Image Space Embeddings and Generalized Convolutional Neural Networks

Nate Strawn

September 20th, 2019

Georgetown University

Table of Contents

Introduction

Inspiration

"When I multiply numbers together, I see two shapes. The image starts to change and evolve, and a third shape emerges. That's the answer. It's mental imagery. It's like maths without having to think."

– Daniel Tammet [6]

Idea

Idea: Embed data into spaces of "smooth" functions over graphs, thereby extending graphical processing techniques to arbitrary datasets.

$$X = \{x_i\}_{i=1}^{N} \subset \mathbb{R}^d$$

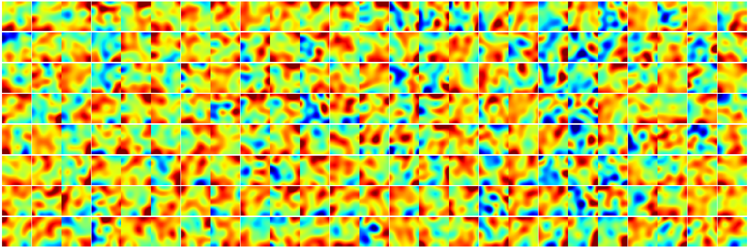
$$\mathbb{R}^d \ni x \xmapsto{\Phi_X} \mathbb{R}^{\mathcal{G}}$$

Implications

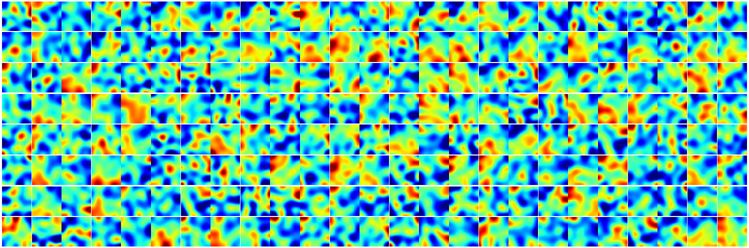
- With $\mathcal{G} = \mathcal{I}_r = \left(\{0, 1, \ldots, r-1\}, \{(k-1, k)\}_{k=1}^{k=r-1}\right)$, Φ_X maps into functions over an interval
- With $\mathcal{G} = \mathcal{I}_r \times \mathcal{I}_r$, Φ_X maps into r by r images
- Wavelet/Curvelet/Shearlet dictionaries for images induce dictionaries for arbitrary datasets
- Convolutional Neural Networks can be applied to arbitrary datasets in a principled manner

Example: Kernel Image Space Embeddings of Tumor Data

Benign Tumors



Malignant Tumors



Smooth Image Space Embeddings

Image Space Embeddings

We will call any isometry $\Phi : \mathbb{R}^d \to C^\infty([0,1]^2)$ or $\Phi : \mathbb{R}^d \to \mathbb{R}^r \otimes \mathbb{R}^r$ an image space embedding.

- $C^\infty([0,1]^2)$ is identified with the space of smooth images with incomplete norm

$$\|f\|^2_{L^2([0,1]^2)} = \int_0^1 \int_0^1 f(x,y)^2 \, dxdy$$

- $\mathbb{R}^r \otimes \mathbb{R}^r$ is identified with the space of r by r matrices, or r by r digital images with norm

$$\|F\|_2^2 = \mathrm{trace}(F^T F).$$

Smoothness of Image Space Embeddings

We will let $\mathcal{D}$ denote:

- the gradient operator on $C^1([0,1]^2)$, or
- the graph derivative $\mathcal{D} : \mathbb{R}^V \to \mathbb{R}^E$ for a graph $\mathcal{G} = (V, E)$ defined by

  $$(\mathcal{D}f)_{(i,\,j)} = f_i - f_j$$

  where $f : \mathbb{R}^V \to \mathbb{R}$ and it is assumed that if $(i, j) \in E$ then $(j, i) \notin E$, and
- the discrete differential $\mathcal{D} : \mathbb{R}^r \otimes \mathbb{R}^r \to \left(\mathbb{R}^r \otimes \mathbb{R}^{r-1}\right) \oplus \left(\mathbb{R}^{r-1} \otimes \mathbb{R}^r\right)$ coincides with the graph derivative on a regular r by r grid

Smoothness of Image Space Embeddings

Given a dataset $X = \{x_i\}_{i=1}^N \subset \mathbb{R}^d$, we measure the smoothness of an image space embedding of X by the mean quadratic variation:

$$MQV(X) = \frac{1}{N}\sum_{i=1}^{N} \|\mathcal{D}(\Phi(x_i))\|^2.$$

Optimally Smooth Image Space Embeddings

We seek the projection which minimizes the mean quadratic variation over the dataset

$$\min_{\Phi} \frac{1}{N} \sum_{i=1}^{N} \|\mathcal{D}(\Phi(x_i))\|_2^2$$

subject to Φ being a linear isometry.

