filter diversity when $R = K$, since $DD^H = \frac{1}{R} I$ implies $D^H D = \frac{1}{K} I$ and each pair of filters is incoherent, i.e., $|\langle d_k, d_{k'} \rangle|^2 = 0$, $\forall k \neq k'$. One often solves (P0) iteratively, by alternating between optimizing D (filter update) and optimizing $\{z_{l,k} : \forall l, k\}$ (sparse code update) [1], i.e., at the i iteration, the current iterates are updated as $\{z_{l,k}^{(i+1)}\} = \operatorname{argmin}_{\{z_{l,k}\}} F(D^{(i)}, \{z_{l,k}\})$ and $D^{(i+1)} = \operatorname{argmin}_{DD^H = \frac{1}{R} \cdot I} F(D, \{z_{l,k}^{(i+1)}\})$.

B. Filter Update in a Matrix Form

The key to our analysis lies in rewriting the filter update for (P0) in matrix form, to which we apply matrix perturbation and

concentration inequalities. Observe first that

$$d_k \circledast x_l = \underbrace{[\Pi^0 x_l, \ldots, \Pi^{R-1} x_l]}_{=:\Psi_l \in \mathbb{C}^{N\times R}} d_k = \Psi_l d_k, \quad l = 1, \ldots, L, \tag{2}$$

where $\Pi := \left[\begin{smallmatrix} 0 & I_{N-1} \\ 1 & 0 \end{smallmatrix}\right] \in \mathbb{C}^{N\times N}$ is the circular shift operator and $(\cdot)^n$ denotes the matrix product of its n copies. We consider a circular boundary condition to simplify the presentation of $\{\Psi_l\}$ in (2), but our entire analysis holds for a general boundary condition with only minor modifications of $\{\Psi_l\}$ as done in [1, §IV-A]. Using (2), the filter update of (P0) is rewritten as

$$D^\star = \underset{D}{\operatorname{argmin}} \sum_{l=1}^{L} \|\Psi_l D - Z_l\|_F^2, \quad \text{subj. to } DD^H = \frac{1}{R} \cdot I, \tag{P1}$$

where $Z_l := [z_{l,1}, \ldots, z_{l,K}] \in \mathbb{C}^{N\times K}$ contains all the current sparse code estimates for the lth sample, and we drop iteration superscript indices $(\cdot)^{(i)}$ throughout. The next section uses this form to characterize the filter update solution $D^\star$.

III. Main Results: Dependence of CAOL on Training Data

The main results in this section illustrate how training with many samples can reduce errors in the filter $D^\star$ from (P1) and characterize the reduction in terms of properties of the training data. Throughout we model the current sparse codes estimates as

$$Z_l = \underbrace{\Psi_l D_{\text{true}}}_{=:Z_{\text{true},l}} + E_l, \quad l = 1, \ldots, L \tag{3}$$

where D_{true} is formed from optimal (orthogonal) filters analogously to (1), and $E_l \in \mathbb{C}^{N\times K}$ captures model mismatch in the current sparse codes, e.g., due to the current iterate being far from convergence or being trapped in local minima.

The following theorem provides a *deterministic* characterization.

Theorem 1: Suppose that both matrices

$$\sum_{l=1}^{L} \Psi_l^H Z_l \in \mathbb{C}^{R\times K} \quad \text{and} \quad \sum_{l=1}^{L} \Psi_l^H Z_{\text{true},l} \in \mathbb{C}^{R\times K} \tag{4}$$

are full row rank, where $\{\Psi_l, Z_l, Z_{\text{true},l} : l = 1, \ldots, L\}$ are defined in (2)–(3). Then, the solution $D^\star$ to (P1) has error with respect to D_{true} bounded as

$$\|D^\star - D_{\text{true}}\|_F^2 \le 5 \frac{\|\sum_{l=1}^{L} \Psi_l^H E_l\|_F^2}{\lambda_{\min}^2(\sum_{l=1}^{L} \Psi_l^H \Psi_l)}, \tag{5}$$

where $\lambda_{\min}(\cdot)$ denotes the smallest eigenvalue of its argument.