Optimally Smooth Discrete Image Space Embeddings

Theorem (S.)

Suppose $r^2 \geq d$*, let* $\{v_j\}_{j=1}^{d} \subset \mathbb{R}^d$ *be the principal components of* X *(ordered by descending singular values), and let* $\{\xi_j\}_{j=1}^{r^2}$ *(ordered by ascending eigenvalues) denote an orthonormal basis of eigenvectors of the graph Laplacian* $\mathcal{L} = \mathcal{D}^T\mathcal{D}$*. Then*

$$\Phi = \sum_{i=1}^{d} \xi_j v_j^T$$

solves the optimal mean quadratic variation embedding program.

Observations

- The optimal isometry pairs highly variable components in $\mathbb{R}^d$ with low-frequency components in $L^2(\mathcal{G})$.
- $x \mapsto F$ by computing the PCA scores of x, arranging them in an r by r matrix, and applying the inverse discrete cosine transform.
- If the data x_i are drawn i.i.d. from a Gaussian, then Φ maps this Gaussian to a Gaussian process with minimal expected quadratic variation.
- The connection with PCA indicates that we can use Kernel PCA to produce nonlinear embeddings into image spaces as well

Optimally Smooth Continuous Image Space Embeddings

Theorem (S.)

Let $\{v_j\}_{j=1}^d \subset \mathbb{R}^d$ *be the principal components of* X *(ordered by descending singular values), and let* $\{k_j\}_{j=1}^d$ *denote the first* d *positive integer vectors ordered by non-decreasing norm. Then*

$$\Phi(x) = \sum_{j=1}^{d} \left(v_j^T x\right) \exp(2\pi i (k_j^T \cdot))$$

solves the optimal mean quadratic variation embedding program

$$\min_{\Phi} \sum_{i=1}^{N} \|\mathcal{D}\Phi(x_i)\|^2_{L^2_{\mathbb{C}}([0,1]^2)}$$

subject to Φ *being a complex isometry.*

Connection with Regularized PCA

Theorem (S.)

In the discrete case, the solution to the minimum quadratic variation program also provides the optimal Φ for the program

$$\min_{C,\Phi} \frac{1}{2}\|X - C\Phi\|_2^2 + \frac{\lambda}{2}\|C\mathcal{D}^*\|_2^2 + \frac{\gamma}{2}\|C\|_2^2$$

subject to Φ being an isometry.

Example: Dictionary Learning

The Sparse Dictionary Learning Problem

Problem: Given a data matrix $X \in \mathbb{R}^N \otimes \mathbb{R}^d$, with d large, find a linear dictionary $\Phi \in M_{k,\, d}$ and coefficients $C \in M_{N,\, k}$ such that $C\Phi \approx X$, and C is sparse/compressible.

Regularized Factorization

The “relaxed” approach attempts to solve the non-convex program:

$$\min_{C,\Phi} \frac{1}{2}\|X - \Phi^T C\|_2^2 + \lambda\|C\|_1.$$

Usual Suspects

$$\min_{C,\Phi} \frac{1}{2}\|X - C\Phi\|_2^2 + \lambda\|C\|_1$$

- Impose $\|\phi_i\|_2^2 = 1$ for each row of

$$\Phi = \begin{pmatrix} -\phi_1- \\ -\phi_2- \\ \vdots \\ -\phi_k- \end{pmatrix}$$

  to deal with the fact that $C\Phi = (qC)\left(\frac{1}{q}\Phi\right)$.
- Program has analytic solution when C is fixed, and is convex optimization with Φ fixed.

Algorithms

- Optimization algorithm for supervised and online learning of dictionaries: Mairal et al. [9, 8]
- Good initialization procedures can lead to provable results: Agarwal et al. [1]

Identifiability

- Exactly sparse and approximation (even for large factors!) is NP-hard: Tillmann [16]
- Probability model-based learning: Remi and Schnass [11], Spielman et al. [14]
- Dictionary is incoherent and coefficients are sufficiently sparse, then original dictionary is a local minimum: Geng and Wright [5], Schnass [12]
- Full spark matrix is also identifiable given sufficient measurements: Garfinkle and Hillar [4]

Caveats

- Many possible local solutions
- Interpretability?
- Large systems require a large amount of computation!

Tight Frame Dictionaries

Recall that $\{\psi_a\}_{a\in\mathcal{A}} \in L^2(\mathbb{R}^2)$ is a frame if there are constants $0 < A \leq B$ such that

$$A\|x\|^2 \leq \sum_{a\in\mathcal{A}} |\langle f,\, \psi_a\rangle|^2 \leq B\|x\|^2 \text{ for all } f \in \mathcal{H},$$

where $\langle\cdot,\cdot\rangle$ and $\|\cdot\|$ are the inner product and induced norm on $L^2(\mathbb{R}^2)$, respectively. If $A = B$, we say that the frame is tight.

Examples of Tight Frames

- Tensor product wavelet systems
- Curvelets
- Shearlets

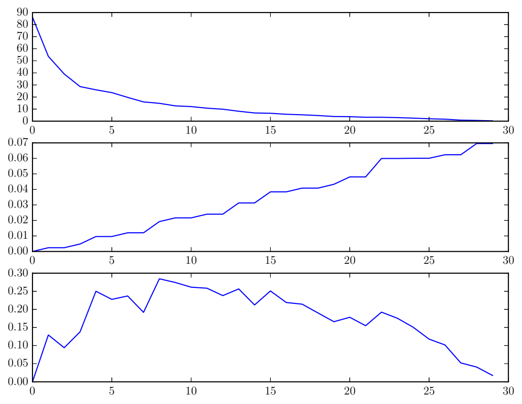
Fact: If $\{\psi_a\}_{a\in\mathcal{A}} \in L^2(\mathbb{R}^2)$ is a tight frame, and $\Phi : \mathbb{R}^d \to L^2(\mathbb{R}^2)$ is an isometry, then $\{\Phi^*\psi_a\}_{a\in\mathcal{A}}$ is a tight frame for $\mathbb{R}^d$.

Example: Wisconsin Breast Cancer Dataset

- 569 examples in $\mathbb{R}^{30}$ describing characteristics of cells obtained from biopsy [15]
- each example is either benign or malignant
- preprocess by removing medians and rescaling by interquartile range in each variable
- image space embedding uses $r = 32$ (images are 32 by 32)

Minimal Mean Quadratic Variation Behavior

PCA Scores vs. eigenvalues of graph Laplacian vs. product



Normalized MMQV ≈ 38

Raw Embeddings of Benign and Malignant Examples

Image Space Embeddings of Benign Tumor Data

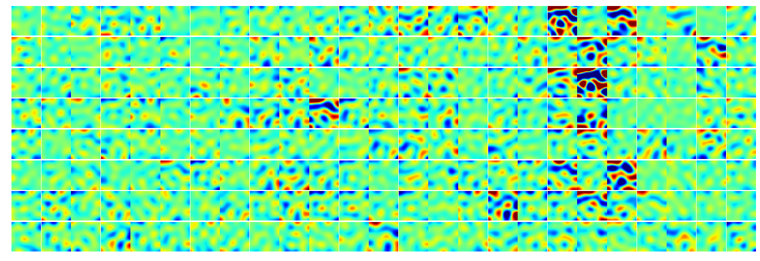
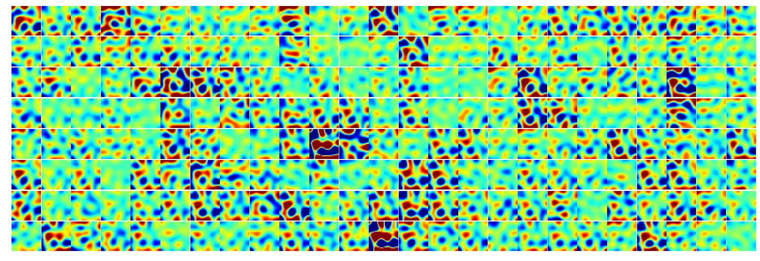