The full row rank condition on (4) ensures that the estimated filters $D^\star$ and the true filters D_{true} are unique, and it further guarantees that the denominator of (5) is strictly positive. When the model mismatches $E_1, \ldots, E_L$ are independent and mean zero, we obtain the following expected error bound:

Corollary 2: Under the construction of Theorem 1, suppose that E_l is a zero-mean random matrix for $l = 1, \ldots, L$, and is independent over l. Then,

$$\mathbb{E}\|D^\star - D_{\text{true}}\|_F^2 \le 5\bar{\sigma}^2\rho^2, \tag{6}$$

where $\mathbb{E}(\cdot)$ denotes the expectation,

$$\bar{\sigma}^2 := \max_{l=1,\ldots,L} \lambda_{\max}(\mathbb{E}\{E_l E_l^H\}),$$

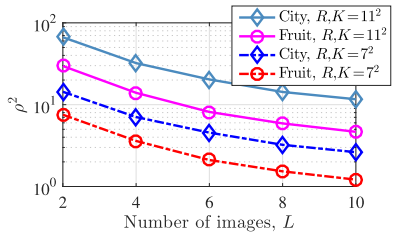

Fig. 1. Empirical values of ρ^2 in (7) show a decrease with L for different datasets and filter dimensions. (The fruit and city datasets with $L=10$ and $N=10^4$ were preprocessed with contrast enhancement and mean subtraction; see details of datasets and experiments in [1], [2] and references therein. For $L < 10$, the results are averaged over 50 datasets randomly selected from the full datasets.) Under the assumptions of Corollary 2, the decrease in this quantity leads to a better expected error bound in (5). Without preprocessing, the quantity ρ^2 increases by a factor of around 10^3.

$$\rho^2 := \frac{\operatorname{tr}(\sum_{l=1}^{L} \Psi_l^H \Psi_l)}{\lambda_{\min}^2(\sum_{l=1}^{L} \Psi_l^H \Psi_l)}, \tag{7}$$

$\lambda_{\max}(\cdot)$ denotes the largest eigenvalue of its argument, and the expectation is taken over the model mismatch.

Given fixed K and R, it is natural to expect that $\bar{\sigma}^2$ is bounded by some constant independent of L, and so the expected error bound in (6) largely depends on ρ^2 in (7). When training samples are i.i.d., one may further expect $(1/L)\sum_{l=1}^{L} \Psi_l^H \Psi_l$ to concentrate around its expectation, roughly resulting in $\rho^2 \propto 1/L$, with a proportionality constant that depends on R and the statistics of the training data. Fig. 1 illustrates ρ^2 for various image datasets, providing empirical evidence of this decrease in real data.

Our second theorem provides a *probabilistic* error bound via concentration inequalities, given i.i.d. training sample and model mismatch pairs $(x_1, E_1), \ldots, (x_L, E_L)$.[1] It removes the zero-mean assumption for the model mismatches $\{E_l : \forall l\}$ in Corollary 2 that might be strong, e.g., if training data are not preprocessed to have zero mean.

Theorem 3: Suppose that training sample and model mismatch pairs $(x_1, E_1), \ldots, (x_L, E_L) \overset{iid}{\sim} (x, E)$, where x and E are almost surely bounded, i.e.,

$$\|x\|_2 \le \gamma \quad \text{and} \quad \|E\|_F \le \sigma, \tag{8}$$

and the matrices in (4) are almost surely full row rank. Then, for any $0 < \delta < \lambda_{\min}(\bar{\Lambda})/(2R\gamma^2)$, the solution $D^\star$ to (P1) has error with respect to D_{true} bounded as