Image Space Embeddings of Malignant Tumor Data



LASSO in the Haar Wavelet Induced Dictionary

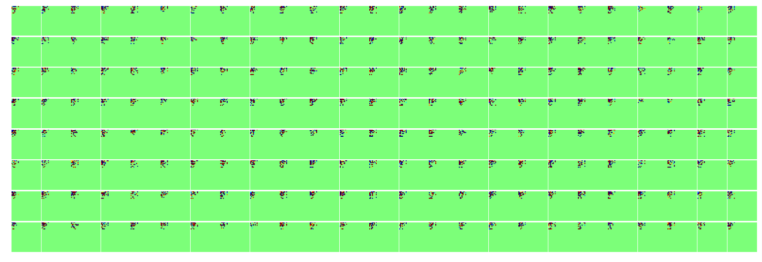
Using the 2D Haar wavelet transform $\mathcal{W}$, we solve

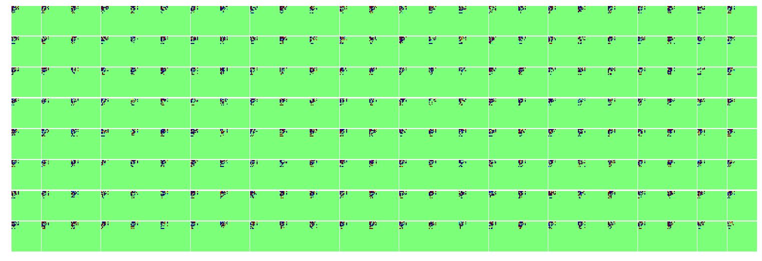
$$\min_{C} \frac{1}{2} \|X - C\mathcal{W}\Phi\|_2^2 + \lambda \|C\|_1$$

where Φ is the image space embedding matrix.

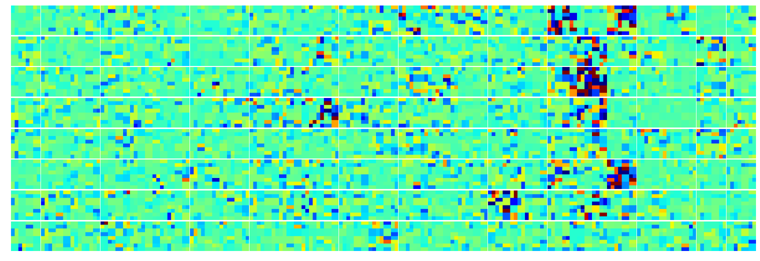
Using BCW dataset, average MSE is 3.4×10^{-3} when $\lambda = 1$.

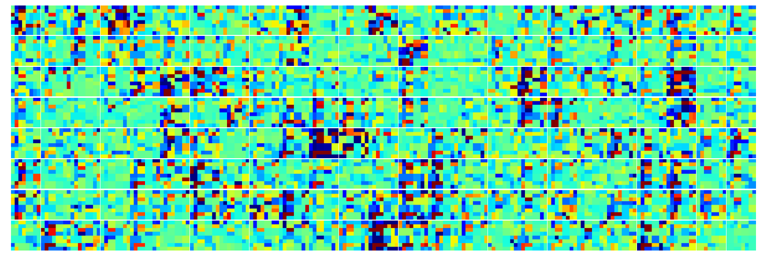
Haar Wavelet Coefficients after LASSO





Inverse DWT of Haar Coefficients

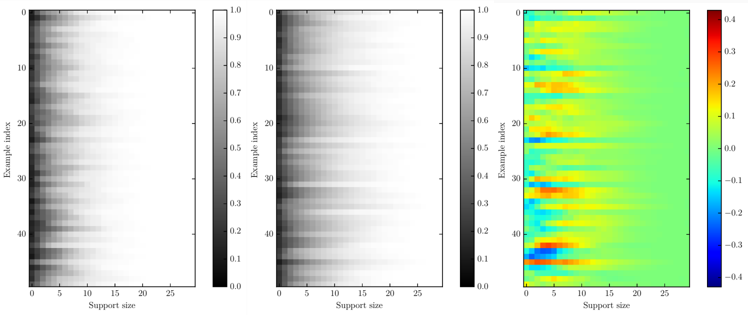




Compression in PCA Basis and Induced Dictionary

Consider best k-term approximations of the first 50 members of the BCW dataset using different dictionaries