$$\|D^\star - D_{\text{true}}\|_F^2 \le 5\left\{ \frac{\sigma\sqrt{\operatorname{tr}(\bar{\Lambda})/L} + \|\mathbb{E}(\Psi^H E)\|_F + 2\sigma\gamma\sqrt{R}\delta}{\lambda_{\min}(\bar{\Lambda}) - 2\gamma^2 R\delta} \right\}^2, \tag{9}$$

with probability at least

$$1 - 3R\exp\left(-L\frac{\delta^2/2}{3 + \delta/3}\right), \tag{10}$$

where $\bar{\Lambda} := \mathbb{E}(\Psi^H\Psi)$ and Ψ is constructed from x as in (2).

Taking δ sufficiently small, the high probability error bound (9) is primarily driven by

$$\bar{\rho} := \frac{\sqrt{\operatorname{tr}(\bar{\Lambda})/L}}{\lambda_{\min}(\bar{\Lambda})} \quad \text{and} \quad \bar{\chi} := \frac{\|\mathbb{E}(\Psi^H E)\|_F}{\lambda_{\min}(\bar{\Lambda})}, \tag{11}$$

[1]We follow the natural convention in sample size analyses of assuming that $\{x_l : \forall l\}$ are i.i.d. samples from an underlying training distribution; see the references cited in Section IV and [5], [6] for other examples. Model mismatches $\{E_l : \forall l\}$ also become i.i.d. across samples at all iterations of CAOL, if "fresh" training samples are used for each update, e.g., as can be done when solving (P1) via mini-batch stochastic optimization.

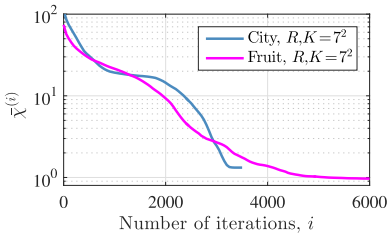

Fig. 2. Empirical estimate of $\bar{\chi}$ in (11) across iterations in the alternating optimization algorithm [1] that solves CAOL (P0) with $\alpha = 10^{-3}$. (The fruit and city datasets with $L = 10$ and $N = 10^4$ were preprocessed with contrast enhancement and mean subtraction; see details of datasets and experiments in [1], [2] and references therein. The model mismatches $\{E_l^{(i)} : \forall l\}$ at the ith iteration were calculated every 50 iterations based on (3), where we use the converged filters for D_{true}.) Observe that $\bar{\chi}^{(i)}$ generally decreases over iterations; when $\bar{\chi}$ is small, the high probability error bound (9) in Theorem 3 depends primarily on $\bar{\rho}$ defined in (11).

where $\bar{\rho}$ is analogous to ρ in (7), and $\bar{\chi}$ captures how correlated the model mismatch is to the training samples. As the number L of training samples increases, $\bar{\rho}$ decreases as $1/\sqrt{L}$. On the other hand, $\bar{\chi}$ is constant with respect to L and provides a floor for the bound. Fig. 2 illustrates $\bar{\chi}$ for CAOL iterates from different image datasets, and provides empirical evidence that this term can indeed be small in real data. If the model mismatch is sufficiently uncorrelated with the training samples, i.e., $\bar{\chi}$ is practically zero, then only the $\bar{\rho}$ term remains and this term decreases with L. Namely, if model mismatch is entirely uncorrelated with the training samples, then using many samples decreases the error bound to (effectively) zero.

IV. Related Works

Sample complexity [7] and synthesis (or reconstruction) error [8] have been studied in the context of synthesis operator learning (e.g., dictionary learning [9]); see the cited papers and references therein. A similar understanding for (C)AOL has however remained largely open; existing works focus primarily on establishing (C)AOL models and their algorithmic challenges [1], [10]–[13]. The authors in [14] studied sample complexity for a patch-based AOL method, but the form of their model differs from that of ours (P0). Specifically, they consider the following AOL problem: $\min_D \sum_l f(D^T \hat{x}_l) + g(D)$, where $f(\cdot)$ is a sparsity promoting function (e.g., a smooth approximation of the ℓ^0-quasi-norm [14]), $g(\cdot)$ is a regularizer or constraint for the filter matrix D, and $\{\hat{x}_l : l = 1, \dots, L\}$ is a set of training *patches* (not images).

V. Proof of Theorem 1

Rewriting (P1) yields that $D^\star$ is a solution of the (scaled) orthogonal Procrustes problem [1, §S.VII]:

$$\underset{D}{\operatorname{argmin}} \|\widetilde{\Psi} D - \widetilde{Z}\|_F^2, \quad \text{subj. to } DD^H = \frac{1}{R} \cdot I, \tag{12}$$

where $\widetilde{\Psi} \in \mathbb{C}^{LN \times R}$ arises by stacking $\Psi_1, \dots, \Psi_L$ vertically and $\widetilde{Z} \in \mathbb{C}^{LN \times K}$ arises likewise from $Z_1, \dots, Z_L$. Similarly, since $\Psi_l D_{\text{true}} = Z_{\text{true},l}$ as in (3), D_{true} is a solution of the analogous (scaled) orthogonal Procrustes problem

$$\underset{D}{\operatorname{argmin}} \|\widetilde{\Psi} D - \widetilde{Z}_{\text{true}}\|_F^2, \quad \text{subj. to } DD^H = \frac{1}{R} \cdot I, \tag{13}$$

where $\widetilde{Z}_{\text{true}} \in \mathbb{C}^{LN \times K}$ arises by stacking $Z_{\text{true},1}, \dots, Z_{\text{true},L}$ vertically.

By assumption, both $\widetilde{\Psi}^H \widetilde{Z}$ and $\widetilde{\Psi}^H \widetilde{Z}_{\text{true}}$ are full row rank and so (12) and (13) have unique solutions given by the unique (scaled) polar factors

$$D^\star = \frac{1}{\sqrt{R}} \mathrm{Q}(\widetilde{Z}^H \widetilde{\Psi})^H \qquad D_{\text{true}} = \frac{1}{\sqrt{R}} \mathrm{Q}(\widetilde{Z}_{\text{true}}^H \widetilde{\Psi})^H \tag{14}$$

where $\mathrm{Q}(\cdot)$ denotes the polar factor of its argument, and can be computed as $\mathrm{Q}(A) = WV^H$ from the (thin) singular value decomposition $A = W\Sigma V^H$.

Thus we have

$$\begin{aligned}
&\|D^\star - D_{\text{true}}\|_F^2 \\
&= \frac{1}{R} \|\mathrm{Q}(\widetilde{Z}_{\text{true}}^H \widetilde{\Psi}) - \mathrm{Q}(\widetilde{Z}^H \widetilde{\Psi})\|_F^2 \\
&\le \frac{1}{R} \|\widetilde{E}^H \widetilde{\Psi}\|_F^2 \left\{ \left[\frac{2}{\sigma_R(\widetilde{Z}_{\text{true}}^H \widetilde{\Psi}) + \sigma_R(\widetilde{Z}^H \widetilde{\Psi})} \right]^2 \right. \\
&\qquad \left. + \left[\frac{1}{\max\{\sigma_R(\widetilde{Z}_{\text{true}}^H \widetilde{\Psi}), \sigma_R(\widetilde{Z}^H \widetilde{\Psi})\}} \right]^2 \right\} \\
&\le \frac{1}{R} \|\widetilde{E}^H \widetilde{\Psi}\|_F^2 \left\{ \left[\frac{2}{\sigma_R(\widetilde{Z}_{\text{true}}^H \widetilde{\Psi})} \right]^2 + \left[\frac{1}{\sigma_R(\widetilde{Z}_{\text{true}}^H \widetilde{\Psi})} \right]^2 \right\} \\
&= \frac{5}{R} \frac{\|\widetilde{\Psi}^H \widetilde{E}\|_F^2}{\sigma_R^2(\widetilde{\Psi}^H \widetilde{Z}_{\text{true}})} = \frac{5}{R} \frac{\|\sum_{l=1}^L \Psi_l^H E_l\|_F^2}{\sigma_R^2(\sum_{l=1}^L \Psi_l^H Z_{\text{true},l})}
\end{aligned} \tag{15}$$