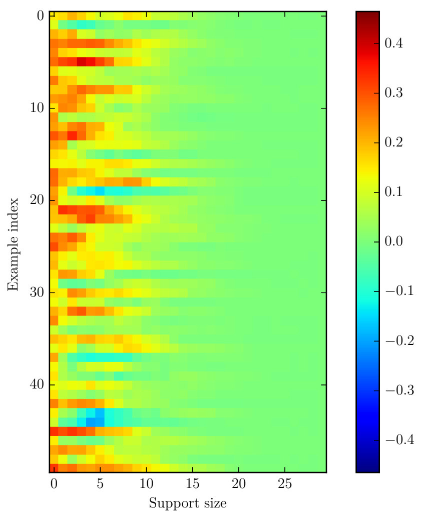
Compression in the dictionary induced by the Haar wavelet system uses orthogonal matching pursuit:



First and second image: Relative SSE for k-term approximations using the PCA basis, Haar-induced dictionary

Third image: First image minus the second image

Comparision with Dictionary Learning



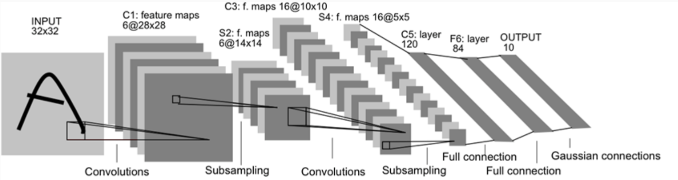
Dictionary learning clearly does better!

Convolutional Neural Networks

People already do this in insane ways!

Convolutional Neural Networks for Arbitrary Datasets

- Exploit image structure to better deal with image collections [7]
- Cutting edge results for image classification tasks



An early (Le-Net5) Convolutional Neural Network design, LeNet-5, used for recognition of digits

Lost in Translation Invariance

- Classification tasks for natural images benefits from translation invariance of class labels
  - Mallat and Bruna [2]
  - Sokolić, Giryes, Sapiro, and Rodrigues [13]
- Almost all image space embeddings of datasets lack this property
- Luckily, translation invariance isn't the whole story
- "Where" features are activated by a convolutional filter may be decisive
  - braille
  - Water and Waffle

More Parameters, More Problems

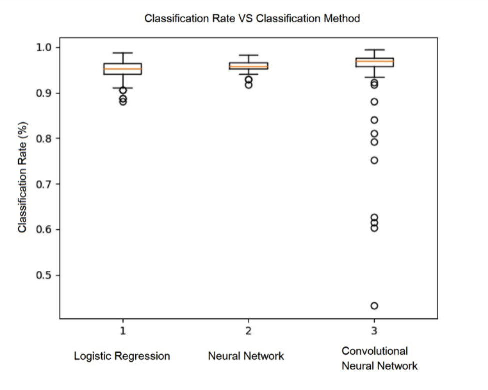
Weight sharing is comparable to regularizing the problem

- Weak evidence via better upper bounds for generalization error [18]
- Precise combinatorial bounds for overfitting? [17]

Experimental Setup

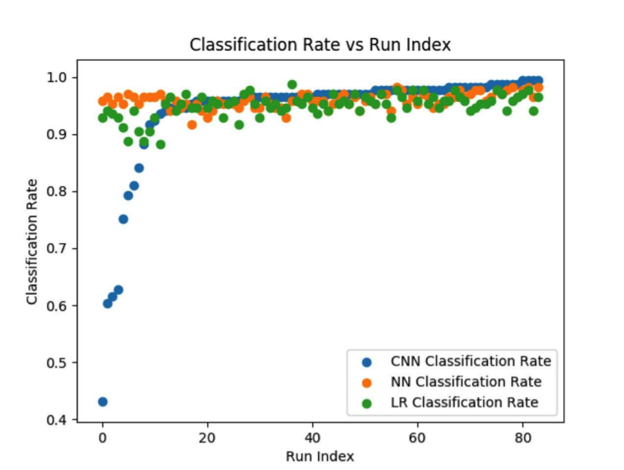
1. Dataset is the image space embedded BCW data
2. For each bootstrap random train/test partition of data, train and test
    - Logistic regression
    - Single hidden layer CNN with softmax activation
    - Single hidden layer NN with softmax activation (same number of units as the CNN)
3. Experiments carried out by Alex Wang of University of Maryland on AWS EC2 GPU instance using TensorFlow

Boxplot Comparision of LR, NN, CNN



Median behavior of CNN is better, but outliers are a problem

Dominance of CNN



CNN generally dominates, but requires more iterations and can sometimes land on bad local minima.