where $\widetilde{E} = \widetilde{Z} - \widetilde{Z}_{\text{true}}$ is exactly $E_1, \dots, E_L$ stacked vertically, and $\sigma_r(\cdot)$ denotes the rth largest singular value of its argument. The first inequality holds by the perturbation bound in [15, Thm. 3], and the second holds since $\sigma_R(\widetilde{Z}^H \widetilde{\Psi}) \ge 0$. Recalling that $Z_{\text{true},l} = \Psi_l D_{\text{true}}$, we rewrite the denominator of (15) as

$$\begin{aligned}
&\sigma_R^2\left(\sum_{l=1}^L \Psi_l^H Z_{\text{true},l} \right) = \sigma_R^2\left(\sum_{l=1}^L \Psi_l^H \Psi_l D_{\text{true}} \right) \\
&= \frac{1}{R} \sigma_R^2\left(\sum_{l=1}^L \Psi_l^H \Psi_l \right) = \frac{1}{R} \lambda_{\min}^2\left(\sum_{l=1}^L \Psi_l^H \Psi_l \right),
\end{aligned} \tag{16}$$

where the second equality holds because $D_{\text{true}} D_{\text{true}}^H = (1/R) I$. Substituting (16) into (15) yields (5).

VI. Proof of Corollary 2

Taking the expectation of (5) over the model mismatch amounts to taking the expectation of the numerator of the upper bound in (5):

$$\begin{aligned}
&\mathbb{E} \Big\| \sum_{l=1}^L \Psi_l^H E_l \Big\|_F^2 \\
&= \sum_{l=1}^L \mathbb{E} \big\| \Psi_l^H E_l \big\|_F^2 = \sum_{l=1}^L \mathrm{tr}\Big(\Psi_l^H \mathbb{E}\{E_l E_l^H\} \Psi_l \Big) \\
&\le \sum_{l=1}^L \lambda_{\max}(\mathbb{E}\{E_l E_l^H\}) \cdot \|\Psi_l\|_F^2 \le \bar{\sigma}^2 \cdot \sum_{l=1}^L \|\Psi_l\|_F^2,
\end{aligned} \tag{17}$$

where the first equality holds by using the assumption that E_l is zero-mean and independent over l, the second equality follows by expanding the Frobenius norm then applying linearity of the trace and expectation, the first inequality holds

since $v^H M v \le \lambda_{\max}(M) \cdot \|v\|_2^2$ for any vector v and Hermitian matrix M, and the last inequality follows from the definition of $\bar{\sigma}^2$. Rewriting (17) using the identity $\sum_{l=1}^L \|\Psi_l\|_F^2 = \sum_{l=1}^L \mathrm{tr}(\Psi_l^H \Psi_l) = \mathrm{tr}(\sum_{l=1}^L \Psi_l^H \Psi_l)$ yields the result (6).

VII. Proof of Theorem 3

We derive two high probability bounds, one each for the numerator and denominator of (5). Then, the bound (9) with probability (10) follows by combining the two via a union bound. Before we begin, note that (8) implies that $\|\Psi\|_2 \le \|\Psi\|_F \le \gamma\sqrt{R}$ almost surely; our proofs use this inequality multiple times.