Proofs and Conclusion

Proof for Discrete Case

1. Minimizing MQV is equivalent to minimizing

$$\|\mathcal{D}\Phi X^T\|^2 = \text{trace}\left(X\Phi^T\mathcal{D}^T\mathcal{D}\Phi X^T\right) = \text{trace}\left(\mathcal{L}\Phi X^T X\Phi^T\right)$$

   where $\mathcal{L}$ is the graph Laplacian.
2. Diagonalization of $\mathcal{L}$ reduces this to trace $\left(\Lambda\widetilde{\Phi}X^T X\widetilde{\Phi}^T\right)$, which is the inner product of diag(Λ) with diag($\widetilde{\Phi}X^T X\widetilde{\Phi}^T$).
3. By Schur-Horn, $\alpha = \text{diag}(\widetilde{\Phi}X^T X\widetilde{\Phi}^T)$ for some $\widetilde{\Phi}$ if and only if α is majorized by the eigenvalues of XX^T
4. This reduces the program to a linear program over the polytope generated by permuting the eigenvalues of $X^T X$, and the rearrangement inequality tells us that the minimum is obtained by pairing the eigenvalues of $\mathcal{L}$ and $X^T X$ in reverse order, multiplying, and summing.
5. Continuous case is morally similar, but requires some more care

Conclusion and Future Directions

- Interesting tool for EDA
- Experiments and theory for dictionary learning
- Exploration of overfitting theory for CNN
- Experiments for more UCI datasets
- Minimal Total Variation embeddings and exploitation of approximation rates (Donoho [3]; Needell and Ward [10])

Questions?

References I

[1] Alekh Agarwal, Animashree Anandkumar, Prateek Jain, Praneeth Netrapalli, and Rashish Tandon. Learning sparsely used overcomplete dictionaries. In *Conference on Learning Theory*, pages 123–137, 2014.

[2] Joan Bruna and Stéphane Mallat. Invariant scattering convolution networks. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1872–1886, 2013.

[3] David L Donoho et al. High-dimensional data analysis: The curses and blessings of dimensionality. *AMS math challenges lecture*, 1 (2000):32, 2000.

[4] Charles J Garfinkle and Christopher J Hillar. Robust identifiability in sparse dictionary learning. *arXiv preprint arXiv:1606.06997*, 2016.

References II

[5] Quan Geng and John Wright. On the local correctness of ?
1-minimization for dictionary learning. In *Information Theory (ISIT), 2014 IEEE International Symposium on*, pages 3180–3184. IEEE, 2014.

[6] Richard Johnson. A genius explains. *The Guardian*, 12, 2005.

[7] Yann LeCun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.

[8] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online dictionary learning for sparse coding. In *Proceedings of the 26th annual international conference on machine learning*, pages 689–696. ACM, 2009.

[9] Julien Mairal, Jean Ponce, Guillermo Sapiro, Andrew Zisserman, and Francis R Bach. Supervised dictionary learning. In *Advances in neural information processing systems*, pages 1033–1040, 2009.

References III

[10] Deanna Needell and Rachel Ward. Stable image reconstruction using total variation minimization. *SIAM Journal on Imaging Sciences*, 6(2):1035–1058, 2013.

[11] Rémi Remi and Karin Schnass. Dictionary identification?sparse matrix-factorization via ℓ_1-minimization. *IEEE Transactions on Information Theory*, 56(7):3523–3539, 2010.

[12] Karin Schnass. Local identification of overcomplete dictionaries. *Journal of Machine Learning Research*, 16:1211–1242, 2015.

[13] Jure Sokolic, Raja Giryes, Guillermo Sapiro, and Miguel RD Rodrigues. Generalization error of invariant classifiers. *arXiv preprint arXiv:1610.04574*, 2016.

[14] Daniel A Spielman, Huan Wang, and John Wright. Exact recovery of sparsely-used dictionaries. In *Conference on Learning Theory*, pages 37–1, 2012.

References IV

[15] W Nick Street, William H Wolberg, and Olvi L Mangasarian. Nuclear feature extraction for breast tumor diagnosis. 1992.

[16] Andreas M Tillmann. On the computational intractability of exact and approximate dictionary learning. *IEEE Signal Processing Letters*, 22(1):45–49, 2015.

[17] KV Vorontsov. Combinatorial probability and the tightness of generalization bounds. *Pattern Recognition and Image Analysis*, 18 (2):243–259, 2008.

[18] Yuchen Zhang, Percy Liang, and Martin J Wainwright. Convexified convolutional neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 4044–4053. JMLR. org, 2017.