A. Upper Bound for Numerator

Observe first that

$$\left\| \sum_{l=1}^L \Psi_l^H E_l \right\|_F = \left\| L\mathbb{E}(\Psi_l^H E_l) + \sum_{l=1}^L \{\Psi_l^H E_l - \mathbb{E}(\Psi_l^H E_l)\} \right\|_F \le L\|\mathbb{E}(\Psi_l^H E_l)\|_F + \left\| \sum_{l=1}^L \xi_l \right\|_2, \tag{18}$$

where $\xi_l := \mathrm{vec}\{\Psi_l^H E_l - \mathbb{E}(\Psi_l^H E_l)\} \in \mathbb{C}^{RK}$ for $l = 1, \ldots, L$. We next bound $\|\sum_{l=1}^L \xi_l\|_2$ via the vector Bernstein inequality [16, Cor. 8.44]. Note that $\xi_1, \ldots, \xi_L$ are i.i.d. with $\mathbb{E}\xi_l = 0$ (by construction). Furthermore, ξ_l is almost surely bounded as

$$\begin{aligned} \|\xi_l\|_2 &= \|\Psi_l^H E_l - \mathbb{E}(\Psi_l^H E_l)\|_F \\ &\le \|\Psi_l^H E_l\|_F + \|\mathbb{E}(\Psi_l^H E_l)\|_F && \text{(Triangle ineq.)} \\ &\le \|\Psi_l^H E_l\|_F + \mathbb{E}\|\Psi_l^H E_l\|_F && \text{(Jensen's ineq.)} \\ &\le \|\Psi_l\|_F \|E_l\|_F + \mathbb{E}\|\Psi_l\|_F \|E_l\|_F \\ &\le 2\sigma\gamma\sqrt{R}. \end{aligned}$$

Thus the vector Bernstein inequality [16, Cor. 8.44] yields that for any $t > 0$,

$$\left\| \sum_{l=1}^L \xi_l \right\|_2 \le \sigma\sqrt{L}\sqrt{\mathrm{tr}(\bar{\Lambda})} + t, \tag{19}$$

with probability at least

$$1 - \exp\left\{ \frac{-t^2/2}{3L(2\sigma\gamma\sqrt{R})^2 + t(2\sigma\gamma\sqrt{R})/3} \right\}. \tag{20}$$

We obtained (19) by the following simplification:

$$\begin{aligned} \mathbb{E}\left\| \sum_{l=1}^L \xi_l \right\|_2 &\le \sqrt{\mathbb{E}\left\| \sum_{l=1}^L \xi_l \right\|_2^2} = \sqrt{L\mathbb{E}\|\xi_l\|_2^2} \\ &= \sqrt{L\mathbb{E}\|\Psi_l^H E_l - \mathbb{E}(\Psi_l^H E_l)\|_F^2} \\ &= \sqrt{L\{\mathbb{E}\|\Psi_l^H E_l\|_F^2 - \|\mathbb{E}(\Psi_l^H E_l)\|_F^2\}} \\ &\le \sqrt{L\mathbb{E}\|\Psi_l^H E_l\|_F^2} \le \sqrt{L\mathbb{E}(\|\Psi_l\|_F^2 \|E_l\|_F^2)} \\ &\le \sqrt{L\sigma^2 \mathbb{E}\|\Psi_l\|_F^2} = \sigma\sqrt{L}\sqrt{\mathrm{tr}(\bar{\Lambda})}, \end{aligned}$$

where the third equality holds by $\mathbb{E}\|A - \mathbb{E}A\|_F^2 = \sum_{i,j} \mathbb{E}(A_{i,j} - \mathbb{E}A_{i,j})^2 = \sum_{i,j} \mathbb{E}A_{i,j}^2 - (\mathbb{E}A_{i,j})^2 = \mathbb{E}\|A\|_F^2 - \|\mathbb{E}A\|_F^2$. We obtained (20) by the following simplifications:

$$\sup_{\|x\|_2 \le 1} \mathbb{E}|x^H \xi_l|^2 \le \mathbb{E}\|\xi_l\|_2^2 \le (2\sigma\gamma\sqrt{R})^2,$$

$$\mathbb{E}\left\| \sum_{l=1}^L \xi_l \right\|_2 \le \mathbb{E}\sum_{l=1}^L \|\xi_l\|_2 \le L\mathbb{E}\|\xi_l\|_2 \le L(2\sigma\gamma\sqrt{R}).$$

Applying (19) and (20) with $t = 2\sigma\gamma\sqrt{R}L\delta$ to the square of (18) yields

$$\left\| \sum_{l=1}^L \Psi_l^H E_l \right\|_F^2 \le L^2 \Big\{ \sigma\sqrt{\mathrm{tr}(\bar{\Lambda})/L} + \|\mathbb{E}(\Psi_l^H E_l)\|_F + 2\sigma\gamma\sqrt{R}\delta \Big\}^2, \tag{21}$$

with probability at least $1 - \exp(-L\frac{\delta^2/2}{3+\delta/3})$.

B. Lower Bound for Denominator

Observe that $\sum_{l=1}^L \Psi_l^H \Psi_l = L\bar{\Lambda} + \sum_{l=1}^L \Lambda_l$, where $\Lambda_l := \Psi_l^H \Psi_l - \bar{\Lambda}$, so Weyl's inequality [17] yields

$$\lambda_{\min}\left(\sum_{l=1}^L \Psi_l^H \Psi_l \right) \ge \lambda_{\min}(L\bar{\Lambda}) - \left\| \sum_{l=1}^L \Lambda_l \right\|_2, \tag{22}$$

and it remains to bound $\|\sum_{l=1}^L \Lambda_l\|_2$. We do so by using the Matrix Bernstein inequality [16, Cor. 8.15].

Note that $\Lambda_1, \ldots, \Lambda_L$ are i.i.d. (since $x_1, \ldots, x_L$ are i.i.d.) and $\mathbb{E}\Lambda_l = 0$. Furthermore, Λ_l is almost surely bounded as

$$\begin{aligned} \|\Lambda_l\|_2 &= \|\Psi_l^H \Psi_l - \mathbb{E}(\Psi_l^H \Psi_l)\|_2 \\ &\le \|\Psi_l^H \Psi_l\|_2 + \|\mathbb{E}(\Psi_l^H \Psi_l)\|_2 && \text{(Triangle ineq.)} \\ &\le \|\Psi_l^H \Psi_l\|_2 + \mathbb{E}\|\Psi_l^H \Psi_l\|_2 && \text{(Jensen's ineq.)} \\ &= \|\Psi_l\|_2^2 + \mathbb{E}\|\Psi_l\|_2^2 \le 2\gamma^2 R. \end{aligned}$$

Thus, the Matrix Bernstein inequality [16, Cor. 8.15] yields that for any $t > 0$,

$$\mathbb{P}\left\{ \left\| \sum_{l=1}^L \Lambda_l \right\|_2 \ge t \right\} \le 2R \exp\left\{ \frac{-t^2/2}{L(2\gamma^2 R)^2 + 2\gamma^2 R t/3} \right\}, \tag{23}$$

where we use the following simplification:

$$\left\| \sum_{l=1}^L \mathbb{E}\Lambda_l^2 \right\|_2 = L\left\|\mathbb{E}\Lambda_l^2\right\|_2 \le L\mathbb{E}\|\Lambda_l\|_2^2 \le L(2\gamma^2 R)^2.$$

Applying (23) with $t = 2\gamma^2 RL\delta$ to the square of (22) yields

$$\lambda_{\min}^2\left(\sum_{l=1}^L \Psi_l^H \Psi_l \right) \ge L^2\{\lambda_{\min}(\bar{\Lambda}) - 2\gamma^2 R\delta\}^2, \tag{24}$$

with probability at least $1 - 2R\exp(-L\frac{\delta^2/2}{1+\delta/3})$.

C. Combined Bound

Combining the bounds (21) and (24) via a union bound yields (9) with probability at least

$$1 - \exp\left(-L\frac{\delta^2/2}{3+\delta/3} \right) - 2R\exp\left(-L\frac{\delta^2/2}{1+\delta/3} \right), \tag{25}$$

which is greater than or equal to (10).

REFERENCES

[1] I. Y. Chun and J. A. Fessler, "Convolutional analysis operator learning: Acceleration and convergence," Jan. 2018, arXiv:1802.05584.

[2] I. Y. Chun and J. A. Fessler, "Convolutional dictionary learning: Acceleration and convergence," *IEEE Trans. Image Process.*, vol. 27, no. 4, pp. 1697–1712, Apr. 2018.

[3] I. Y. Chun and J. A. Fessler, "Convergent convolutional dictionary learning using adaptive contrast enhancement (CDL-ACE): Application of CDL to image denoising," in *Proc. Sampling Theory Appl.*, Tallinn, Estonia, Jul. 2017, pp. 460–464.

[4] I. Y. Chun and J. A. Fessler, "Convolutional analysis operator learning: Application to sparse-view CT," in *Proc. Asilomar Conf. Signals, Syst., Comput.*, Pacific Grove, CA, USA, Oct. 2018, pp. 1631–1635.

[5] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer Series in Statistics). New York, NY, USA: Springer, 2009.

[6] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning*. Cambridge, MA, USA: MIT Press, 2018.

[7] Z. Shakeri, A. D. Sarwate, and W. U. Bajwa, "Sample complexity bounds for dictionary learning from vector- and tensor-valued data," in *Information Theoretic Methods in Data Science*, M. Rodrigues and Y. Eldar, Eds., Cambridge, U.K.: Cambridge Univ. Press, 2019, ch. 5.

[8] S. Singh, B. Póczos, and J. Ma, "Minimax reconstruction risk of convolutional sparse dictionary learning," in *Proc. Int. Conf. Artif. Int. Stat.*, Playa Blanca, Lanzarote, Canary Islands, Apr. 2018, vol. 84, pp. 1327–1336.

[9] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.

[10] M. Yaghoobi, S. Nam, R. Gribonval, and M. E. Davies, "Constrained overcomplete analysis operator learning for cosparse signal modelling," *IEEE Trans. Signal Process.*, vol. 61, no. 9, pp. 2341–2355, Mar. 2013.

[11] S. Hawe, M. Kleinsteuber, and K. Diepold, "Analysis operator learning and its application to image reconstruction," *IEEE Trans. Image Process.*, vol. 22, no. 6, pp. 2138–2150, Jun. 2013.

[12] J.-F. Cai, H. Ji, Z. Shen, and G.-B. Ye, "Data-driven tight frame construction and image denoising," *Appl. Comput. Harmon. Anal.*, vol. 37, no. 1, pp. 89–105, Oct. 2014.

[13] S. Ravishankar and Y. Bresler, "ℓ_0 sparsifying transform learning with efficient optimal updates and convergence guarantees," *IEEE Trans. Signal Process.*, vol. 63, no. 9, pp. 2389–2404, May 2015.

[14] M. Seibert, J. Wörmann, R. Gribonval, and M. Kleinsteuber, "Learning cosparse analysis operators with separable structures," *IEEE Trans. Signal Process.*, vol. 64, no. 1, pp. 120–130, Jan. 2016.

[15] R.-C. Li, "New perturbation bounds for the unitary polar factor," *SIAM J. Matrix Anal. Appl.*, vol. 16, no. 1, pp. 327–332, Jan. 1995.

[16] S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing*. New York, NY, USA: Springer, 2013.

[17] H. Weyl, "Das asymptotische verteilungsgesetz der eigenwerte linearer partieller differentialgleichungen (mit einer anwendung auf die theorie der hohlraumstrahlung)," *Mathematische Annalen*, vol. 71, no. 4, pp. 441–479, Dec. 1912.